Multi-SimLex: A Large-Scale Evaluation of Multilingual and Cross-Lingual Lexical Semantic Similarity

https://multisimlex.com/

Ivan Vulić [∗♠]
LTL, University of Cambridge

Simon Baker [∗♠]
LTL, University of Cambridge

Edoardo Maria Ponti [∗♠]
LTL, University of Cambridge

Ulla Petti [∗]
LTL, University of Cambridge

Ira Leviant [∗∗]
Faculty of Industrial Engineering and Management, Technion, IIT

Kelly Wing [∗]
LTL, University of Cambridge

Olga Majewska [∗]
LTL, University of Cambridge

Eden Bar [∗∗]
Faculty of Industrial Engineering and Management, Technion, IIT

Matt Malone [∗]
LTL, University of Cambridge

Thierry Poibeau [†]
LATTICE Lab, CNRS and ENS/PSL and Univ. Sorbonne nouvelle/USPC

Roi Reichart [∗∗]
Faculty of Industrial Engineering and Management, Technion, IIT

Anna Korhonen [∗]
LTL, University of Cambridge

We introduce Multi-SimLex, a large-scale lexical resource and evaluation benchmark covering datasets for 12 typologically diverse languages, including major languages (e.g., Mandarin Chinese, Spanish, Russian) as well as less-resourced ones (e.g., Welsh, Kiswahili). Each language dataset is annotated for the lexical relation of semantic similarity and contains 1,888 semantically aligned concept pairs, providing a representative coverage of word classes (nouns, verbs, adjectives, adverbs), frequency ranks, similarity intervals, lexical fields, and concreteness levels. Additionally, owing to the alignment of concepts across languages, we provide a suite of 66 cross-lingual semantic similarity datasets. Due to its extensive size and language coverage, Multi-SimLex provides entirely novel opportunities for experimental evaluation and analysis. On its monolingual

[∗] ♠*Equal contribution*; English Faculty Building, 9 West Road Cambridge CB3 9DA, United Kingdom. E-mail: `{iv250,sb895,ep490,om304,alk23}@cam.ac.uk`

[∗∗] Technion City, Haifa 3200003, Israel. E-mail: `{ira.leviant,edenb}@campus.technion.ac.il`, `roiri@ie.technion.ac.il`

[†] Rue Maurice Arnoux, 92120 Montrouge, France. E-mail: `thierry.poibeau@ens.fr`

and cross-lingual benchmarks, we evaluate and analyze a wide array of recent state-of-the-art monolingual and cross-lingual representation models, including static and contextualized word embeddings (such as fastText, M-BERT and XLM), externally informed lexical representations, as well as fully unsupervised and (weakly) supervised cross-lingual word embeddings. We also present a step-by-step dataset creation protocol for creating consistent, Multi-Simlex -style resources for additional languages. We make these contributions - the public release of Multi-SimLex datasets, their creation protocol, strong baseline results, and in-depth analyses which can be be helpful in guiding future developments in multilingual lexical semantics and representation learning - available via a website which will encourage community effort in further expansion of Multi-Simlex to many more languages. Such a large-scale semantic resource could inspire significant further advances in NLP across languages.

1. Introduction

The lack of annotated training and evaluation data for many tasks and domains hinders the development of computational models for the majority of the world's languages (Snyder and Barzilay 2010; Adams et al. 2017; Ponti et al. 2019a). The necessity to guide and advance multilingual and cross-lingual NLP through annotation efforts that follow cross-lingually consistent guidelines has been recently recognized by collaborative initiatives such as the Universal Dependency (UD) project (Nivre et al. 2019). The latest version of UD (as of March 2020) covers more than 70 languages. Crucially, this resource continues to steadily grow and evolve through the contributions of annotators from across the world, extending the UD's reach to a wide array of typologically diverse languages. Besides steering research in multilingual parsing (Zeman et al. 2018; Kondratyuk and Straka 2019; Doitch et al. 2019) and cross-lingual parser transfer (Rasooli and Collins 2017; Lin et al. 2019; Rotman and Reichart 2019), the consistent annotations and guidelines have also enabled a range of insightful comparative studies focused on the languages' syntactic (dis)similarities (Bjerva and Augenstein 2018; Ponti et al. 2018a; Pires, Schlinger, and Garrette 2019).

Inspired by the UD work and its substantial impact on research in (multilingual) syntax, in this article we introduce **Multi-SimLex**, a suite of manually and consistently annotated **semantic datasets** for 12 different languages, focused on the fundamental lexical relation of **semantic similarity** (Budanitsky and Hirst 2006; Hill, Reichart, and Korhonen 2015). For any pair of words, this relation measures whether their referents share the same (functional) features, as opposed to general cognitive association captured by co-occurrence patterns in texts (i.e., the distributional information). Datasets that quantify the strength of true semantic similarity between concept pairs such as SimLex-999 (Hill, Reichart, and Korhonen 2015) or SimVerb-3500 (Gerz et al. 2016) have been instrumental in improving models for distributional semantics and representation learning. Discerning between semantic similarity and relatedness/association is not only crucial for theoretical studies on lexical semantics (see §2), but has also been shown to benefit a range of language understanding tasks in NLP. Examples include dialog state tracking (Mrkšić et al. 2017; Ren et al. 2018), spoken language understanding (Kim et al. 2016; Kim, de Marneffe, and Fosler-Lussier 2016), text simplification (Glavaš and Vulić 2018; Ponti et al. 2018b; Lauscher et al. 2019), dictionary and thesaurus construction (Cimiano, Hotho, and Staab 2005; Hill et al. 2016).

Despite the proven usefulness of semantic similarity datasets, they are available only for a small and typologically narrow sample of resource-rich languages such as German, Italian, and Russian (Leviant and Reichart 2015), whereas some language types and low-resource languages typically lack similar evaluation data. Even if some resources do exist, they are limited in their *size* (e.g., 500 pairs in Turkish (Ercan and Yıldız 2018), 500 in Farsi (Camacho-Collados et al. 2017), or 300 in Finnish (Venekoski and Vankka 2017)) and *coverage* (e.g., all datasets which originated from the original English SimLex-999 contain only high-frequent concepts, and are dominated by nouns). This is why, as our departure point, we introduce a **larger and more comprehensive** English word similarity dataset spanning 1,888 concept pairs (see §4).

Most importantly, semantic similarity datasets in different languages have been created using heterogeneous construction procedures with different guidelines for translation and annotation, as well as different rating scales. For instance, some datasets were obtained by directly translating the English SimLex-999 in its entirety (Leviant and Reichart 2015; Mrkšić et al. 2017) or in part (Venekoski and Vankka 2017). Other datasets were created from scratch (Ercan and Yıldız 2018) and yet others sampled English concept pairs differently from SimLex-999 and then translated and reannotated them in target languages (Camacho-Collados et al. 2017). This heterogeneity makes these datasets incomparable and precludes systematic cross-linguistic analyses. In this article, consolidating the lessons learned from previous dataset construction paradigms, we propose a carefully designed **translation and annotation protocol** for developing monolingual Multi-SimLex datasets with aligned concept pairs for typologically diverse languages. We apply this protocol to a set of 12 languages, including a mixture of major languages (e.g., Mandarin, Russian, and French) as well as several low-resource ones (e.g., Kiswahili, Welsh, and Yue Chinese). We demonstrate that our proposed dataset creation procedure yields data with high inter-annotator agreement rates (e.g., the average mean inter-annotator agreement for Welsh is 0.742).

The unified construction protocol and alignment between concept pairs enables a series of quantitative analyses. Preliminary studies on the influence that polysemy and cross-lingual variation in lexical categories (see §2.3) have on similarity judgments are provided in §5. Data created according to Multi-SimLex protocol also allow for probing into whether similarity judgments are universal across languages, or rather depend on linguistic affinity (in terms of linguistic features, phylogeny, and geographical location). We investigate this question in §5.4. Naturally, Multi-SimLex datasets can be used as an intrinsic evaluation benchmark to assess the quality of lexical representations based on monolingual, joint multilingual, and transfer learning paradigms. We conduct a systematic evaluation of several state-of-the-art representation models in §7, showing that there are large gaps between human and system performance in all languages. The proposed construction paradigm also supports the automatic creation of 66 cross-lingual Multi-SimLex datasets by interleaving the monolingual ones. We outline the construction of the cross-lingual datasets in §6, and then present a quantitative evaluation of a series of cutting-edge cross-lingual representation models on this benchmark in §8.

Contributions. We now summarize the main contributions of this work:

1) Building on lessons learned from prior work, we create a more comprehensive lexical semantic similarity dataset for the English language spanning a total of 1,888 concept pairs balanced with respect to similarity, frequency, and concreteness, and covering four word classes: nouns, verbs, adjectives and, for the first time, adverbs. This dataset serves as the main source for the creation of equivalent datasets in several other languages.

2) We present a carefully designed and rigorous language-agnostic translation and annotation protocol. These well-defined guidelines will facilitate the development of future Multi-SimLex datasets for other languages. The proposed protocol eliminates

some crucial issues with prior efforts focused on the creation of multi-lingual semantic resources, namely: i) limited coverage; ii) heterogeneous annotation guidelines; and iii) concept pairs which are semantically incomparable across different languages.

3) We offer to the community manually annotated evaluation sets of 1,888 concept pairs across 12 typologically diverse languages, and 66 large cross-lingual evaluation sets. To the best of our knowledge, Multi-SimLex is the most comprehensive evaluation resource to date focused on the relation of semantic similarity.

4) We benchmark a wide array of recent state-of-the-art monolingual and cross-lingual word representation models across our sample of languages. The results can serve as strong baselines that lay the foundation for future improvements.

5) We present a first large-scale evaluation study on the ability of encoders pretrained on language modeling (such as BERT (Devlin et al. 2019) and XLM (Conneau and Lample 2019)) to reason over word-level semantic similarity in different languages. To our own surprise, the results show that monolingual pretrained encoders, even when presented with word types out of context, are sometimes competitive with static word embedding models such as fastText (Bojanowski et al. 2017) or `word2vec` (Mikolov et al. 2013). The results also reveal a huge gap in performance between massively multilingual pretrained encoders and language-specific encoders in favor of the latter: our findings support other recent empirical evidence related to the "curse of multilinguality" (Conneau et al. 2019; Bapna and Firat 2019) in representation learning.

6) We make all of these resources available on a website which facilitates easy creation, submission and sharing of Multi-Simlex-style datasets for a larger number of languages. We hope that this will yield an even larger repository of semantic resources that inspire future advances in NLP within and across languages.

In light of the success of Universal Dependencies (Nivre et al. 2019), we hope that our initiative will instigate a collaborative public effort with established and clear-cut guidelines that will result in additional Multi-SimLex datasets in a large number of languages in the near future. Moreover, we hope that it will provide means to advance our understanding of distributional and lexical semantics across a large number of languages. All monolingual and cross-lingual Multi-SimLex datasets–along with detailed translation and annotation guidelines–are available online at: `https://multisimlex.com/`.

2. Lexical Semantic Similarity

2.1 Similarity and Association

The focus of the Multi-SimLex initiative is on the lexical relation of pure *semantic similarity*. For any pair of words, this relation measures whether their referents share the same features. For instance, *graffiti* and *frescos* are similar to the extent that they are both forms of painting and appear on walls. This relation can be contrasted with the cognitive *association* between two words, which often depends on how much their referents interact in the real world, or are found in the same situations. For instance, a *painter* is easily associated with *frescos*, although they lack any physical commonalities. Association is also known in the literature under other names: relatedness (Budanitsky and Hirst 2006), topical similarity (McKeown et al. 2002), and domain similarity (Turney 2012).

Semantic similarity and association overlap to some degree, but do not coincide (Kiela, Hill, and Clark 2015; Vulić, Kiela, and Korhonen 2017). In fact, there exist plenty of pairs that are intuitively associated but not similar. Pairs where the converse is true can also be encountered, although more rarely. An example are synonyms where a word is common and the other infrequent, such as *to seize* and *to commandeer*. Hill, Reichart, and Korhonen (2015) revealed that while similarity measures based on the WordNet graph (Wu and Palmer 1994) and human judgments of association in the University of South Florida Free Association Database (Nelson, McEvoy, and Schreiber 2004) do correlate, a number of pairs follow opposite trends. Several studies on human cognition also point in the same direction. For instance, semantic priming can be triggered by similar words without association (Lucas 2000). On the other hand, a connection with cue words is established more quickly for topically related words rather than for similar words in free association tasks (De Deyne and Storms 2008).

A key property of semantic similarity is its *gradience*: pairs of words can be similar to a different degree. On the other hand, the relation of *synonymy* is binary: pairs of words are synonyms if they can be substituted in all contexts (or most contexts, in a looser sense), otherwise they are not. While synonyms can be conceived as lying on one extreme of the semantic similarity continuum, it is crucial to note that their definition is stated in purely relational terms, rather than invoking their referential properties (Lyons 1977; Cruse 1986; Coseriu 1967). This makes behavioral studies on semantic similarity fundamentally different from lexical resources like WordNet (Miller 1995), which include paradigmatic relations (such as synonymy).

2.2 Similarity for NLP: Intrinsic Evaluation and Semantic Specialization

The ramifications of the distinction between similarity and association are profound for distributional semantics. This paradigm of lexical semantics is grounded in the distributional hypothesis, formulated by Firth (1957) and Harris (1951). According to this hypothesis, the meaning of a word can be recovered empirically from the contexts in which it occurs within a collection of texts. Since both pairs of topically related words and pairs of purely similar words tend to appear in the same contexts, their associated meaning confounds the two distinct relations (Hill, Reichart, and Korhonen 2015; Schwartz, Reichart, and Rappoport 2015; Vulić et al. 2017b). As a result, distributional methods obscure a crucial facet of lexical meaning.

This limitation also reflects onto word embeddings (WEs), representations of words as low-dimensional vectors that have become indispensable for a wide range of NLP applications (Collobert et al. 2011; Chen and Manning 2014; Melamud et al. 2016, *inter alia*). In particular, it involves both *static* WEs learned from co-occurrence patterns (Mikolov et al. 2013; Levy and Goldberg 2014; Bojanowski et al. 2017) and *contextualized* WEs learned from modeling word sequences (Peters et al. 2018; Devlin et al. 2019, *inter alia*). As a result, in the induced representations, geometrical closeness (measured e.g. through cosine distance) conflates genuine similarity with broad relatedness. For instance, the vectors for antonyms such as *sober* and *drunk*, by definition dissimilar, might be neighbors in the semantic space under the distributional hypothesis. Turney (2012), Kiela and Clark (2014), and Melamud et al. (2016) demonstrated that different choices of hyper-parameters in WE algorithms (such as context window) emphasize different relations in the resulting representations. Likewise, Agirre et al. (2009) and Levy and Goldberg (2014) discovered that WEs learned from texts annotated with syntactic information mirror similarity better than simple local bag-of-words neighborhoods.

The failure of WEs to capture semantic similarity, in turn, affects model performance in several NLP applications where such knowledge is crucial. In particular, Natural Language Understanding tasks such as statistical dialog modeling, text simplification, or semantic text similarity (Mrkšić et al. 2016; Kim et al. 2016; Ponti et al. 2019c), among others, suffer the most. As a consequence, resources providing clean information on semantic similarity are key in mitigating the side effects of the distributional signal. In particular, such databases can be employed for the *intrinsic evaluations* of specific WE models as a proxy of their reliability for downstream applications (Collobert and Weston 2008; Baroni and Lenci 2010; Hill, Reichart, and Korhonen 2015); intuitively, the more WEs are misaligned with human judgments of similarity, the more their performance on actual tasks is expected to be degraded. Moreover, word representations can be *specialized* (a.k.a. retrofitted) by disentangling word relations of similarity and association. In particular, linguistic constraints sourced from external databases (such as synonyms from WordNet) can be injected into WEs (Faruqui et al. 2015; Wieting et al. 2015; Mrkšić et al. 2017; Lauscher et al. 2019; Kamath et al. 2019, *inter alia*) in order to enforce a particular relation in a distributional semantic space while preserving the original adjacency properties.

