Syn2Vec: Synset Colexification Graphs for Lexical Semantic Similarity

John Harvill[1], **Roxana Girju**[2], **Mark Hasegawa-Johnson**[1]
[1]Department of Electrical and Computer Engineering,
[2]Departments of Linguistics and Computer Science,
University of Illinois at Urbana-Champaign
{harvill2, girju, jhasegaw}@illinois.edu

Abstract

In this paper we focus on patterns of colexification (co-expressions of form–meaning mapping in the lexicon) as an aspect of lexical-semantic organization, and use them to build large scale synset graphs across BabelNet's typologically diverse set of 499 world languages. We introduce and compare several approaches: *monolingual* and *cross-lingual colexification graphs*, popular *distributional models*, and *fusion approaches*. The models are evaluated against human judgments on a semantic similarity task for nine languages. Our strong empirical findings also point to the importance of universality of our graph synset embedding representations with no need for any language-specific adaptation when evaluated on the lexical similarity task. The insights of our exploratory investigation of large-scale colexification graphs could inspire significant advances in NLP across languages, especially for tasks involving languages which lack dedicated lexical resources, and can benefit from language transfer from large shared cross-lingual semantic spaces.

1 Introduction

Distributional models like word embeddings have been widely used in Natural Language Processing (NLP) (Iacobacci et al., 2016; Devlin et al., 2014; Hewlett et al., 2016). They operate under the assumption that words appearing in similar contexts have similar meanings, and thus close representations. However, as do other unsupervised learning models, they suffer from classic limitations – i.e., there is no guarantee that all context words contribute to the meaning of the target word, while, in fact, it is possible that some low frequency words, with poorly-trained embeddings, are highly semantically connected. Also, they don't distinguish between topically related words and near synonyms.

Dictionaries and thesauruses, on the other hand, have traditionally offered an alternative approach, through their discrete lists of fine-grained senses, textual definitions, and relationships with other senses. Given a sufficiently large dictionary of many fine-grained sense representations in many of the world's languages, one could perform sophisticated semantic tasks on word senses (Conia and Navigli, 2020). In fact, investigating universal and areal cross-linguistic variations in the lexicon has been the focus of lexical typology. One increasingly popular empirical method of investigating senses based on cross-linguistic comparison in typological studies has been that of colexification patterns. "A given language is said to colexify two functionally distinct senses if, and only if, it can associate them with the same lexical form" (François, 2008), reflecting natural semantic connections (Haspelmath, 2003). For example, 'fire' and 'firewood' are colexified in Kamoro (New Guinea language) as 'uta' and in Wayuu (Arawakan language) as 'siki', but receive distinct lexemes in English and Romanian. Each polysemous lexeme as a whole is language-specific, yet a great number of lexical polysemies are each attested across multiple languages.

In this paper, we investigate semantic structures in the lexicon as manifested by colexification patterns in a large number of languages and assess, at a large scale, their usefulness to the Lexical Semantic Similarity (LSIM) task (Vulić et al., 2020a). We used one of the largest digital lexical resources to date, BabelNet 5.0 (953.4M (concept) lexicalizations in 499 languages[1]) to build and process colexification graphs.

Specifically, we make the following contributions: **1)** Propose a simple, yet effective algorithm to automatically construct large-scale synset similarity graphs based on the principle of colexification, and use the graphs to generate high-quality

[1]BabelNet 5.0's claim of supporting 500 languages seems to be a typo. There are only 499 (see "Languages and Coverage" tab) https://babelnet.org/statistics

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5259 - 5270

synset and word representations. **2)** Demonstrate the importance of universality of our graph synset embedding representations with no need for any language-specific adaptation when evaluated on LSIM. **3)** Show that our proposed approach significantly outperforms state-of-the-art synset and word embedding techniques on the LSIM task. **4)** When combined with knowledge-based approaches like our cross-lingual colexification patterns, purely unsupervised distributional models like fastText and BERT result in better alignment with human perception, as measured on the LSIM task.

Our findings and models contribute to advances in computational modeling of natural language understanding across languages, and offer new insights into linguistic typology.

2 Related Work

The concept of colexification has been introduced by Haspelmath (2003) to distinguish senses in the grammatical domain, but has been formalized for the field of lexicon by François (2008) who, in a cross-linguistic study of the world's lexicons, investigated colexification patterns captured in a semantic map. Unlike Haspelmath who showed that 12 diverse languages are sufficient to build a stable semantic map, Francois posits that, in fact, the number of distinctions between senses increases with the number and variety of considered languages. Following these studies, List et al. (2018) built a weighted colexification graph using data from 195 languages in 44 language families, with subsequent improved language coverage versions (Rzymski et al., 2020a). Here, closely-related or similar concepts tend to be often densely connected (List et al., 2018; Georgakopoulos et al., 2021). Youn et al. (2016) constructed colexification graphs in the domain of natural objects (celestial and landscape) and investigated their polysemy distributions for the task of semantic similarity. We also take advantage of recurrent patterns in semantic structure across different language families. However, unlike them, we found some evidence that geographical and cultural differences matter in the human perception of our cross-linguistically connected concepts. Pericliev et al. (2015) distinguished between homonymy and polysemy by investigating the colexifications of 100 basic concepts. Georgakopoulos et al. (2021) discovered cross-linguistic similarities based on colexification patterns of verbs of vision and hearing in the the domain of perception-cognition. Jackson et al. (2019) relied on colexification patterns to test the universality of emotion perception. Like us, they show that, while there are shared structures of (affective) meanings across cultures, there are also some variations. Di Natale et al. (2021) tested whether colexification patterns in multilingual resources are correlated with affective meaning similarity between words. Bao et al. (2021) showed that no two concepts are colexified in every language by analyzing colexification data from three resources: BabelNet, Open Multilingual WordNet (Bond and Foster, 2013), and the Database of Cross-Linguistic Colexifications (CLICS3) (Rzymski et al., 2020b).