2.3 Similarity and Language Variation: Semantic Typology

In this work, we tackle the concept of (true) semantic similarity from a multilingual perspective. While the same meaning representations may be shared by all human speakers at a deep cognitive level, there is no one-to-one mapping between the words in the lexicons of different languages. This makes the comparison of similarity judgments across languages difficult, since the meaning overlap of translationally equivalent words is sometimes far less than exact. This results from the fact that the way languages 'partition' semantic fields is partially arbitrary (Trier 1931), although constrained cross-lingually by common cognitive biases (Majid et al. 2007). For instance, consider the field of colors: English distinguishes between *green* and *blue*, whereas Murle (South Sudan) has a single word for both (Kay and Maffi 2013).

In general, *semantic typology* studies the variation in lexical semantics across the world's languages. According to (Evans 2011), the ways languages categorize concepts into the lexicon follow three main axes: 1) *granularity*: what is the number of categories in a specific domain?; 2) *boundary location*: where do the lines marking different categories lie?; 3) *grouping and dissection*: what are the membership criteria of a category; which instances are considered to be more prototypical? Different choices with respect to these axes lead to different lexicalization patterns.[1] For instance, distinct senses in a polysemous word in English, such as *skin* (referring to both the body and fruit), may be assigned separate words in other languages such as Italian *pelle* and *buccia*, respectively (Rzymski et al. 2020). We later analyze whether similarity scores obtained from native speakers also loosely follow the patterns described by semantic typology.

3. Previous Work and Evaluation Data

Word Pair Datasets. Rich expert-created resources such as WordNet (Miller 1995; Fellbaum 1998), VerbNet (Kipper Schuler 2005; Kipper et al. 2008), or FrameNet (Baker, Fillmore, and Lowe 1998) encode a wealth of semantic and syntactic information, but are expensive and time-consuming to create. The scale of this problem gets multiplied by the number of languages in consideration. Therefore, crowd-sourcing with non-expert annotators has been adopted as a quicker alternative to produce smaller and more focused semantic

[1] More formally, *colexification* is a phenomenon when different meanings can be expressed by the same word in a language (François 2008). For instance, the two senses which are distinguished in English as *time* and *weather* are co-lexified in Croatian: the word *vrijeme* is used in both cases.

resources and evaluation benchmarks. This alternative practice has had a profound impact on distributional semantics and representation learning (Hill, Reichart, and Korhonen 2015). While some prominent English word pair datasets such as WordSim-353 (Finkelstein et al. 2002), MEN (Bruni, Tran, and Baroni 2014), or Stanford Rare Words (Luong, Socher, and Manning 2013) did not discriminate between similarity and relatedness, the importance of this distinction was established by Hill, Reichart, and Korhonen (2015, see again the discussion in §2.1) through the creation of SimLex-999. This inspired other similar datasets which focused on different lexical properties. For instance, SimVerb-3500 (Gerz et al. 2016) provided similarity ratings for 3,500 English verbs, whereas CARD-660 (Pilehvar et al. 2018) aimed at measuring the semantic similarity of infrequent concepts.

Semantic Similarity Datasets in Other Languages. Motivated by the impact of datasets such as SimLex-999 and SimVerb-3500 on representation learning in English, a line of related work focused on creating similar resources in other languages. The dominant approach is translating and reannotating the entire original English SimLex-999 dataset, as done previously for German, Italian, and Russian (Leviant and Reichart 2015), Hebrew and Croatian (Mrkšić et al. 2017), and Polish (Mykowiecka, Marciniak, and Rychlik 2018). Venekoski and Vankka (2017) apply this process only to a subset of 300 concept pairs from the English SimLex-999. On the other hand, Camacho-Collados et al. (2017) sampled a new set of 500 English concept pairs to ensure wider topical coverage and balance across similarity spectra, and then translated those pairs to German, Italian, Spanish, and Farsi (SEMEVAL-500). A similar approach was followed by Ercan and Yıldız (2018) for Turkish, by Huang et al. (2019) for Mandarin Chinese, and by Sakaizawa and Komachi (2018) for Japanese. Netisopakul, Wohlgenannt, and Pulich (2019) translated the concatenation of SimLex-999, WordSim-353, and the English SEMEVAL-500 into Thai and then reannotated it. Finally, Barzegar et al. (2018) translated English SimLex-999 and WordSim-353 to 11 resource-rich target languages (German, French, Russian, Italian, Dutch, Chinese, Portuguese, Swedish, Spanish, Arabic, Farsi), but they did not provide details concerning the translation process and the resolution of translation disagreements. More importantly, they also did not reannotate the translated pairs in the target languages. As we discussed in § 2.3 and reiterate later in §5, semantic differences among languages can have a profound impact on the annotation scores; particulary, we show in §5.4 that these differences even roughly define language clusters based on language affinity.

A core issue with the current datasets concerns a lack of one unified procedure that ensures the comparability of resources in different languages. Further, concept pairs for different languages are sourced from different corpora (e.g., direct translation of the English data versus sampling from scratch in the target language). Moreover, the previous SimLex-based multilingual datasets inherit the main deficiencies of the English original version, such as the focus on nouns and highly frequent concepts. Finally, prior work mostly focused on languages that are widely spoken and do not account for the variety of the world's languages. Our long-term goal is devising a standardized methodology to extend the coverage also to languages that are resource-lean and/or typologically diverse (e.g., Welsh, Kiswahili as in this work).

Multilingual Datasets for Natural Language Understanding. The Multi-SimLex initiative and corresponding datasets are also aligned with the recent efforts on procuring multilingual benchmarks that can help advance computational modeling of natural language understanding across different languages. For instance, pretrained multilingual language models such as multilingual BERT (Devlin et al. 2019) or XLM (Conneau and Lample 2019) are typically probed on XNLI test data (Conneau et al. 2018b) for cross-lingual natural language inference. XNLI was created by translating examples from the English MultiNLI dataset, and projecting its sentence labels (Williams, Nangia, and Bowman 2018). Other recent multilingual datasets target the task of question answering based on reading comprehension: i) MLQA (Lewis et al. 2019) includes 7 languages ii) XQuAD (Artetxe, Ruder, and Yogatama 2019) 10 languages; iii) TyDiQA (Clark et al. 2020) 9 widely spoken typologically diverse languages. While MLQA and XQuAD result from the translation from an English dataset, TyDiQA was built independently in each language. Another multilingual dataset, PAWS-X (Yang et al. 2019), focused on the paraphrase identification task and was created translating the original English PAWS (Zhang, Baldridge, and He 2019) into 6 languages. We believe that Multi-SimLex can substantially contribute to this endeavor by offering a comprehensive multilingual benchmark for the fundamental lexical level relation of semantic similarity. In future work, Multi-SimLex also offers an opportunity to investigate the correlations between word-level semantic similarity and performance in downstream tasks such as QA and NLI across different languages.

4. The Base for Multi-SimLex: Extending English SimLex-999

In this section, we discuss the design principles behind the English (ENG) Multi-SimLex dataset, which is the basis for all the Multi-SimLex datasets in other languages, as detailed in §5. We first argue that a new, more balanced, and more comprehensive evaluation resource for lexical semantic similarity in English is necessary. We then describe how the 1,888 word pairs contained in the ENG Multi-SimLex were selected in such a way as to represent various linguistic phenomena within a single integrated resource.

Construction Criteria. The following criteria have to be satisfied by any high-quality semantic evaluation resource, as argued by previous studies focused on the creation of such resources (Hill, Reichart, and Korhonen 2015; Gerz et al. 2016; Vulić et al. 2017a; Camacho-Collados et al. 2017, *inter alia*):

(C1) Representative and diverse. The resource must cover the full range of diverse concepts occurring in natural language, including different word classes (e.g., nouns, verbs, adjectives, adverbs), concrete and abstract concepts, a variety of lexical fields, and different frequency ranges.

(C2) Clearly defined. The resource must provide a clear understanding of which semantic relation exactly is annotated and measured, possibly contrasting it with other relations. For instance, the original SimLex-999 and SimVerb-3500 explicitly focus on true semantic similarity and distinguish it from broader relatedness captured by datasets such as MEN (Bruni, Tran, and Baroni 2014) or WordSim-353 (Finkelstein et al. 2002).

(C3) Consistent and reliable. The resource must ensure consistent annotations obtained from non-expert native speakers following simple and precise annotation guidelines.

In choosing the word pairs and constructing ENG Multi-SimLex, we adhere to these requirements. Moreover, we follow good practices established by the research on related resources. In particular, since the introduction of the original SimLex-999 dataset (Hill, Reichart, and Korhonen 2015), follow-up works have improved its construction protocol across several aspects, including: 1) coverage of more lexical fields, e.g., by relying on a diverse set of Wikipedia categories (Camacho-Collados et al. 2017), 2) infrequent/rare words (Pilehvar et al. 2018), 3) focus on particular word classes, e.g., verbs (Gerz et al. 2016), 4) annotation quality control (Pilehvar et al. 2018). Our goal is to make use of these improvements towards a larger, more representative, and more reliable lexical similarity dataset in English and, consequently, in all other languages.

The Final Output: English Multi-SimLex. In order to ensure that the criterion C1 is satisfied, we consolidate and integrate the data already carefully sampled in prior work into a single, comprehensive, and representative dataset. This way, we can control for diversity, frequency, and other properties while avoiding to perform this time-consuming selection process from scratch. Note that, on the other hand, the word pairs chosen for English are scored from scratch as part of the entire Multi-SimLex annotation process, introduced later in §5. We now describe the external data sources for the final set of word pairs:

1) *Source: SimLex-999*. (Hill, Reichart, and Korhonen 2015). The English Multi-SimLex has been initially conceived as an extension of the original SimLex-999 dataset. Therefore, we include all 999 word pairs from SimLex, which span 666 noun pairs, 222 verb pairs, and 111 adjective pairs. While SimLex-999 already provides examples representing different POS classes, it does not have a sufficient coverage of different linguistic phenomena: for instance, it contains only very frequent concepts, and it does not provide a representative set of verbs (Gerz et al. 2016).

2) *Source: SemEval-17: Task 2* (henceforth SEMEVAL-500; Camacho-Collados et al. 2017). We start from the full dataset of 500 concept pairs to extract a total of 334 concept pairs for English Multi-SimLex a) which contain only single-word concepts, b) which are not named entities, c) where POS tags of the two concepts are the same, d) where both concepts occur in the top 250K most frequent word types in the English Wikipedia, and e) do not already occur in SimLex-999. The original concepts were sampled as to span all the 34 domains available as part of BabelDomains (Camacho-Collados and Navigli 2017), which roughly correspond to the main high-level Wikipedia categories. This ensures topical diversity in our sub-sample.

3) *Source: CARD-660* (Pilehvar et al. 2018). 67 word pairs are taken from this dataset focused on rare word similarity, applying the same selection criteria a) to e) employed for SEMEVAL-500. Words are controlled for frequency based on their occurrence counts from the Google News data and the ukWaC corpus (Baroni et al. 2009). CARD-660 contains some words that are very rare (*logboat*), domain-specific (*erythroleukemia*) and slang (*2mrw*), which might be difficult to translate and annotate across a wide array of languages. Hence, we opt for retaining only the concept pairs above the threshold of top 250K most frequent Wikipedia concepts, as above.

4) *Source: SimVerb-3500* (Gerz et al. 2016) Since both CARD-660 and SEMEVAL-500 are heavily skewed towards noun pairs, and nouns also dominate the original SimLex-999, we also extract additional verb pairs from the verb-specific similarity dataset SimVerb-3500. We randomly sample 244 verb pairs from SimVerb-3500 that represent all similarity spectra. In particular, we add 61 verb pairs for each of the similarity intervals: $[0, 1.5), [1.5, 3), [3, 4.5), [4.5, 6]$. Since verbs in SimVerb-3500 were originally chosen from VerbNet (Kipper, Snyder, and Palmer 2004; Kipper et al. 2008), they cover a wide range of verb classes and their related linguistic phenomena.

5) *Source: University of South Florida* (USF; Nelson, McEvoy, and Schreiber 2004) norms, the largest database of free association for English. In order to improve the representation of different POS classes, we sample additional adjectives and adverbs from the USF norms following the procedure established by Hill, Reichart, and Korhonen (2015); Gerz et al. (2016). This yields additional 122 adjective pairs, but only a limited number of adverb pairs (e.g., *later – never*, *now – here*, *once – twice*). Therefore, we also create a set

of adverb pairs semi-automatically by sampling adjectives that can be derivationally transformed into adverbs (e.g. adding the suffix *-ly*) from the USF, and assessing the correctness of such derivation in WordNet. The resulting pairs include, for instance, *primarily – mainly*, *softly – firmly*, *roughly – reliably*, etc. We include a total of 123 adverb pairs into the final English Multi-SimLex. Note that this is the first time adverbs are included into any semantic similarity dataset.

Fulfillment of Construction Criteria. The final ENG Multi-SimLex dataset spans 1,051 noun pairs, 469 verb pairs, 245 adjective pairs, and 123 adverb pairs.[2] As mentioned above, the criterion C1 has been fulfilled by relying only on word pairs that already underwent meticulous sampling processes in prior work, integrating them into a single resource. As a consequence, Multi-SimLex allows for fine-grained analyses over different POS classes, concreteness levels, similarity spectra, frequency intervals, relation types, morphology, lexical fields, and it also includes some challenging orthographically similar examples (e.g., *infection – inflection*).[3] We ensure that the criteria C2 and C3 are satisfied by using similar annotation guidelines as Simlex-999, SimVerb-3500, and SEMEVAL-500 that explicitly target semantic similarity. In what follows, we outline the carefully tailored process of translating and annotating Multi-SimLex datasets in all target languages.

5. Multi-SimLex: Translation and Annotation

We now detail the development of the final Multi-SimLex resource, describing our language selection process, as well as translation and annotation of the resource, including the steps taken to ensure and measure the quality of this resource. We also provide key data statistics and preliminary cross-lingual comparative analyses.

Language Selection. Multi-SimLex comprises eleven languages in addition to English. The main objective for our inclusion criteria has been to balance language prominence (by number of speakers of the language) for maximum impact of the resource, while simultaneously having a diverse suite of languages based on their typological features (such as morphological type and language family). Table 1 summarizes key information about the languages currently included in Multi-SimLex. We have included a mixture of fusional, agglutinative, isolating, and introflexive languages that come from eight different language families. This includes languages that are very widely used such as Chinese Mandarin and Spanish, and low-resource languages such as Welsh and Kiswahili. We hope to further include additional languages and inspire other researchers to contribute to the effort over the lifetime of this project.

The work on data collection can be divided into two crucial phases: 1) a translation phase where the extended English language dataset with 1,888 pairs (described in §4) is translated into eleven target languages, and 2) an annotation phase where human raters scored each pair in the translated set as well as the English set. Detailed guidelines for both phases are available online at: `https://multisimlex.com`.

[2] There is a very small number of adjective and verb pairs extracted from CARD-660 and SEMEVAL-500 as well. For instance, the total number of verbs is 469 since we augment the original 222 SimLex-999 verb pairs with 244 SimVerb-3500 pairs and 3 SEMEVAL-500 pairs; and similarly for adjectives.

[3] Unlike SEMEVAL-500 and CARD-660, we do not explicitly control for the equal representation of concept pairs across each similarity interval for several reasons: a) Multi-SimLex contains a substantially larger number of concept pairs, so it is possible to extract balanced samples from the full data; b) such balance, even if imposed on the English dataset, would be distorted in all other monolingual and cross-lingual datasets; c) balancing over similarity intervals arguably does not reflect a true distribution “in the wild” where most concepts are only loosely related or completely unrelated.

Language	ISO 639-3	Family	Type	# Speakers
Chinese Mandarin	CMN	Sino-Tibetan	Isolating	1.116 B
Welsh	CYM	IE: Celtic	Fusional	0.7 M
English	ENG	IE: Germanic	Fusional	1.132 B
Estonian	EST	Uralic	Agglutinative	1.1 M
Finnish	FIN	Uralic	Agglutinative	5.4 M
French	FRA	IE: Romance	Fusional	280 M
Hebrew	HEB	Afro-Asiatic	Introflexive	9 M
Polish	POL	IE: Slavic	Fusional	50 M
Russian	RUS	IE: Slavic	Fusional	260 M
Spanish	SPA	IE: Romance	Fusional	534.3 M
Kiswahili	SWA	Niger-Congo	Agglutinative	98 M
Yue Chinese	YUE	Sino-Tibetan	Isolating	73.5 M

Table 1: The list of 12 languages in the Multi-SimLex multilingual suite along with their corresponding language family (IE = Indo-European), broad morphological type, and their ISO 639-3 code. The number of speakers is based on the total count of L1 and L2 speakers, according to ethnologue.com.

5.1 Word Pair Translation

Translators for each target language were instructed to find direct or approximate translations for the 1,888 word pairs that satisfy the following rules. (1) All pairs in the translated set must be unique (i.e., no duplicate pairs); (2) Translating two words from the same English pair into the same word in the target language is not allowed (e.g., it is not allowed to translate *car* and *automobile* to the same Spanish word *coche*). (3) The translated pairs must preserve the semantic relations between the two words when possible. This means that, when multiple translations are possible, the translation that best conveys the semantic relation between the two words found in the original English pair is selected. (4) If it is not possible to use a single-word translation in the target language, then a multi-word expression (MWE) can be used to convey the nearest possible semantics given the above points (e.g., the English word *homework* is translated into the Polish MWE *praca domowa*).

Satisfying the above rules when finding appropriate translations for each pair–while keeping to the spirit of the intended semantic relation in the English version–is not always straightforward. For instance, kinship terminology in Sinitic languages (Mandarin and Yue) uses different terms depending on whether the family member is older or younger, and whether the family member comes from the mother's side or the father's side. In Mandarin, *brother* has no direct translation and can be translated as either: 哥哥(*older brother*) or 弟弟(*younger brother*). Therefore, in such cases, the translators are asked to choose the best option given the semantic context (relation) expressed by the pair in English, otherwise select one of the translations arbitrarily. This is also used to remove duplicate pairs in the translated set, by differentiating the duplicates using a variant at each instance. Further, many translation instances were resolved using near-synonymous terms in the translation. For example, the words in the pair: *wood – timber* can only be directly translated in Estonian to *puit*, and are not distinguishable. Therefore, the translators approximated the translation for *timber* to the compound noun *puitmaterjal* (literally: *wood material*) in order to produce a valid pair in the target language. In some cases, a direct transliteration from English is used. For example, the pair: *physician* and

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.

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
R1: Start	13	12	14	12	13	10	11	12	12	12	11	13
R3: End	11	10	13	10	10	10	10	10	10	10	10	11

Table 3: Number of human annotators. R1 = Annotation Round 1, R3 = Round 3.

Round 2: We compare the scores of all annotators and identify the pairs for each annotator that have shown the most disagreement. We ask the annotators to reconsider the assigned scores for those pairs only. The annotators may chose to either change or keep the scores. As in the case with Round 1, the annotators have no access to the scores of the other annotators, and the process is anonymous. This process gives a chance for annotators to correct errors or reconsider their judgments, and has been shown to be very effective in reaching consensus, as reported by Pilehvar et al. (2018). We used a very similar procedure as Pilehvar et al. (2018) to identify the pairs with the most disagreement; for each annotator, we marked the ith pair if the rated score s_i falls within: $s_i \geq \mu_i + 1.5$ or $s_i \leq \mu_i - 1.5$, where μ_i is the mean of the other annotators' scores.

Round 3: We compute the average agreement for each annotator (with the other annotators), by measuring the average Spearman's correlation against all other annotators. We discard the scores of annotators that have shown the least average agreement with all other annotators, while we maintain at least ten annotators per language by the end of this round. The actual process is done in multiple iterations: (S1) we measure the average agreement for each annotator with every other annotator (this corresponds to the APIAA measure, see later); (S2) if we still have more than 10 valid annotators and the lowest average score is higher than in the previous iteration, we remove the lowest one, and rerun S1. Table 3 shows the number of annotators at both the start (Round 1) and end (Round 3) of our process for each language.

We measure the agreement between annotators using two metrics, average pairwise inter-annotator agreement (APIAA), and average mean inter-annotator agreement (AMIAA). Both of these use Spearman's correlation (ρ) between annotators scores, the only difference is how they are averaged. They are computed as follows:

$$1)\text{APIAA} = \frac{2\sum_{i,j}\rho(s_i,s_j)}{N(N-1)} \qquad 2)\text{AMIAA} = \frac{\sum_i \rho(s_i,\mu_i)}{N}\text{, where: } \mu_i = \frac{\sum_{j,j\neq i} s_j}{N-1} \tag{1}$$

where $\rho(s_i, s_j)$ is the Spearman's correlation between annotators i and j's scores (s_i,s_j) for all pairs in the dataset, and N is the number of annotators. APIAA has been used widely as the standard measure for inter-annotator agreement, including in the original SimLex paper (Hill, Reichart, and Korhonen 2015). It simply averages the pairwise Spearman's correlation between all annotators. On the other hand, AMIAA compares the average Spearman's correlation of one held-out annotator with the average of all the other $N-1$ annotators, and then averages across all N 'held-out' annotators. It smooths individual annotator effects and arguably serves as a better upper bound than APIAA (Gerz et al. 2016; Vulić et al. 2017a; Pilehvar et al. 2018, *inter alia*).

We present the respective APIAA and AMIAA scores in Table 4 and Table 5 for all part-of-speech subsets, as well as the agreement for the full datasets. As reported in prior work (Gerz et al. 2016; Vulić et al. 2017a), AMIAA scores are typically higher than APIAA scores. Crucially, the results indicate 'strong agreement' (across all languages) using both

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
Nouns	0.661	0.622	0.659	0.558	0.647	0.698	0.538	0.606	0.524	0.582	0.626	0.727
Adjectives	0.757	0.698	0.823	0.695	0.721	0.741	0.683	0.699	0.625	0.64	0.658	0.785
Verbs	0.694	0.604	0.707	0.58	0.644	0.691	0.615	0.593	0.555	0.588	0.631	0.76
Adverbs	0.699	0.593	0.695	0.579	0.646	0.595	0.561	0.543	0.535	0.563	0.562	0.716
Overall	0.68	0.619	0.698	0.583	0.646	0.697	0.572	0.609	0.53	0.576	0.623	0.733

Table 4: Average pairwise inter-annotator agreement (APIAA). A score of 0.6 and above indicates strong agreement.

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
Nouns	0.757	0.747	0.766	0.696	0.766	0.809	0.68	0.717	0.657	0.71	0.725	0.804
Adjectives	0.800	0.789	0.865	0.79	0.792	0.831	0.754	0.792	0.737	0.743	0.686	0.811
Verbs	0.774	0.733	0.811	0.715	0.757	0.808	0.72	0.722	0.69	0.71	0.702	0.784
Adverbs	0.749	0.693	0.777	0.697	0.748	0.729	0.645	0.655	0.608	0.671	0.623	0.716
Overall	0.764	0.742	0.794	0.715	0.76	0.812	0.699	0.723	0.667	0.703	0.71	0.792

Table 5: Average mean inter-annotator agreement (AMIAA). A score of 0.6 and above indicates strong agreement.

measurements. The languages with the highest annotator agreement were French (FRA) and Yue Chinese (YUE), while Russian (RUS) had the lowest overall IAA scores. These scores, however, are still considered to be 'moderately strong agreement'.

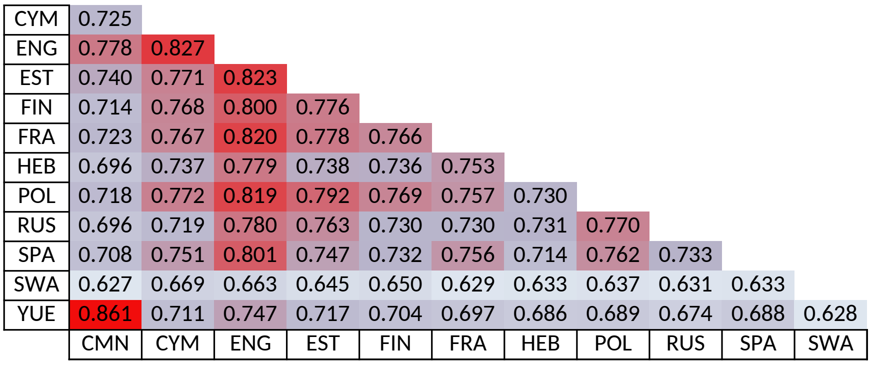
5.3 Data Analysis

Similarity Score Distributions. Across all languages, the average score (mean $= 1.61$, median$= 1.1$) is on the lower side of the similarity scale. However, looking closer at the scores of each language in Table 6, we indicate notable differences in both the averages and the spread of scores. Notably, French has the highest average of similarity scores (mean$= 2.61$, median$= 2.5$), while Kiswahili has the lowest average (mean$= 1.28$, median$= 0.5$). Russian has the lowest spread ($\sigma = 1.37$), while Polish has the largest ($\sigma = 1.62$). All of the languages are strongly correlated with each other, as shown in Figure 1, where all of the Spearman's correlation coefficients are greater than 0.6 for all language pairs. Languages that share the same language family are highly correlated (e.g, CMN-YUE, RUS-POL, EST-FIN). In addition, we observe high correlations between English and most other languages, as expected. This is due to the effect of using English as the base/anchor language to create the dataset. In simple words, if one translates to two languages L_1 and L_2 starting from the same set of pairs in English, it is higly likely that L_1 and L_2 will diverge from English in different ways. Therefore, the similarity between L_1-ENG and L_2-ENG is expected to be higher than between L_1-L_2, especially if L_1 and L_2 are typologically dissimilar languages (e.g., HEB-CMN, see Figure 1). This phenomenon is well documented in related prior work (Leviant and Reichart 2015; Camacho-Collados et al. 2017; Mrkšić et al. 2017; Vulić, Ponzetto, and Glavaš 2019). While we acknowledge this as a slight artifact of the dataset design, it would otherwise be impossible to construct a semantically aligned and comprehensive dataset across a large number of languages.

We also report differences in the distribution of the frequency of words among the languages in Multi-SimLex. Figure 2 shows six example languages, where each bar segment shows the proportion of words in each language that occur in the given frequency range. For example, the 10K-20K segment of the bars represents the proportion

Lang:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
Interval												
$[0, 1)$	56.99	52.01	50.95	35.01	47.83	17.69	28.07	49.36	50.21	43.96	61.39	57.89
$[1, 2)$	8.74	19.54	17.06	30.67	21.35	20.39	35.86	17.32	22.40	22.35	11.86	7.84
$[2, 3)$	13.72	11.97	12.66	16.21	12.02	22.03	16.74	11.86	11.81	14.83	9.11	11.76
$[3, 4)$	11.60	8.32	8.16	10.22	10.17	17.64	8.47	8.95	8.10	9.38	7.10	12.98
$[4, 5)$	6.41	5.83	6.89	6.25	5.61	12.55	6.62	7.57	5.88	6.78	6.30	6.89
$[5, 6]$	2.54	2.33	4.29	1.64	2.97	9.64	4.24	4.93	1.59	2.70	4.24	2.65

Table 6: Fine-grained distribution of concept pairs over different rating intervals in each Multi-SimLex language, reported as percentages. The total number of concept pairs in each dataset is 1,888.



CYM	0.725										
ENG	0.778	0.827									
EST	0.740	0.771	0.823								
FIN	0.714	0.768	0.800	0.776							
FRA	0.723	0.767	0.820	0.778	0.766						
HEB	0.696	0.737	0.779	0.738	0.736	0.753					
POL	0.718	0.772	0.819	0.792	0.769	0.757	0.730				
RUS	0.696	0.719	0.780	0.763	0.730	0.730	0.731	0.770			
SPA	0.708	0.751	0.801	0.747	0.732	0.756	0.714	0.762	0.733		
SWA	0.627	0.669	0.663	0.645	0.650	0.629	0.633	0.637	0.631	0.633	
YUE	0.861	0.711	0.747	0.717	0.704	0.697	0.686	0.689	0.674	0.688	0.628
	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA

Figure 1: Spearman's correlation coefficient (ρ) of the similarity scores for all languages in Multi-SimLex.

of words in the dataset that occur in the list of most frequent words between the frequency rank of 10,000 and 20,000 in that language; likewise with other intervals. Frequency lists for the presented languages are derived from Wikipedia and Common Crawl corpora.[4] While many concept pairs are direct or approximate translations of English pairs, we can see that the frequency distribution does vary across different languages, and is also related to inherent language properties. For instance, in Finnish and Russian, while we use infinitive forms of all verbs, conjugated verb inflections are often more frequent in raw corpora than the corresponding infinitive forms. The variance can also be partially explained by the difference in monolingual corpora size used to derive the frequency rankings in the first place: absolute vocabulary sizes are expected to fluctuate across different languages. However, it is also important to note that the datasets also contain subsets of lower-frequency and rare words, which can be used for rare word evaluations in multiple languages, in the spirit of Pilehvar et al. (2018)'s English rare word dataset.

Cross-Linguistic Differences. Table 7 shows some examples of average similarity scores of English, Spanish, Kiswahili and Welsh concept pairs. Remember that the scores range from 0 to 6: the higher the score, the more similar the participants found the concepts in the pair. The examples from Table 7 show evidence of both the stability of

[4] Frequency lists were obtained from fastText word vectors which are sorted by frequency: https://fasttext.cc/docs/en/crawl-vectors.html

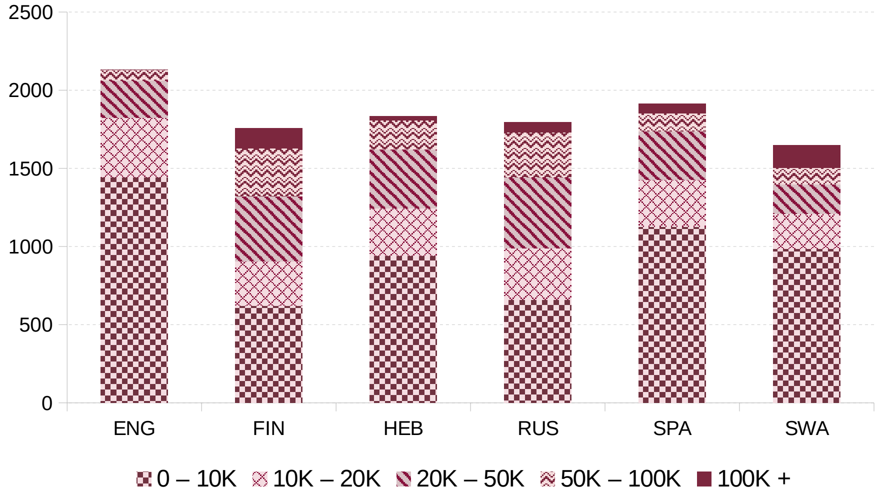


Figure 2: A distribution over different frequency ranges for words from Multi-SimLex datasets for selected languages. Multi-word expressions are excluded from the analysis.