Although the scope of research on colexification varies across projects, most studies have assumed that colexification captures some degree of semantic similarity. Indeed, this is implied to some extent by the very definition of colexification, and supported by previous results in linguistics and NLP, suggesting that more commonly colexified meanings across languages require less cognitive effort to relate and recall (Xu et al., 2020). However, such a connection between cross-linguistic colexification patterns and semantic similarity has not been fully assessed at a large scale. Given that features of the lexicon are not easily identifiable across many languages, one solution is to rely on large data repositories to unveil cross-linguistic generalizations. However, since most languages lack dedicated lexical resources for semantic similarity (henceforth, *low-resource languages*), one option is to *transfer lexico-semantic knowledge from large shared cross-lingual semantic spaces*. In this paper, as part of a large scale empirical study, we show that lexico-semantic associations captured by cross-lingual colexification patterns in BabelNet contribute significantly in assessing if two words are semantically similar.

3 Synset Cross-Lingual Colexification

Under the colexification framework, the primary unit of observation for lexical typology is no longer the word, but the sense – a functionally-based criterion of concept definition (François, 2008). In this project, however, we use *the synset* as the primary unit. For us, this is also a technical consideration, since only synsets can be compared across languages, especially in BabelNet. Lexical concepts are grouped into sets of cognitive (near) synonyms, called synsets, each encoding a distinct

meaning. Synsets are connected through lexical relations and conceptual/semantic relations (i.e., hyperonymy, hyponymy, meronymy, etc.). In this paper, we use only lexical relations to capture colexifications between meanings. For instance, the lexical relation between the senses of 'fire' and 'firewood' of the word uta (in Kamaro) is a case of *strict colexification*. *Loose colexifications* like derivationally-related forms can show interesting semantic associations, but are not considered here.

Two synset concepts that are colexified in at least one language are usually perceived as somehow semantically connected, either directly or indirectly. However, proving such connectedness is by no means an easy task. The accurate description of lexical data often requires taking into account the many functional properties of real-world referents as well as culture-specific aspects of a language or geographic area. Such cases might capture underlying linguistic phenomena such as metaphor, metonymy, hyperonymy, analogical extension, and a rich set of cases of semantic shifts unique to each language or language family (Juvonen and Koptjevskaja-Tamm, 2016) – whose analysis falls within the scope of semantic or etymological studies, and beyond our goal here. Instead, our purpose is to organize cross-linguistic sense information in a way that captures various semantic connections between senses, allowing one to zoom in and out on aspects of the lexicon in cross-linguistic comparative studies. We rely here on the powerful structure of BabelNet that maps concepts (i.e., synsets) across a large, typologically diverse set of languages. This allows us to empirically examine at a large scale the contribution of such a rich body of knowledge to the task of semantic similarity - such empirical evidence is still lacking in the field. Our model of semantic connection between synsets (we call *Syn2Vec*) is simple: Given the set of concepts (synsets) of a lexeme, we assign a semantic link between every synset-synset pair. We want to investigate the idea that, as more and more languages are explored, and more and more senses are amassed, the resulting graph of cross-linguistic inter-connected synset concepts will capture aspects of semantic knowledge that might be missing in one language alone.

Given this intuition, next we briefly introduce the lexical resource used and our proposed algorithm to construct colexification graphs which model the synset semantic connections. We hypothesize that the conceptual representations of lexical typology captured by our cross-lingual colexification patterns do match, to some extent, the language-internal perception of native speakers, and test this hypothesis empirically on the LSIM task.

Algorithm 1 Construction of Colexification Graph: Given a set of languages L and corresponding vocabularies V, create graph edges between all colexified synset pairs (nodes).

```
function ConstructGraph(L, V)
    CSP ← {}                                  ▷ Colexified Synset Pairs.
    for l ∈ L do
        for x ∈ V_l do
            if |S_x| ≥ 2 then
                for {s1, s2} ∈ (S_x choose 2) do
                    CSP ← CSP ∪ {s1, s2}
                end for
            end if
        end for
    end for
    G ← graph
    for {s1, s2} ∈ CSP do
        G(s1, s2) ← 1
    end for
    return G
end function
```

A. The Lexical Resource, BabelNet. To collect synset information, we use BabelNet (Navigli and Ponzetto, 2010), to our knowledge, the largest cross-linguistic semantic network that extends the popular WordNet (Miller, 1995) by integrating other resources (Wikipedia, Wiktionary, etc). Every BabelNet synset (20.3M total) is identified as either a concept or named entity, and we only use concept synsets (7.2M) for our analysis. The maximum and minimum numbers of (concept[2]) lexemes in a language are 6.1M (English) and 1.8M (Parthian), respectively.