Word Pair	POS	ENG	SPA	SWA	CYM
Similar average rating					
unlikely – friendly	ADV	0	0	0	0
book – literature	N	2.5	2.3	2.1	2.3
vanish – disappear	V	5.2	5.3	5.5	5.3
Different average rating					
regular – average	ADJ	4	4.1	0.5	0.8
care – caution	N	4.1	5.7	0.2	3.1
One language higher					
large – big	ADJ	5.9	2.7	3.8	3.8
bank – seat	N	0	5.1	0	0.1
sunset - evening	N	1.6	1.5	5.5	2.8
purely – completely	ADV	2.3	2.3	1.1	5.4
One language lower					
woman – wife	N	0.9	2.9	4.1	4.8
amazingly – fantastically	ADV	5.1	0.4	4.1	4.1
wonderful – terrific	ADJ	5.3	5.4	0.9	5.7
promise – swear	V	4.8	5.3	4.3	0

Table 7: Examples of concept pairs with their similarity scores from four languages. For brevity, only the original English concept pair is included, but note that the pair is translated to all target languages, see §5.1.

average similarity scores across languages (*unlikely – friendly*, *book – literature*, and *vanish – disappear*), as well as language-specific differences (*care – caution*). Some differences in similarity scores seem to group languages into clusters. For example, the word pair *regular – average* has an average similarity score of 4.0 and 4.1 in English and Spanish, respectively, whereas in Kiswahili and Welsh the average similarity score of this pair is 0.5 and 0.8. We analyze this phenomenon in more detail in §5.4.

There are also examples for each of the four languages having a notably higher or lower similarity score for the same concept pair than the three other languages. For example, *large – big* in English has an average similarity score of 5.9, whereas Spanish, Kiswahili and Welsh speakers rate the closest concept pair in their native language to have a similarity score between 2.7 and 3.8. What is more, *woman – wife* receives an average similarity of 0.9 in English, 2.9 in Spanish, and greater than 4.0 in Kiswahili and Welsh. The examples from Spanish include *banco – asiento* (*bank – seat*) which receives an average similarity score 5.1, while in the other three languages the similarity score for this word pair does not exceed 0.1. At the same time, the average similarity score of *espantosamente – fantásticamente* (*amazingly – fantastically*) is much lower in Spanish (0.4) than in other languages (4.1 – 5.1). In Kiswahili, an example of a word pair with a higher similarity score than the rest would be *machweo – jioni* (*sunset – evening*), having an average score of 5.5, while the other languages receive 2.8 or less, and a notably lower similarity score is given to *wa ajabu - mkubwa sana* (*wonderful – terrific*), getting 0.9, while the other languages receive 5.3 or more. Welsh examples include *yn llwyr - yn gyfan gwbl* (*purely – completely*), which scores 5.4 among Welsh speakers but 2.3 or less in other languages, while *addo – tyngu* (*promise – swear*) is rated as 0 by all Welsh annotators, but in the other three languages 4.3 or more on average.

There can be several explanations for the differences in similarity scores across languages, including but not limited to cultural context, polysemy, metonymy, translation, regional and generational differences, and most commonly, the fact that words and meanings do not exactly map onto each other across languages. For example, it is likely that the other three languages do not have two separate words for describing the concepts in the concept pair: *big – large*, and the translators had to opt for similar lexical items that were more distant in meaning, explaining why in English the concept pair received a much higher average similarity score than in other languages. A similar issue related to the mapping problem across languages arose in the Welsh concept pair *yn llwye – yn gyfan gwbl*, where Welsh speakers agreed that the two concepts are very similar. When asked, bilingual speakers considered the two Welsh concepts more similar than English equivalents *purely – completely*, potentially explaining why a higher average similarity score was reached in Welsh. The example of *woman – wife* can illustrate cultural differences or another translation-related issue where the word 'wife' did not exist in some languages (for example, Estonian), and therefore had to be described using other words, affecting the comparability of the similarity scores. This was also the case with the *football – soccer* concept pair. The pair *bank – seat* demonstrates the effect of the polysemy mismatch across languages: while 'bank' has two different meanings in English, neither of them is similar to the word 'seat', but in Spanish, '*banco*' can mean 'bank', but it can also mean 'bench'. Quite naturally, Spanish speakers gave the pair *banco – asiento* a higher similarity score than the speakers of languages where this polysemy did not occur.

An example of metonymy affecting the average similarity score can be seen in the Kiswahili version of the word pair: *sunset – evening* (*machweo – jioni*). The average similarity score for this pair is much higher in Kiswahili, likely because the word 'sunset' can act as a metonym of 'evening'. The low similarity score of *wonderful – terrific* in Kiswahili (*wa ajabu - mkubwa sana*) can be explained by the fact that while '*mkubwa sana*' can be used as 'terrific' in Kiswahili, it technically means 'very big', adding to the examples of translation- and mapping-related effects. The word pair *amazingly – fantastically* (*espantosamente – fantásticamente*) brings out another translation-related problem: the accuracy of the translation. While '*espantosamente*' could arguably be translated to 'amazingly', more common meanings include: 'frightfully', 'terrifyingly', and 'shockingly', explaining why the average similarity score differs from the rest of the languages. Another problem was brought out by *addo – tyngu* (*promise – swear*) in Welsh,

Language	Word Pair	POS	Rating all participants agree with
ENG	trial – test	N	4-5
SWA	archbishop – bishop	N	4-5
SPA, CYM	start – begin	V	5-6
ENG	smart – intelligent	ADJ	5-6
ENG, SPA	quick – rapid	ADJ	5-6
SPA	circumstance – situation	N	5-6
CYM	football – soccer	N	5-6
SWA	football – soccer	N	6
SWA	pause – wait	V	6
SWA	money – cash	N	6
CYM	friend – buddy	N	6

Table 8: Examples of concept pairs with their similarity scores from four languages where all participants show strong agreement in their rating.

where the '*tyngu*' may not have been a commonly used or even a known word choice for annotators, pointing out potential regional or generational differences in language use.

Table 8 presents examples of concept pairs from English, Spanish, Kiswahili, and Welsh on which the participants agreed the most. For example, in English all participants rated the similarity of *trial – test* to be 4 or 5. In Spanish and Welsh, all participants rated *start – begin* to correspond to a score of 5 or 6. In Kiswahili, *money – cash* received a similarity rating of 6 from every participant. While there are numerous examples of concept pairs in these languages where the participants agreed on a similarity score of 4 or higher, it is worth noting that none of these languages had a single pair where all participants agreed on either 1-2, 2-3, or 3-4 similarity rating. Interestingly, in English all pairs where all the participants agreed on a 5-6 similarity score were adjectives.

5.4 Effect of Language Affinity on Similarity Scores

Based on the analysis in Figure 1 and inspecting the anecdotal examples in the previous section, it is evident that the correlation between similarity scores across languages is not random. To corroborate this intuition, we visualize the vectors of similarity scores for each single language by reducing their dimensionality to 2 via Principal Component Analysis (Pearson 1901). The resulting scatter plot in Figure 3 reveals that languages from the same family or branch have similar patterns in the scores. In particular, Russian and Polish (both Slavic), Finnish and Estonian (both Uralic), Cantonese and Mandarin Chinese (both Sinitic), and Spanish and French (both Romance) are all neighbors.

In order to quantify exactly the effect of language affinity on the similarity scores, we run correlation analyses between these and language features. In particular, we extract feature vectors from URIEL (Littell et al. 2017), a massively multilingual typological database that collects and normalizes information compiled by grammarians and field linguists about the world's languages. In particular, we focus on information about *geography* (the areas where the language speakers are concentrated), *family* (the phylogenetic tree each language belongs to), and typology (including *syntax*, phonological *inventory*, and *phonology*).[5] Moreover, we consider typological representations of languages that are not manually crafted by experts, but rather learned from texts. Malaviya, Neubig, and

[5] For the extraction of these features, we employed `lang2vec`: `github.com/antonisa/lang2vec`

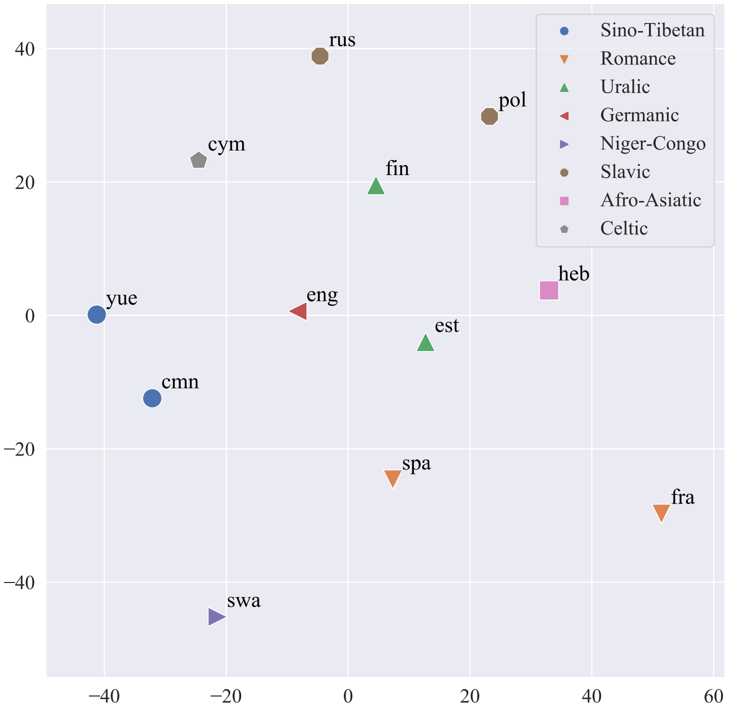


Figure 3: PCA of the language vectors resulting from the concatenation of similarity judgments for all pairs.

Littell (2017) proposed to construct such representations by training language-identifying vectors end-to-end as part of neural machine translation models.

The vector for similarity judgments and the vector of linguistic features for a given language have different dimensionality. Hence, we first construct a distance matrix for each vector space, such that both columns and rows are language indices, and each cell value is the cosine distance between the vectors of the corresponding language pair. Given a set of L languages, each resulting matrix S has dimensionality of $\mathbb{R}^{|L|\times|L|}$ and is symmetrical. To estimate the correlation between the matrix for similarity judgments and each of the matrices for linguistic features, we run a Mantel test (Mantel 1967), a non-parametric statistical test based on matrix permutations that takes into account inter-dependencies among pairwise distances.

The results of the Mantel test reported in Table 3 show that there exist statistically significant correlations between similarity judgments and geography, family, and syntax, given that $p < 0.05$ and $z > 1.96$. The correlation coefficient is particularly strong for geography ($r = 0.647$) and syntax ($r = 0.649$). The former result is intuitive, because languages in contact easily borrow and loan lexical units, and cultural interactions may result in similar cognitive categorizations. The result for syntax, instead, cannot be explained so easily, as formal properties of language do not affect lexical semantics. Instead, we conjecture that, while no causal relation is present, both syntactic features and similarity judgments might be linked to a common explanatory variable (such as geography). In fact, several syntactic properties are not uniformly spread across the globe. For instance, verbs with Verb–Object–Subject word order are mostly concentrated in

Features	Dimension	Mantel r	Mantel p	Mantel z
geography	299	0.647	0.007*	3.443
family	3718	0.329	0.023*	2.711
syntax	103	0.649	0.007*	3.787
inventory	158	0.155	0.459	0.782
phonology	28	0.397	0.046	1.943
Malaviya, Neubig, and Littell (2017)	512	-0.431	0.264	-1.235

Table 9: Mantel test on the correlation between similarity judgments from Multi-SimLex and linguistic features from typological databases.

Oceania (Dryer 2013). In turn, geographical proximity leads to similar judgment patterns, as mentioned above. On the other hand, we find no correlation with phonology and inventory, as expected, nor with the bottom-up typological features from Malaviya, Neubig, and Littell (2017).

6. Cross-Lingual Multi-SimLex Datasets

A crucial advantage of having semantically aligned monolingual datasets across different languages is the potential to create *cross-lingual semantic similarity datasets*. Such datasets allow for probing the quality of cross-lingual representation learning algorithms (Camacho-Collados et al. 2017; Conneau et al. 2018a; Chen and Cardie 2018; Doval et al. 2018; Ruder, Vulić, and Søgaard 2019; Conneau and Lample 2019; Ruder, Søgaard, and Vulić 2019) as an intrinsic evaluation task. However, the cross-lingual datasets previous work relied upon (Camacho-Collados et al. 2017) were limited to a homogeneous set of high-resource languages (e.g., English, German, Italian, Spanish) and a small number of concept pairs (all less than 1K pairs). We address both problems by 1) using a typologically more diverse language sample, and 2) relying on a substantially larger English dataset as a source for the cross-lingual datasets: 1,888 pairs in this work versus 500 pairs in the work of Camacho-Collados et al. (2017). As a result, each of our cross-lingual datasets contains a substantially larger number of concept pairs, as shown in Table 11. The cross-lingual Multi-Simlex datasets are constructed automatically, leveraging word pair translations and annotations collected in all 12 languages. This yields a total of 66 cross-lingual datasets, one for each possible combination of languages. Table 11 provides the final number of concept pairs, which lie between 2,031 and 3,480 pairs for each cross-lingual dataset, whereas Table 10 shows some sample pairs with their corresponding similarity scores.

The automatic creation and verification of cross-lingual datasets closely follows the procedure first outlined by Camacho-Collados, Pilehvar, and Navigli (2015) and later adopted by Camacho-Collados et al. (2017) (for semantic similarity) and Vulić, Ponzetto, and Glavaš (2019) (for graded lexical entailment). First, given two languages, we intersect their aligned concept pairs obtained through translation. For instance, starting from the aligned pairs *attroupement – foule* in French and *rahvasumm – rahvahulk* in Estonian, we construct two cross-lingual pairs *attroupement – rahvaluk* and *rahvasumm – foule*. The scores of cross-lingual pairs are then computed as averages of the two corresponding monolingual scores. Finally, in order to filter out concept pairs whose semantic meaning was not preserved during this operation, we retain only cross-lingual pairs for which the

Pair	Concept-1	Concept-2	Score	Pair	Concept-1	Concept-2	Score
CYM-ENG	rhyddid	liberty	5.37	CMN-EST	可能	optimistlikult	0.83
CYM-POL	plentynaidd	niemądry	2.15	FIN-SWA	psykologia	sayansi	2.20
SWA-ENG	kutimiza	accomplish	5.24	ENG-FRA	normally	quotidiennement	2.41
CMN-FRA	有弹性	flexible	4.08	FIN-SPA	auto	bicicleta	0.85
FIN-SPA	tietämättömyys	inteligencia	0.55	CMN-YUE	使灰心	使气馁	4.78
SPA-FRA	ganador	candidat	2.15	CYM-SWA	sefyllfa	mazingira	1.90
EST-YUE	takso	巴士	2.08	EST-SPA	armee	legión	3.25
ENG-FIN	orange	sitrushedelmä	3.43	FIN-EST	halveksuva	põlglik	5.55
SPA-POL	palabra	wskazówka	0.55	CMN-CYM	学生	disgybl	4.45
POL-SWA	prawdopodobnie	uwezekano	4.05	POL-ENG	grawitacja	meteor	0.27

Table 10: Example concept pairs with their scores from a selection of cross-lingual Multi-SimLex datasets.

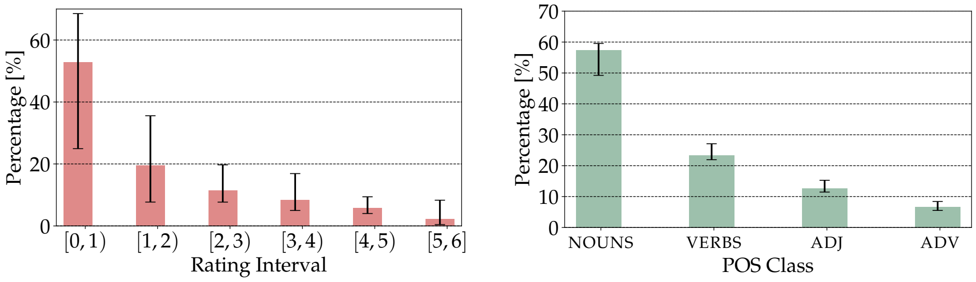
	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
CMN	1,888	–	–	–	–	–	–	–	–	–	–	–
CYM	3,085	1,888	–	–	–	–	–	–	–	–	–	–
ENG	3,151	3,380	1,888	–	–	–	–	–	–	–	–	–
EST	3,188	3,305	3,364	1,888	–	–	–	–	–	–	–	–
FIN	3,137	3,274	3,352	3,386	1,888	–	–	–	–	–	–	–
FRA	2,243	2,301	2,284	2,787	2,682	1,888	–	–	–	–	–	–
HEB	3,056	3,209	3,274	3,358	3,243	2,903	1,888	–	–	–	–	–
POL	3,009	3,175	3,274	3,310	3,294	2,379	3,201	1,888	–	–	–	–
RUS	3,032	3,196	3,222	3,339	3,257	2,219	3,226	3,209	1,888	–	–	–
SPA	3,116	3,205	3,318	3,312	3,256	2,645	3,256	3,250	3,189	1,888	–	–
SWA	2,807	2,926	2,828	2,845	2,900	2,031	2,775	2,819	2,855	2,811	1,888	–
YUE	3,480	3,062	3,099	3,080	3,063	2,313	3,005	2,950	2,966	3,053	2,821	1,888

Table 11: The sizes of all monolingual (main diagonal) and cross-lingual datasets.

corresponding monolingual scores (s_s, s_t) differ at most by one fifth of the full scale (i.e., $\mid s_s - s_t \mid \leq 1.2$). This heuristic mitigates the noise due to cross-lingual semantic shifts (Camacho-Collados et al. 2017; Vulić, Ponzetto, and Glavaš 2019). We refer the reader to the work of Camacho-Collados, Pilehvar, and Navigli (2015) for a detailed technical description of the procedure.

To assess the quality of the resulting cross-lingual datasets, we have conducted a verification experiment similar to Vulić, Ponzetto, and Glavaš (2019). We randomly sampled 300 concept pairs in the English-Spanish, English-French, and English-Mandarin cross-lingual datasets. Subsequently, we asked bilingual native speakers to provide similarity judgments of each pair. The Spearman's correlation score ρ between automatically induced and manually collected ratings achieves $\rho \geq 0.90$ on all samples, which confirms the viability of the automatic construction procedure.

Score and Class Distributions. The summary of score and class distributions across all 66 cross-lingual datasets are provided in Figure 4a and Figure 4b, respectively. First, it is obvious that the distribution over the four POS classes largely adheres to that of the original monolingual Multi-SimLex datasets, and that the variance is quite low: e.g., the ENG-FRA dataset contains the lowest proportion of nouns (49.21%) and the highest proportion of verbs (27.1%), adjectives (15.28%), and adverbs (8.41%). On the other hand, the distribution over similarity intervals in Figure 4a shows a much greater variance. This is again expected as this pattern resurfaces in monolingual datasets (see Table 6). It is also evident that the data are skewed towards lower-similarity concept pairs. However, due to the joint size of all cross-lingual datasets (see Table 11), even the least represented



(a) Rating distribution

(b) Distribution over POS classes

Figure 4: **(a)** Rating distribution and **(b)** distribution of pairs over the four POS classes in cross-lingual Multi-SimLex datasets averaged across each of the 66 language pairs (y axes plot percentages as the total number of concept pairs varies across different cross-lingual datasets). Minimum and maximum percentages for each rating interval and POS class are also plotted.

intervals contain a substantial number of concept pairs. For instance, the RUS-YUE dataset contains the least highly similar concept pairs (in the interval $[4, 6]$) of all 66 cross-lingual datasets. Nonetheless, the absolute number of pairs (138) in that interval for RUS-YUE is still substantial. If needed, this makes it possible to create smaller datasets which are balanced across the similarity spectra through sub-sampling.

7. Monolingual Evaluation of Representation Learning Models

After the numerical and qualitative analyses of the Multi-SimLex datasets provided in §§ 5.3–5.4, we now benchmark a series of representation learning models on the new evaluation data. We evaluate standard static word embedding algorithms such as fastText (Bojanowski et al. 2017), as well as a range of more recent text encoders pretrained on language modeling such as multilingual BERT (Devlin et al. 2019). These experiments provide strong baseline scores on the new Multi-SimLex datasets and offer a first large-scale analysis of pretrained encoders on word-level semantic similarity across diverse languages. In addition, the experiments now enabled by Multi-SimLex aim to answer several important questions. **(Q1)** Is it viable to extract high-quality word-level representations from pretrained encoders receiving subword-level tokens as input? Are such representations competitive with standard static word-level embeddings? **(Q2)** What are the implications of monolingual pretraining versus (massively) multilingual pretraining for performance? **(Q3)** Do lightweight unsupervised post-processing techniques improve word representations consistently across different languages? **(Q4)** Can we effectively transfer available external lexical knowledge from resource-rich languages to resource-lean languages in order to learn word representations that distinguish between true similarity and conceptual relatedness (see the discussion in §2.3)?

7.1 Models in Comparison

Static Word Embeddings in Different Languages. First, we evaluate a standard method for inducing non-contextualized (i.e., static) word embeddings across a plethora of different languages: FASTTEXT (FT) vectors (Bojanowski et al. 2017) are currently the most popular and robust choice given 1) the availability of pretrained vectors in a large number of languages (Grave et al. 2018) trained on large Common Crawl (CC) plus Wikipedia (Wiki) data, and 2) their superior performance across a range of NLP tasks (Mikolov et al. 2018). In fact, FASTTEXT is an extension of the standard word-level CBOW and skip-gram `word2vec` models (Mikolov et al. 2013) that takes into account subword-level information, i.e. the contituent character n-grams of each word (Zhu, Vulić, and Korhonen 2019). For this reason, FASTTEXT is also more suited for modeling rare words and morphologically rich languages.[6]

We rely on 300-dimensional FT word vectors trained on CC+Wiki and available online for 157 languages.[7] The word vectors for all languages are obtained by CBOW with position-weights, with character n-grams of length 5, a window of size 5, 10 negative examples, and 10 training epochs. We also probe another (older) collection of FT vectors, pretrained on full Wikipedia dumps of each language.[8]. The vectors are 300-dimensional, trained with the skip-gram objective for 5 epochs, with 5 negative examples, a window size set to 5, and relying on all character n-grams from length 3 to 6. Following prior work, we trim the vocabularies for all languages to the 200K most frequent words and compute representations for multi-word expressions by averaging the vectors of their constituent words.

Unsupervised Post-Processing. Further, we consider a variety of *unsupervised post-processing* steps that can be applied post-training on top of any pretrained input word embedding space *without* any external lexical semantic resource. So far, the usefulness of such methods has been verified only on the English language through benchmarks for lexical semantics and sentence-level tasks (Mu, Bhat, and Viswanath 2018). In this paper, we assess if unsupervised post-processing is beneficial also in other languages. To this end, we apply the following post-hoc transformations on the initial word embeddings:

1) *Mean centering* (MC) is applied after unit length normalization to ensure that all vectors have a zero mean, and is commonly applied in data mining and analysis (Bro and Smilde 2003; van den Berg et al. 2006).

2) *All-but-the-top* (ABTT) (Mu, Bhat, and Viswanath 2018; Tang, Mousavi, and de Sa 2019) eliminates the common mean vector and a few top dominating directions (according to PCA) from the input distributional word vectors, since they do not contribute towards distinguishing the actual semantic meaning of different words. The method contains a single (tunable) hyper-parameter dd_A which denotes the number of the dominating directions to remove from the initial representations. Previous work has verified the usefulness of ABTT in several English lexical semantic tasks such as semantic similarity, word analogies, and concept categorization, as well as in sentence-level text classification tasks (Mu, Bhat, and Viswanath 2018).

3) UNCOVEC (Artetxe et al. 2018) adjusts the similarity order of an arbitrary input word embedding space, and can emphasize either syntactic or semantic information in the transformed vectors. In short, it transforms the input space $\boldsymbol{X}$ into an adjusted space $\boldsymbol{X}\boldsymbol{W}_\alpha$ through a linear map $\boldsymbol{W}_\alpha$ controlled by a single hyper-parameter α. The n^{th}-

[6] We have also trained standard word-level CBOW and skip-gram with negative sampling (SGNS) on full Wikipedia dumps for several languages, but our preliminary experiments have verified that they under-perform compared to FASTTEXT. This finding is consistent with other recent studies demonstrating the usefulness of subword-level information (Vania and Lopez 2017; Mikolov et al. 2018; Zhu, Vulić, and Korhonen 2019; Zhu et al. 2019). Therefore, we do not report the results with CBOW and SGNS for brevity.

[7] https://fasttext.cc/docs/en/crawl-vectors.html

[8] https://fasttext.cc/docs/en/pretrained-vectors.html

order similarity transformation of the input word vector space $\boldsymbol{X}$ (for which $n = 1$) can be obtained as $\boldsymbol{M}_n(\boldsymbol{X}) = \boldsymbol{M}_1(\boldsymbol{X}\boldsymbol{W}_{(n-1)/2})$, with $\boldsymbol{W}_\alpha = \boldsymbol{Q}\boldsymbol{\Gamma}^\alpha$, where $\boldsymbol{Q}$ and $\boldsymbol{\Gamma}$ are the matrices obtained via eigendecomposition of $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Gamma}\boldsymbol{Q}^T$. $\boldsymbol{\Gamma}$ is a diagonal matrix containing eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$; $\boldsymbol{Q}$ is an orthogonal matrix with eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ as columns. While the motivation for the UNCOVEC methods does originate from adjusting discrete similarity orders, note that α is in fact a continuous real-valued hyper-parameter which can be carefully tuned. For more technical details we refer the reader to the original work of Artetxe et al. (2018).

As mentioned, all post-processing methods can be seen as unsupervised retrofitting methods that, given an arbitrary input vector space $\boldsymbol{X}$, produce a perturbed/transformed output vector space $\boldsymbol{X}'$, but unlike common retrofitting methods (Faruqui et al. 2015; Mrkšić et al. 2017), the perturbation is completely unsupervised (i.e., self-contained) and does not inject any external (semantic similarity oriented) knowledge into the vector space. Note that different perturbations can also be stacked: e.g., we can apply UNCOVEC and then use ABTT on top the output UNCOVEC vectors. When using UNCOVEC and ABTT we always length-normalize and mean-center the data first (i.e., we apply the simple MC normalization). Finally, we tune the two hyper-parameters d_A (for ABTT) and α (UNCOVEC) on the English Multi-SimLex and use the same values on the datasets of all other languages; we report results with $dd_A = 3$ or $dd_A = 10$, and $\alpha = -0.3$.

Contextualized Word Embeddings. We also evaluate the capacity of unsupervised pretraining architectures based on language modeling objectives to reason over lexical semantic similarity. To the best of our knowledge, our article is the first study performing such analyses. State-of-the-art models such as BERT (Devlin et al. 2019), XLM (Conneau and Lample 2019), or ROBERTA (Liu et al. 2019b) are typically very deep neural networks based on the Transformer architecture (Vaswani et al. 2017). They receive subword-level tokens as inputs (such as WordPieces (Schuster and Nakajima 2012)) to tackle data sparsity. In output, they return contextualized embeddings, dynamic representations for words in context.

To represent words or multi-word expressions through a pretrained model, we follow prior work (Liu et al. 2019a) and compute an input item's representation by 1) feeding it to a pretrained model *in isolation*; then 2) averaging the H last hidden representations for each of the item's constituent subwords; and then finally 3) averaging the resulting subword representations to produce the final d-dimensional representation, where d is the embedding and hidden-layer dimensionality (e.g., $d = 768$ with BERT). We opt for this approach due to its proven viability and simplicity (Liu et al. 2019a), as it does not require any additional corpora to condition the induction of contextualized embeddings.[9] Other ways to extract the representations from pretrained models (Aldarmaki and Diab 2019; Wu et al. 2019; Cao, Kitaev, and Klein 2020) are beyond the scope of this work, and we will experiment with them in the future.

In other words, we treat each pretrained encoder ENC as a black-box function to encode a single word or a multi-word expression x in each language into a d-dimensional contextualized representation $\mathbf{x}_{\text{ENC}} \in \mathbb{R}^d = \text{ENC}(x)$ (e.g., $d = 768$ with BERT). As multilingual pretrained encoders, we experiment with the multilingual BERT model (M-BERT) (Devlin et al. 2019) and XLM (Conneau and Lample 2019). M-BERT is pretrained

[9] We also tested another encoding method where we fed pairs instead of single words/concepts into the pretrained encoder. The rationale is that the other concept in the pair can be used as disambiguation signal. However, this method consistently led to sub-par performance across all experimental runs.

on monolingual Wikipedia corpora of 102 languages (comprising all Multi-SimLex languages) with a 12-layer Transformer network, and yields 768-dimensional representations. Since the concept pairs in Multi-SimLex are lowercased, we use the uncased version of M-BERT.[10] M-BERT comprises all Multi-SimLex languages, and its evident ability to perform cross-lingual transfer (Pires, Schlinger, and Garrette 2019; Wu and Dredze 2019; Wang et al. 2020) also makes it a convenient baseline model for cross-lingual experiments later in §8. The second multilingual model we consider, XLM-100,[11] is pretrained on Wikipedia dumps of 100 languages, and encodes each concept into a $1,280$-dimensional representation. In contrast to M-BERT, XLM-100 drops the next-sentence prediction objective and adds a cross-lingual masked language modeling objective. For both encoders, the representations of each concept are computed as averages over the last $H = 4$ hidden layers in all experiments, as suggested by Wu et al. (2019).[12]

Besides M-BERT and XLM, covering multiple languages, we also analyze the performance of "language-specific" BERT and XLM models for the languages where they are available: Finnish, Spanish, English, Mandarin Chinese, and French. The main goal of this comparison is to study the differences in performance between multilingual "one-size-fits-all" encoders and language-specific encoders. For all experiments, we rely on the pretrained models released in the Transformers repository (Wolf et al. 2019).[13]

Unsupervised post-processing steps devised for static word embeddings (i.e., mean-centering, ABTT, UNCOVEC) can also be applied on top of contextualized embeddings if we predefine a vocabulary of word types V that will be represented in a word vector space $\mathbf{X}$. We construct such V for each language as the intersection of word types covered by the corresponding CC+Wiki fastText vectors and the (single-word or multi-word) expressions appearing in the corresponding Multi-SimLex dataset.

Finally, note that it is not feasible to evaluate a full range of available pretrained encoders within the scope of this work. Our main intention is to provide the first set of baseline results on Multi-SimLex by benchmarking a sample of most popular encoders, at the same time also investigating other important questions such as performance of static versus contextualized word embeddings, or multilingual versus language-specific pretraining. Another purpose of the experiments is to outline the wide potential and applicability of the Multi-SimLex datasets for multilingual and cross-lingual representation learning evaluation.

7.2 Results and Discussion

The results we report are Spearman's ρ coefficients of the correlation between the ranks derived from the scores of the evaluated models and the human scores provided in each Multi-SimLex dataset. The main results with static and contextualized word vectors for all test languages are summarized in Table 12. The scores reveal several interesting patterns, and also pinpoint the main challenges for future work.

[10] https://github.com/google-research/bert/blob/master/multilingual.md

[11] https://github.com/facebookresearch/XLM

[12] In our preliminary experiments on several language pairs, we have also verified that this choice is superior to: a) using the output of only the last hidden layer (i.e., $H = 1$) and b) averaging over all hidden layers (i.e., $H = 12$ for the BERT-BASE architecture). Likewise, using the special prepended '[CLS]' token rather than the constituent sub-words to encode a concept also led to much worse performance across the board.

[13] github.com/huggingface/transformers. The full list of currently supported pretrained encoders is available here: huggingface.co/models.

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
FASTTEXT (CC+Wiki)	*(272)*	*(151)*	*(12)*	*(319)*	*(347)*	*(43)*	*(66)*	*(326)*	*(291)*	*(46)*	*(222)*	*(–)*
(1) FT:INIT	.534	.363	.528	.469	.607	.578	.450	.405	.422	.511	.439	–
(2) FT:+MC	.539	**.393**	.535	.473	.621	.584	.480	.412	.424	.516	.469	–
(3) FT:+ABTT (-3)	.557	.389	.536	**.495**	.642	.610	.501	.427	.459	.523	**.473**	–
(4) FT:+ABTT (-10)	**.583**	.384	**.551**	.476	.651	.623	.503	.455	**.500**	**.542**	.462	–
(5) FT:+UNCOVEC	.572	.387	.550	.465	.642	.595	.501	.435	.437	.525	.437	–
(1)+(2)+(5)+(3)	.574	.386	.549	.476	**.655**	.604	.503	.442	.452	.528	.432	–
(1)+(2)+(5)+(4)	.577	.376	.542	.455	.652	**.613**	**.510**	**.466**	.491	.540	.424	–
FASTTEXT (Wiki)	*(429)*	*(282)*	*(6)*	*(343)*	*(345)*	*(73)*	*(62)*	*(354)*	*(343)*	*(57)*	*(379)*	*(677)*
(1) FT:INIT	.315	.318	.436	.400	.575	.444	.428	.370	.359	.432	.332	.376
(2) FT:+MC	.373	.337	.445	**.404**	.583	.463	.447	.383	.378	.447	.373	.427
(3) FT:+ABTT (-3)	.459	**.343**	.453	**.404**	**.584**	.487	.447	.387	.394	.456	**.423**	**.429**
(4) FT:+ABTT (-10)	.496	.323	.460	.385	.581	.494	.460	**.401**	**.400**	**.477**	.406	.399
(5) FT:+UNCOVEC	.518	.328	.469	.375	.568	.483	.449	.389	.387	.469	.386	.394
(1)+(2)+(5)+(3)	**.526**	.323	.470	.369	.564	**.495**	.448	.392	.392	.473	.388	.388
(1)+(2)+(5)+(4)	**.526**	.307	**.471**	.355	.548	**.495**	**.450**	.394	.394	.476	.382	.396
M-BERT	*(0)*	*(0)*	*(0)*	*(0)*	*(0)*	*(0)*	*(0)*	*(0)*	*(0)*	*(0)*	*(0)*	*(0)*
(1) M-BERT:INIT	.408	.033	.138	.085	.162	.115	.104	.069	.085	.145	.125	.404
(2) M-BERT:+MC	.458	.044	.256	.122	.173	.183	.128	.097	.123	.203	.128	.469
(3) M-BERT:+ABTT (-3)	**.487**	.056	.321	.137	.200	.287	.144	.126	.197	.299	.135	**.492**
(4) M-BERT:+ABTT (-10)	.456	.056	**.329**	.122	.164	**.306**	.121	.126	.183	**.315**	.136	.467
(5) M-BERT:+UNCOVEC	.464	.063	.317	**.144**	**.213**	.288	**.164**	**.144**	.198	.287	.143	.464
(1)+(2)+(5)+(3)	.464	.083	.326	.130	.201	.304	.149	.122	**.199**	.295	**.148**	.456
(1)+(2)+(5)+(4)	.444	**.086**	.326	.112	.179	.305	.135	.127	.187	.285	.119	.447

Table 12: A summary of results (Spearman's ρ correlation scores) on the full monolingual Multi-SimLex datasets for 12 languages. We benchmark fastText word embeddings trained on two different corpora (CC+Wiki and only Wiki) as well the multilingual M-BERT model (see §7.1). Results with the initial word vectors are reported (i.e., without any unsupervised post-processing), as well as with different unsupervised post-processing methods, described in §7.1. The language codes are provided in Table 1. The numbers in the parentheses (gray rows) refer to the number of OOV concepts excluded from the computation. The highest scores for each language and per model are in **bold**.

State-of-the-Art Representation Models. The absolute scores of CC+Wiki FT, Wiki FT, and M-BERT are not directly comparable, because these models have different coverage. In particular, Multi-SimLex contains some out-of-vocabulary (OOV) words whose static FT embeddings are not available.[14] On the other hand, M-BERT has perfect coverage. A general comparison between CC+Wiki and Wiki FT vectors, however, supports the intuition that larger corpora (such as CC+Wiki) yield higher correlations. Another finding is that a single massively multilingual model such as M-BERT cannot produce semantically rich word-level representations. Whether this actually happens because the training objective is different—or because the need to represent 100+ languages reduces its language-specific capacity—is investigated further below.

The overall results also clearly indicate that (i) there are differences in performance across different monolingual Multi-SimLex datasets, and (ii) unsupervised post-

[14] We acknowledge that it is possible to approximate word-level representations of OOVs with FT by summing the constituent n-gram embeddings as proposed by Bojanowski et al. (2017). However, we do not perform this step as the resulting embeddings are typically of much lower quality than non-OOV embeddings (Zhu, Vulić, and Korhonen 2019).

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
XLM-100						XLM:+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, XLM-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)

summarized in Table 12 are the first evidence that also confirms its importance for semantic similarity in a wide array of languages. In sum, as a general rule of thumb, we suggest to always mean-center representations for semantic tasks.

The results further indicate that additional post-processing methods such as ABTT and UNCOVEC on top of mean-centered vector spaces can lead to further gains in most languages. The gains are even visible for languages which start from high correlation scores: for instance., CMN with CC+Wiki FT increases from 0.534 to 0.583, from 0.315 to 0.526 with Wiki FT, and from 0.408 to 0.487 with M-BERT. Similarly, for RUS with CC+Wiki FT we can improve from 0.422 to 0.500, and for FRA the scores improve from 0.578 to 0.613. There are additional similar cases reported in Table 12.

Overall, the unsupervised post-processing techniques seem universally useful across languages, but their efficacy and relative performance does vary across different languages. Note that we have not carefully fine-tuned the hyper-parameters of the evaluated post-processing methods, so additional small improvements can be expected for some languages. The main finding, however, is that these post-processing techniques are robust to semantic similarity computations beyond English, and are truly language independent. For instance, removing dominant latent (PCA-based) components from word vectors emphasizes semantic differences between different concepts, as only shared non-informative latent semantic knowledge is removed from the representations.

In summary, pretrained word embeddings do contain more information pertaining to semantic similarity than revealed in the initial vectors. This way, we have corroborated the hypotheses from prior work (Mu, Bhat, and Viswanath 2018; Artetxe et al. 2018) which were not previously empirically verified on other languages due to a shortage of evaluation data; this gap has now been filled with the introduction of the Multi-SimLex datasets. In all follow-up experiments, we always explicitly denote which post-processing configuration is used in evaluation.

POS-Specific Subsets. We present the results for subsets of word pairs grouped by POS class in Table 13. Prior work based on English data showed that representations for nouns are typically of higher quality than those for the other POS classes (Schwartz, Reichart, and Rappoport 2015, 2016; Vulić et al. 2017b). We observe a similar trend in other languages as well. This pattern is consistent across different representation models and can be attributed to several reasons. First, verb representations need to express a rich range of syntactic and semantic behaviors rather than purely referential features (Gruber 1976; Levin 1993; Kipper et al. 2008). Second, low correlation scores on the adjective and adverb subsets in some languages (e.g., POL, CYM, SWA) might be due to their low frequency in monolingual texts, which yields unreliable representations. In general, the variance in performance across different word classes warrants further research in class-specific representation learning (Baker, Reichart, and Korhonen 2014; Vulić et al. 2017b). The scores further attest the usefulness of unsupervised post-processing as almost all class-specific correlation scores are improved by applying mean-centering and ABTT. Finally, the results for M-BERT and XLM-100 in Table 13 further confirm that massively multilingual pretraining cannot yield reasonable semantic representations for many languages: in fact, for some classes they display no correlation with human ratings at all.

Differences across Languages. Naturally, the results from Tables 12 and 13 also reveal that there is variation in performance of both static word embeddings and pretrained encoders across different languages. Among other causes, the lowest absolute scores with FT are reported for languages with least resources available to train monolingual word embeddings, such as Kiswahili, Welsh, and Estonian. The low performance on Welsh is especially indicative: Figure 1 shows that the ratings in the Welsh dataset match up very well with the English ratings, but we cannot achieve the same level of correlation in Welsh with Welsh FT word embeddings. Difference in performance between two closely related languages, EST (low-resource) and FIN (high-resource), provides additional evidence in this respect.

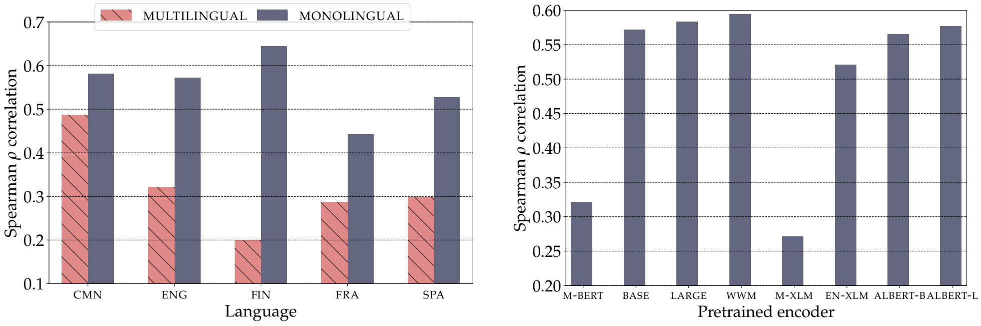
The highest reported scores with M-BERT and XLM-100 are obtained for Mandarin Chinese and Yue Chinese: this effectively points to the weaknesses of massively multilingual training with a joint subword vocabulary spanning 102 and 100 languages. Due to the difference in scripts, "language-specific" subwords for YUE and CMN do not need to be shared across a vast amount of languages and the quality of their representation remains unscathed. This effectively means that M-BERT's subword vocabulary contains plenty of CMN-specific and YUE-specific subwords which are exploited by the encoder when producing M-BERT-based representations. Simultaneously, higher scores with M-BERT (and XLM in Table 13) are reported for resource-rich languages such as French, Spanish, and English, which are better represented in M-BERT's training data. We also observe lower absolute scores (and a larger number of OOVs) for languages with very rich and productive morphological systems such as the two Slavic languages (Polish and Russian) and Finnish. Since Polish and Russian are known to have large Wikipedias and Common Crawl data (Conneau et al. 2019) (e.g., their Wikipedias are in the top 10 largest Wikipedias worldwide), the problem with coverage can be attributed exactly to the proliferation of morphological forms in those languages.

Finally, while Table 12 does reveal that unsupervised post-processing is useful for all languages, it also demonstrates that peak scores are achieved with different post-processing configurations. This finding suggests that a more careful language-specific fine-tuning is indeed needed to refine word embeddings towards semantic similarity. We plan to inspect the relationship between post-processing techniques and linguistic properties in more depth in future work.

Multilingual vs. Language-Specific Contextualized Embeddings. Recent work has shown that—despite the usefulness of massively multilingual models such as M-BERT and XLM-100 for zero-shot cross-lingual transfer (Pires, Schlinger, and Garrette 2019; Wu and Dredze 2019)—stronger results in downstream tasks for a particular language can be achieved by pretraining language-specific models on language-specific data.

In this experiment, motivated by the low results of M-BERT and XLM-100 (see again Table 13), we assess if monolingual pretrained encoders can produce higher-quality word-level representations than multilingual models. Therefore, we evaluate language-specific BERT and XLM models for a subset of the Multi-SimLex languages for which such models are currently available: Finnish (Virtanen et al. 2019) (BERT-BASE architecture, uncased), French (Le et al. 2019) (the FlauBERT model based on XLM), English (BERT-BASE, uncased), Mandarin Chinese (BERT-BASE) (Devlin et al. 2019) and Spanish (BERT-BASE, uncased). In addition, we also evaluate a series of pretrained encoders available for English: (i) BERT-BASE, BERT-LARGE, and BERT-LARGE with whole word masking (WWM) from the original work on BERT (Devlin et al. 2019), (ii) monolingual "English-specific" XLM (Conneau and Lample 2019), and (iii) two models which employ parameter reduction techniques to build more compact encoders: ALBERT-B uses a configuration similar to BERT-BASE, while ALBERT-L is similar to BERT-LARGE, but with an 18× reduction in the number of parameters (Lan et al. 2020).[16]

[16] All models and their further specifications are available at the following link: https://huggingface.co/models.



(a) Monolingual vs multilingual

(b) Pretrained ENG encoders

Figure 5: **(a)** A performance comparison between monolingual pretrained language encoders and massively multilingual encoders. For four languages (CMN, ENG, FIN, SPA), we report the scores with monolingual uncased BERT-BASE architectures and multilingual uncased M-BERT model, while for FRA we report the results of the multilingual XLM-100 architecture and a monolingual French FlauBERT model (Le et al. 2019), which is based on the same architecture as XLM-100. **(b)** A comparison of various pretrained encoders available for English. All these models are post-processed via ABTT (-3).

From the results in Table 5, it is clear that monolingual pretrained encoders yield much more reliable word-level representations. The gains are visible even for languages such as CMN which showed reasonable performance with M-BERT and are substantial on all test languages. This further confirms the validity of language-specific pretraining in lieu of multilingual training, if sufficient monolingual data are available. Moreover, a comparison of pretrained English encoders in Figure 5b largely follows the intuition: the larger BERT-LARGE model yields slight improvements over BERT-BASE, and we can improve a bit more by relying on word-level (i.e., lexical-level) masking.Finally, light-weight ALBERT model variants are quite competitive with the original BERT models, with only modest drops reported, and ALBERT-L again outperforms ALBERT-B. Overall, it is interesting to note that the scores obtained with monolingual pretrained encoders are on a par with or even outperform static FT word embeddings: this is a very intriguing finding per se as it shows that such subword-level models trained on large corpora can implicitly capture rich lexical semantic knowledge.

Similarity-Specialized Word Embeddings. Conflating distinct lexico-semantic relations is a well-known property of distributional representations (Turney and Pantel 2010; Melamud et al. 2016). Semantic specialization fine-tunes distributional spaces to emphasize a particular lexico-semantic relation in the transformed space by injecting external lexical knowledge (Glavaš, Ponti, and Vulić 2019). Explicitly discerning between true semantic similarity (as captured in Multi-SimLex) and broad conceptual relatedness benefits a number of tasks, as discussed in §2.1.[17] Since most languages lack dedicated lexical resources, however, one viable strategy to steer monolingual word vector spaces to emphasize semantic similarity is through cross-lingual transfer of lexical knowledge, usually through a shared cross-lingual word vector space (Ruder, Vulić, and Søgaard

[17] For an overview of specialization methods for semantic similarity, we refer the interested reader to the recent tutorial (Glavaš, Ponti, and Vulić 2019).

2019). Therefore, we evaluate the effectiveness of specialization transfer methods using Multi-SimLex as our multilingual test bed.

We evaluate a current state-of-the-art cross-lingual specialization transfer method with minimal requirements, put forth recently by Ponti et al. (2019c).[18] In a nutshell, their LI-POSTSPEC method is a multi-step procedure that operates as follows. First, the knowledge about semantic similarity is extracted from WordNet in the form of triplets, that is, linguistic constraints (w_1, w_2, r), where w_1 and w_2 are two concepts, and r is a relation between them obtained from WordNet (e.g., synonymy or antonymy). The goal is to "attract" synonyms closer to each other in the transformed vector space as they reflect true semantic similarity, and "repel" antonyms further apart. In the second step, the linguistic constraints are translated from English to the target language via a shared cross-lingual word vector space. To this end, following Ponti et al. (2019c) we rely on cross-lingual word embeddings (CLWEs) (Joulin et al. 2018) available online, which are based on Wiki FT vectors.[19] Following that, a constraint refinement step is applied in the target language which aims to eliminate the noise inserted during the translation process. This is done by training a relation classification tool: it is trained again on the English linguistic constraints and then used on the translated target language constraints, where the transfer is again enabled via a shared cross-lingual word vector space.[20] Finally, a state-of-the-art monolingual specialization procedure from Ponti et al. (2018b) injects the (now target language) linguistic constraints into the target language distributional space.

The scores are summarized in Table 14. Semantic specialization with LI-POSTSPEC leads to substantial improvements in correlation scores for the majority of the target languages, demonstrating the importance of external semantic similarity knowledge for semantic similarity reasoning. However, we also observe deteriorated performance for the three target languages which can be considered the lowest-resource ones in our set: CYM, SWA, YUE. We hypothesize that this occurs due to the inferior quality of the underlying monolingual Wikipedia word embeddings, which generates a chain of error accumulations. In particular, poor distributional word estimates compromise the alignment of the embedding spaces, which in turn results in increased translation noise, and reduced refinement ability of the relation classifier. On a high level, this "poor get poorer" observation again points to the fact that one of the primary causes of low performance of resource-low languages in semantic tasks is the sheer lack of even unlabeled data for distributional training. On the other hand, as we see from Table 13, typological dissimilarity between the source and the target does not deteriorate the effectiveness of semantic specialization. In fact, LI-POSTSPEC does yield substantial gains also for the typologically distant targets such as HEB, CMN, and EST. The critical problem indeed seems to be insufficient raw data for monolingual distributional training.

8. Cross-Lingual Evaluation

Similar to monolingual evaluation in §7, we now evaluate several state-of-the-art cross-lingual representation models on the suite of 66 automatically constructed cross-lingual Multi-SimLex datasets. Again, note that evaluating a full range of cross-lingual models

[18] We have also evaluated other specialization transfer methods, e.g., (Glavaš and Vulić 2018; Ponti et al. 2018b), but they are consistently outperformed by the method of Ponti et al. (2019c).

[19] `https://fasttext.cc/docs/en/aligned-vectors.html`; for target languages for which there are no pretrained CLWEs, we induce them following the same procedure of Joulin et al. (2018).

[20] We again follow Ponti et al. (2019c) and use a state-of-the-art relation classifier (Glavaš and Vulić 2018). We refer the reader to the original work for additional technical details related to the classifier design.

Languages:	CMN	CYM	ENG	EST	FIN	FRA	HEB	POL	RUS	SPA	SWA	YUE
FASTTEXT (Wiki)	*(429)*	*(282)*	*(6)*	*(343)*	*(345)*	*(73)*	*(62)*	*(354)*	*(343)*	*(57)*	*(379)*	*(677)*
FT:INIT	.315	**.318**	–	.400	.575	.444	.428	.370	.359	.432	**.332**	**.376**
LI-POSTSPEC	**.584**	.204	–	**.515**	**.619**	**.601**	**.510**	**.531**	**.547**	**.635**	.238	.267

Table 14: The impact of vector space specialization for semantic similarity. The scores are reported using the current state-of-the-art specialization transfer LI-POSTSPEC method of Ponti et al. (2019c), relying on English as a resource-rich source language and the external lexical semantic knowledge from the English WordNet.

available in the rich prior work on cross-lingual representation learning is well beyond the scope of this article. We therefore focus our cross-lingual analyses on several well-established and indicative state-of-the-art cross-lingual models, again spanning both static and contextualized cross-lingual word embeddings.

8.1 Models in Comparison

Static Word Embeddings. We rely on a state-of-the-art mapping-based method for the induction of cross-lingual word embeddings (CLWEs): VECMAP (Artetxe, Labaka, and Agirre 2018b). The core idea behind such mapping-based or projection-based approaches is to learn a post-hoc alignment of independently trained monolingual word embeddings (Ruder, Vulić, and Søgaard 2019). Such methods have gained popularity due to their conceptual simplicity and competitive performance coupled with reduced bilingual supervision requirements: they support CLWE induction with only as much as a few thousand word translation pairs as the bilingual supervision (Mikolov, Le, and Sutskever 2013; Xing et al. 2015; Upadhyay et al. 2016; Ruder, Søgaard, and Vulić 2019). More recent work has shown that CLWEs can be induced with even weaker supervision from small dictionaries spanning several hundred pairs (Vulić and Korhonen 2016; Vulić et al. 2019), identical strings (Smith et al. 2017), or even only shared numerals (Artetxe, Labaka, and Agirre 2017). In the extreme, *fully unsupervised* projection-based CLWEs extract such seed bilingual lexicons from scratch on the basis of monolingual data only (Conneau et al. 2018a; Artetxe, Labaka, and Agirre 2018b; Hoshen and Wolf 2018; Alvarez-Melis and Jaakkola 2018; Chen and Cardie 2018; Mohiuddin and Joty 2019, *inter alia*).

Recent empirical studies (Glavaš et al. 2019; Vulić et al. 2019; Doval et al. 2019) have compared a variety of unsupervised and weakly supervised mapping-based CLWE methods, and VECMAP emerged as the most robust and very competitive choice. Therefore, we focus on 1) its fully unsupervised variant (UNSUPER) in our comparisons. For several language pairs, we also report scores with two other VECMAP model variants: 2) a supervised variant which learns a mapping based on an available seed lexicon (SUPER), and 3) a supervised variant *with self-learning* (SUPER+SL) which iteratively increases the seed lexicon and improves the mapping gradually. For a detailed description of these variants, we refer the reader to recent work (Artetxe, Labaka, and Agirre 2018b; Vulić et al. 2019). We again use CC+Wiki FT vectors as initial monolingual word vectors, except for YUE where Wiki FT is used. The seed dictionaries of two different sizes (1k and 5k translation pairs) are based on PanLex (Kamholz, Pool, and Colowick 2014), and are

	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.

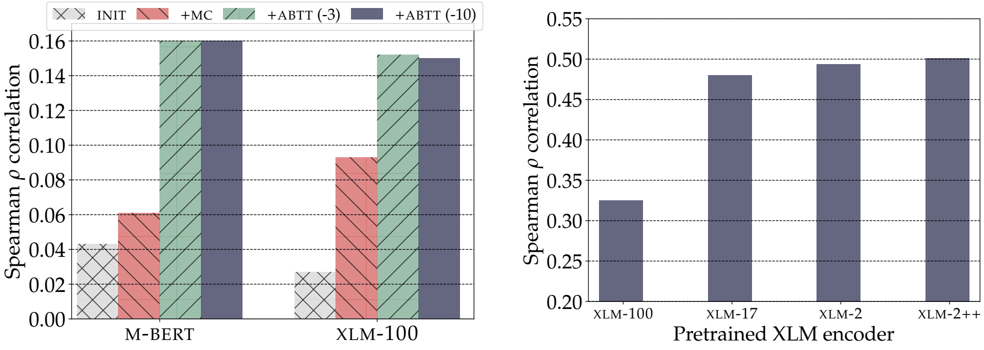
As in monolingual settings, we can apply unsupervised post-processing steps such as ABTT to both static and contextualized cross-lingual word embeddings.

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



(a) Average scores (b) Scores on ENG-FRA

Figure 6: Further performance analyses of cross-lingual Multi-SimLex datasets. (a) Spearman's ρ correlation scores averaged over all 66 cross-lingual Multi-SimLex datasets for two pretrained multilingual encoders (M-BERT and XLM). The scores are obtained with different configurations that exclude (INIT) or enable unsupervised post-processing. (b) A comparison of various pretrained encoders available for the English-French language pair, see the main text for a short description of each benchmarked pretrained encoder.

of Table 15): many correlation scores are in fact no-correlation results, accentuating the problem of fully unsupervised cross-lingual learning for typologically diverse languages and with fewer amounts of monolingual data (Vulić et al. 2019). The scores are particularly low across the board for lower-resource languages such as Welsh and Kiswahili. It also seems that the lack of monolingual data is a larger problem than typological dissimilarity between language pairs, as we do observe reasonably high correlation scores with VECMAP for language pairs such as CMN-SPA, HEB-EST, and RUS-FIN. However, typological differences (e.g., morphological richness) still play an important role as we observe very low scores when pairing CMN with morphologically rich languages such FIN, EST, POL, and RUS. Similar to prior work of Vulić et al. (2019) and Doval et al. (2019), given the fact that unsupervised VECMAP is the most robust unsupervised CLWE method at present (Glavaš et al. 2019), our results again question the usefulness of fully unsupervised approaches for a large number of languages, and call for further developments in the area of unsupervised and weakly supervised cross-lingual representation learning.

The scores of M-BERT and XLM-100[22] lead to similar conclusions as in the monolingual settings. Reasonable correlation scores are achieved only for a small subset of resource-rich language pairs (e.g., ENG, FRA, SPA, CMN) which dominate the multilingual M-BERT training. Interestingly, the scores indicate a much higher performance of language pairs where YUE is one of the languages when we use M-BERT instead of VECMAP. This boils down again to the fact that YUE, due to its specific language script, has a good representation of its words and subwords in the shared M-BERT vocabulary. At the same time, a reliable VECMAP mapping between YUE and other languages cannot be found due to a small monolingual YUE corpus. In cases when VECMAP does not yield a degenerate

[22] The XLM-100 scores are not reported for brevity; they largely follow the patterns observed with M-BERT. The aggregated scores between the two encoders are also very similar as indicated by Figure 6a.

cross-lingual vector space starting from two monolingual ones, the final correlation scores seem substantially higher than the ones obtained by the single massively multilingual M-BERT model.

Finally, the results in Figure 6a again verify the usefulness of unsupervised post-processing also in cross-lingual settings. We observe improved performance with both M-BERT and XLM-100 when mean centering (+MC) is applied, and further gains can be achieved by using ABTT on the mean-centered vector spaces. A similar finding also holds for static cross-lingual word embeddings[23], where applying ABBT (-10) yields higher scores on 61/66 language pairs.

Fully Unsupervised vs. Weakly Supervised Cross-Lingual Embeddings. The results in Table 15 indicate that fully unsupervised cross-lingual learning fails for a large number of language pairs. However, recent work (Vulić et al. 2019) has noted that these sub-optimal non-alignment solutions with the UNSUPER model can be avoided by relying on (weak) cross-lingual supervision spanning only several thousands or even hundreds of word translation pairs. Therefore, we examine 1) if we can further improve the results on cross-lingual Multi-SimLex resorting to (at least some) cross-lingual supervision for resource-rich language pairs; and 2) if such available word-level supervision can also be useful for a range of languages which displayed near-zero performance in Table 15. In other words, we test if recent "tricks of the trade" used in the rich literature on CLWE learning reflect in gains on cross-lingual Multi-SimLex datasets.

First, we reassess the findings established on the bilingual lexicon induction task (Søgaard, Ruder, and Vulić 2018; Vulić et al. 2019): using at least some cross-lingual supervision is always beneficial compared to using no supervision at all. We report improvements over the UNSUPER model for all 10 language pairs in Table 16, even though the UNSUPER method initially produced strong correlation scores. The importance of self-learning increases with decreasing available seed dictionary size, and the +SL model always outperforms UNSUPER with 1k seed pairs; we observe the same patterns also with even smaller dictionary sizes than reported in Table 16 (250 and 500 seed pairs). Along the same line, the results in Table 17 indicate that at least some supervision is crucial for the success of static CLWEs on resource-leaner language pairs. We note substantial improvements on all language pairs; in fact, the VECMAP model is able to learn a more reliable mapping starting from clean supervision. We again note large gains with self-learning.

Multilingual vs. Bilingual Contextualized Embeddings. Similar to the monolingual settings, we also inspect if massively multilingual training in fact dilutes the knowledge necessary for cross-lingual reasoning on a particular language pair. Therefore, we compare the 100-language XLM-100 model with i) a variant of the same model trained on a smaller set of 17 languages (XLM-17); ii) a variant of the same model trained specifically for the particular language pair (XLM-2); and iii) a variant of the bilingual XLM-2 model that also leverages bilingual knowledge from parallel data during joint training (XLM-2++). We again use the pretrained models made available by Conneau and Lample (2019), and we refer to the original work for further technical details.

The results are summarized in Figure 6b, and they confirm the intuition that massively multilingual pretraining can damage performance even on resource-rich languages and language pairs. We observe a steep rise in performance when the

[23] Note that VECMAP does mean centering by default as one of its preprocessing steps prior to learning the mapping function (Artetxe, Labaka, and Agirre 2018b; Vulić et al. 2019).

	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 additional languages. This can help and create a hugely valuable, large-scale semantic resource for multilingual NLP research.

The core Multi-SimLex we release with this paper already enables researchers to carry out novel linguistic analysis as well as establishes a benchmark for evaluating representation learning models. Based on our preliminary analyses, we found that speakers of closely related languages tend to express equivalent similarity judgments. In particular, geographical proximity seems to play a greater role than family membership in determining the similarity of judgments across languages. Moreover, we tested several state-of-the-art word embedding models, both static and contextualized representations, as well as several (supervised and unsupervised) post-processing techniques, on the newly released Multi-SimLex. This enables future endeavors to improve multilingual representation learning with challenging baselines. In addition, our results provide several important insights for research on both monolingual and cross-lingual word representations:

1) Unsupervised post-processing techniques (mean centering, elimination of top principal components, adjusting similarity orders) are always beneficial independently of the language, although the combination leading to the best scores is language-specific and hence needs to be tuned.

2) Similarity rankings obtained from word embeddings for nouns are better aligned with human judgments than all the other part-of-speech classes considered here (verbs, adjectives, and, for the first time, adverbs). This confirms previous generalizations based on experiments on English.

3) The factor having the greatest impact on the quality of word representations is the availability of raw texts to train them in the first place, rather than language properties (such as family, geographical area, typological features).

4) Massively multilingual pretrained encoders such as M-BERT (Devlin et al. 2019) and XLM-100 (Conneau and Lample 2019) fare quite poorly on our benchmark, whereas pretrained encoders dedicated to a single language are more competitive with static word embeddings such as fastText (Bojanowski et al. 2017). Moreover, for language-specific encoders, parameter reduction techniques reduce performance only marginally.

5) Techniques to inject clean lexical semantic knowledge from external resources into distributional word representations were proven to be effective in emphasizing the relation of semantic similarity. In particular, methods capable of transferring such knowledge from resource-rich to resource-lean languages (Ponti et al. 2019c) increased the correlation with human judgments for most languages, except for those with limited unlabelled data.

Future work can expand our preliminary, yet large-scale study on the ability of pretrained encoders to reason over word-level semantic similarity in different languages. For instance, we have highlighted how sharing the same encoder parameters across multiple languages may harm performance. However, it remains unclear if, and to what extent, the input language embeddings present in XLM-100 but absent in M-BERT help mitigate this issue. In addition, pretrained language embeddings can be obtained both from typological databases (Littell et al. 2017) and from neural architectures (Malaviya, Neubig, and Littell 2017). Plugging these embeddings into the encoders in lieu of embeddings trained end-to-end as suggested by prior work (Tsvetkov et al. 2016; Ammar et al. 2016; Ponti et al. 2019b) might extend the coverage to more resource-lean languages.

Another important follow-up analysis might involve the comparison of the performance of representation learning models on multilingual datasets for both word-level semantic similarity and sentence-level Natural Language Understanding. In particular, Multi-SimLex fills a gap in available resources for multilingual NLP and might help understand how lexical and compositional semantics interact if put alongside existing resources such as XNLI (Conneau et al. 2018b) for natural language inference or PAWS-X (Yang et al. 2019) for cross-lingual paraphrase identification. Finally, the Multi-SimLex annotation could turn out to be a unique source of evidence to study the effects of polysemy in human judgments on semantic similarity: for equivalent word pairs in multiple languages, are the similarity scores affected by how many senses the two words (or multi-word expressions) incorporate?

In light of the success of initiatives like Universal Dependencies for multilingual treebanks, we hope that making Multi-SimLex and its guidelines available will encourage other researchers to expand our current sample of languages. We particularly encourage creation and submission of comparable Multi-SimLex datasets for under-resourced and typologically diverse languages in future work. In particular, we have made a Multi-Simlex community website available to facilitate easy creation, gathering, dissemination, and use of annotated datasets: `https://multisimlex.com/`.

Acknowledgments

This work is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909). Thierry Poibeau is partly supported by a PRAIRIE 3IA Institute fellowship ("Investissements d’avenir" program, reference ANR-19-P3IA-0001).

References

Adams, Oliver, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pages 937–947.

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalová, 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.

Aldarmaki, Hanan and Mona Diab. 2019. Context-aware cross-lingual mapping. In *Proceedings of NAACL-HLT*, pages 3906–3911.

Alvarez-Melis, David and Tommi Jaakkola. 2018. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of EMNLP*, pages 1881–1890.

Ammar, Waleed, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the ACL*, 4:431–444.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*, pages 451–462.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of AAAI*, pages 5012–5019.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

Artetxe, Mikel, Gorka Labaka, Iñigo Lopez-Gazpio, and Eneko Agirre. 2018. Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. In *Proceedings of CoNLL*, pages 282–291.

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of ACL*, pages 86–90.

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

Bapna, Ankur and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of EMNLP*, pages 1538–1548.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The

WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Barzegar, Siamak, Brian Davis, Manel Zarrouk, Siegfried Handschuh, and André Freitas. 2018. SemR-11: A multi-lingual gold-standard for semantic similarity and relatedness for eleven languages. In *Proceedings of LREC*, pages 3912–3916.

van den Berg, Robert A., Huub C.J. Hoefsloot, Johan A. Westerhuis, Age K. Smilde, and Mariët J. van der Werf. 2006. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7(1):142.

Bjerva, Johannes and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of NAACL-HLT*, pages 907–916.

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

Bro, Rasmus and Age K. Smilde. 2003. Centering and scaling in component analysis. *Journal of Chemometrics*, 17(1):16–33.

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

Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Camacho-Collados, Jose and Roberto Navigli. 2017. BabelDomains: Large-scale domain labeling of lexical resources. In *Proceedings of EACL*, pages 223–228.

Camacho-Collados, Jose, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of SEMEVAL*, pages 15–26.

Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of ACL*, pages 1–7.

Cao, Steven, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *Proceedings of ICLR*.

Chen, Danqi and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*, pages 740–750.

Chen, Xilun and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of EMNLP*, pages 261–270.

Cimiano, Philipp, Andreas Hotho, and Steffen Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24:305–339.

Clark, Jonathan H., Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the ACL*.

Collobert, Ronan and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, pages 160–167.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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

Conneau, Alexis, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *Proceedings of ICLR*.

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.

Coseriu, Eugenio. 1967. Lexikalische solidaritäten. *Poetica*, 1:293–303.

Cruse, David Alan. 1986. *Lexical Semantics*. Cambridge University Press.

De Deyne, Simon and Gert Storms. 2008. Word associations: Network and semantic properties. *Behavior Research Methods*, 40(1):213–231.

Devlin, Jacob, 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.

Doitch, Amichay, Ram Yazdi, Tamir Hazan, and Roi Reichart. 2019. Perturbation based learning for structured NLP tasks with application to dependency parsing. *Transactions of the ACL*, 7:643–659.

Doval, Yerai, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2018. Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of EMNLP*, pages 294–304.

Doval, Yerai, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2019. On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. *CoRR*, abs/1908.07742.

Dryer, Matthew S. 2013. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ercan, Gökhan and Olcay Taner Yıldız. 2018. AnlamVer: Semantic model evaluation dataset for Turkish - Word similarity and relatedness. In *Proceedings of COLING*, pages 3819–3836.

Ethayarajh, Kawin. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP*, pages 55–65.

Evans, Nicholas. 2011. Semantic Typology. In *The Oxford Handbook of Linguistic Typology*. Oxford University Press, pages 504–533.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pages 1606–1615.

Fellbaum, Christiane. 1998. *WordNet*. MIT Press.

Finkelstein, Lev, 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.

Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

François, Alexandre. 2008. Semantic maps and the typology of colexification. *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 106:163.

Gerz, Daniela, 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.

Glavaš, Goran, Edoardo Maria Ponti, and Ivan Vulić. 2019. Semantic specialization of distributional word vectors. In *Proceedings of EMNLP: Tutorial Abstracts*.

Glavaš, Goran and Ivan Vulić. 2018. Discriminating between lexico-semantic relations with the specialization tensor model. In *Proceedings of NAACL-HLT*, pages 181–187.

Glavaš, Goran and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of ACL*, pages 34–45.

Glavaš, Goran, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL*, pages 710–721.

Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of LREC*, pages 3483–3487.

Gruber, Jeffrey. 1976. *Lexical Structures in Syntax and Semantics*, volume 25. North-Holland.

Harris, Zellig S. 1951. *Methods in Structural Linguistics*. University of Chicago Press.

Hill, Felix, KyungHyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the ACL*, 4:17–30.

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

Hoshen, Yedid and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of EMNLP*, pages 469–478.

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

Joulin, Armand, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP*, pages 2979–2984.

Kamath, Aishwarya, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP*, pages 72–83.

Kamholz, David, Jonathan Pool, and Susan M. Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of LREC*, pages 3145–3150.

Kay, Paul and Luisa Maffi. 2013. Green and blue. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kiela, Douwe and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 21–30.

Kiela, Douwe, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of EMNLP*, pages 2044–2048.

Kim, Joo-Kyung, Marie-Catherine de Marneffe, and Eric Fosler-Lussier. 2016. Adjusting word embeddings with semantic intensity orders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 62–69.

Kim, Joo-Kyung, Gokhan Tur, Asli Celikyilmaz, Bin Cao, and Ye-Yi Wang. 2016. Intent detection using semantically enriched word embeddings. In *SLT*.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Kipper, Karin, Benjamin Snyder, and Martha Palmer. 2004. Extending a verb-lexicon using a semantically annotated corpus. In *Proceedings of LREC*.

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

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

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of ICLR*.

Lauscher, Anne, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2019. Informing unsupervised pretraining with external linguistic knowledge. *arXiv preprint arXiv:1909.02339*.

Lazaridou, Angeliki, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of ACL*, pages 270–280.

Le, Hang, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. FlauBERT: Unsupervised language model pre-training for french. *CoRR*, abs/1912.05372.

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

Levin, Beth. 1993. *English verb classes and alternation, A preliminary investigation*. The University of Chicago Press.

Levy, Omer and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308.

Lewis, Patrick S. H., Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: Evaluating cross-lingual extractive question answering. *CoRR*, abs/1910.07475.

Lin, Yu-Hsiang, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL*, pages 3125–3135.

Littell, Patrick, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of EACL*, pages 8–14.

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

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lucas, Margery. 2000. Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, 7(4):618–630.

Luong, Thang, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, pages 104–113.

Lyons, John. 1977. *Semantics*, volume 2. Cambridge University Press.

Majid, Asifa, Melissa Bowerman, Miriam van Staden, and James S Boster. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics*, 18(2):133–152.

Malaviya, Chaitanya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of EMNLP*, pages 2529–2535.

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

McKeown, Kathleen R., Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbia's newsblaster. In *Proceedings of HLT*, page 280–285.

Melamud, Oren, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of NAACL-HLT*, pages 1030–1040.

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

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint, CoRR*, abs/1309.4168.

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

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM*, pages 39–41.

Mimno, David and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of EMNLP*, pages 2873–2878.

Mohiuddin, Tasnim and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In *Proceedings of NAACL-HLT*, pages 3857–3867.

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, pages 142–148.

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

Mu, Jiaqi, Suma Bhat, and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *Proceedings of ICLR*.

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

Nelson, Douglas L., Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, 36(3):402–407.

Netisopakul, Ponrudee, Gerhard Wohlgenannt, and Aleksei Pulich. 2019. Word similarity datasets for Thai: Construction and evaluation. *CoRR*, abs/1904.04307.

Nivre, Joakim, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, et al. 2019. Universal Dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Pearson, Karl. 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Pilehvar, Mohammad Taher, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge rare word dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of EMNLP*, pages 1391–1401.

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

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019a. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Ponti, Edoardo Maria, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018a. Isomorphic transfer of syntactic structures in cross-lingual nlp. In *Proceedings of ACL*, pages 1531–1542.

Ponti, Edoardo Maria, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019b. Towards zero-shot language modeling. In *Proceedings of EMNLP-IJCNLP*, pages 2893–2903.

Ponti, Edoardo Maria, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018b. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP*, pages 282–293.

Ponti, Edoardo Maria, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019c. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of EMNLP*, pages 2206–2217.

Radovanović, Miloš, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531.

Rasooli, Mohammad Sadegh and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *Transactions of the ACL*, 5:279–293.

Ren, Liliang, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of EMNLP*, pages 2780–2786.

Rotman, Guy and Roi Reichart. 2019. Deep contextualized self-training for low resource dependency parsing. *Transactions of the ACL*, 7:695–713.

Ruder, Sebastian, Anders Søgaard, and Ivan Vulić. 2019. Unsupervised cross-lingual representation learning. In *Proceedings of ACL: Tutorial Abstracts*, pages 31–38.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Rzymski, Christoph, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(1):1–12.

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

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of ACL*, pages 732–746.

Schuster, Mike and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152.

Schwartz, Roy, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267.

Schwartz, Roy, Roi Reichart, and Ari Rappoport. 2016. Symmetric patterns and coordinations: Fast and enhanced representations of verbs and adjectives. In *Proceedings of NAACL-HLT*, pages 499–505.

Smith, Samuel L., David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR (Conference Track)*.

Snyder, Benjamin and Regina Barzilay. 2010. Climbing the tower of Babel: Unsupervised multilingual learning. In *Proceedings of ICML*, pages 29–36.

Søgaard, Anders, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL*, pages 778–788.

Suzuki, Ikumi, Kazuo Hara, Masashi Shimbo, Marco Saerens, and Kenji Fukumizu. 2013. Centering similarity measures to reduce hubs. In *Proceedings of EMNLP*, pages 613–623.

Tang, Shuai, Mahta Mousavi, and Virginia R. de Sa. 2019. An empirical study on post-processing methods for word embeddings. *CoRR*, abs/1905.10971.

Trier, Jost. 1931. *Der Deutsche Wortschatz im Sinnbezirk des Verstandes: Die Geschichte eines sprachlichen Feldes. 1. Von den Anfängen bis zum Beginn des 13. Jahrhunderts*. Ph.D. thesis, University of Bonn.

Tsvetkov, Yulia, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W. Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of NAACL-HLT*, pages 1357–1366.

Turney, Peter D. 2012. Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, 44:533–585.

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artifical Intelligence Research*, 37(1):141–188.

Upadhyay, Shyam, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of ACL*, pages 1661–1670.

Vania, Clara and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of ACL*, pages 2016–2027.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 6000–6010.

Venekoski, Viljami and Jouko Vankka. 2017. Finnish resources for evaluating language model semantics. In *Proceedings of NODALIDA*, pages 231–236.

Virtanen, Antti, 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.

Vulić, Ivan, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017a. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Vulić, Ivan, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP*, pages 4407–4418.

Vulić, Ivan, Douwe Kiela, and Anna Korhonen. 2017. Evaluation by association: A systematic study of quantitative word association evaluation. In *Proceedings of EACL*, pages 163–175.

Vulić, Ivan and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of ACL*, pages 247–257.

Vulić, Ivan, Simone Paolo Ponzetto, and Goran Glavaš. 2019. Multilingual and cross-lingual graded lexical entailment. In *Proceedings of ACL*, pages 4963–4974.

Vulić, Ivan, Roy Schwartz, Ari Rappoport, Roi Reichart, and Anna Korhonen. 2017b. Automatic selection of context configurations for improved class-specific word representations. In *Proceedings of CoNLL*, pages 112–122.

Wang, Zihan, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *Proceedings of ICLR*.

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358.

Williams, Adina, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*, pages 1112–1122.

Wolf, Thomas, 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.

Wu, Shijie, Alexis Conneau, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Emerging cross-lingual structure in pretrained language models. *CoRR*, abs/1911.01464.

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

Wu, Zhibiao and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of ACL*, pages 133–138.

Xing, Chao, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL-HLT*, pages 1006–1011.

Yang, Yinfei, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP*, pages 3687–3692.

Zeman, Daniel, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.

Zhang, Mozhi, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. 2019. Are girls neko or shōjo? Cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of ACL*, pages 3180–3189.

Zhang, Yuan, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of NAACL-HLT*, pages 1298–1308.

Zhu, Yi, Benjamin Heinzerling, Ivan Vulić, Michael Strube, Roi Reichart, and Anna Korhonen. 2019. On the importance of subword information for morphological tasks in truly low-resource languages. In *Proceedings of CoNLL*, pages 216–226.

Zhu, Yi, Ivan Vulić, and Anna Korhonen. 2019. A systematic study of leveraging subword information for learning word representations. In *Proceedings of NAACL-HLT*, pages 912–932.