B. Building the Colexification Graphs. We denote the set of languages as $L = \{l_1, l_2, ..., l_N\}$, and the language vocabulary lists as $V = \{V_{l_1}, V_{l_2}, ..., V_{l_N}\}$. The vocabulary for each language is represented by $V_{l_n} = \{x_1, x_2, ..., x_{|V_{l_n}|}\}$ where the elements x_k are lexemes. For each lexeme x_k we have a corresponding set of synsets $S_{x_k} = \{s_1, s_2, ..., s_{|S_{x_k}|}\}$. For a colexification graph G, nodes represent synsets, and edge weights the semantic connection between two synsets. We denote the weight for each edge $\{s_1, s_2\}$ in the graph as $G(s_1, s_2) = G(s_2, s_1)$ (undirected graph). Algorithm 1 details the graph construction.

In this study, we examine two types of colexification graphs: (1) Monolingual, and (2) Cross-

[2]We filter out lexemes that have no concept synsets.

lingual. For monolingual graphs, we choose one language l and provide $L = \{l\}$ and $V = \{V_l\}$ to Algorithm 1. For the cross-lingual graph, we use all languages, i.e. $L = \{l_1, l_2, ..., l_N\}$, $V = \{V_{l_1}, V_{l_2}, ..., V_{l_N}\}$. In BabelNet, the same concept in each language will be mapped to a common synset and thus be represented as a node in the colexification graph.[3]

C. Creating Synset and Word Embeddings. As we like to capture lexico-semantic associations, given a colexification graph G, we assume that vector representations of the nodes (synsets) that are close to one another are similar as computed by cosine distance in the embedding space. We use a recently-developed node embedding approach, ProNE (Zhang et al., 2019), which, compared to other popular node embedding techniques like Deepwalk and Node2Vec (Perozzi et al., 2014; Grover and Leskovec, 2016), is much faster and demonstrates superior node representations on several classification datasets using a lib-linear classifier. We use the Python implementation from `nodevectors`[4] with all default hyperparameters.

In predicting the lexical similarity of two words, we assume that their perceptual similarity is determined by summing the synset embeddings of each word, then comparing the results. Thus, a word embedding w is computed as:

$$w = \sum_{s \in S_{Babel(w)}} s_{emb} \quad (1)$$

where s_{emb} is the embedding for synset s and $S_{Babel(w)}$ is the set of synsets of word w in BabelNet. Prior to each semantic similarity task, we normalize each word embedding to have magnitude one. Next, we take all evaluation words and perform mean centering, then Principal Component Analysis[5] (PCA) following (Ghannay et al., 2016), as we empirically found this improves performance.

4 Baselines

We evaluate the quality of our BabelNet dictionary approach by comparing it to high-quality and popular word and synset embedding approaches. We want to test whether the structural regularities observed in distributed text representations provide a route past some of the limitations of dictionaries, whether these two representations are comparable, and whether their combination might benefit the task of lexical similarity. Specifically, we compare to the well-known static word embedding approach fastText (Joulin et al., 2016), and an approach that extracts contextualized representations of words from a pretrained BERT language model (Vulić et al., 2020b), which we call "BERT." We also compare to "ARES", a recent synset/sense embedding model (Scarlini et al., 2020) that builds representations for each synset by collecting relevant contexts and extracting contextual embeddings of lemmas belonging to each synset from a pretrained language model (BERT). Similar to (Scarlini et al., 2020), we rely on ARES synset embeddings for our multilingual analysis. To compute ARES word embeddings, we follow Equation (1), but use the pretrained ARES multilingual synset embeddings[6]. We re-implement the BERT baseline[7], but use pretrained word embeddings for fastText[8].

5 Experimental Setup

Lexical semantic similarity (LSIM) seeks to accurately measure the perceived similarity in meaning between two words and does so by the Spearman's rank correlation[9] between similarity scores of human judgments and those computed automatically (cosine similarity of the words' vector representations). We rely here on Multi-SimLex (Vulić et al., 2020a), arguably the most comprehensive semantic similarity evaluation resource to date, which contains monolingual lists of 1,888 word pairs, with aligned concepts in 13 typologically diverse languages[10]. Diverse criteria were used here to test whether two words are semantically similar, and not vaguely associated. Thus, the Multi-SimLex datasets could be used as "an intrinsic evaluation benchmark to assess the quality of lexical representations based on monolingual, joint multilingual, and transfer learning paradigms" (Vulić et al., 2020a). Of the original 13 languages, we limit our

[3] We removed the three BabelNet noisy lexemes (with >1,000 synsets): the empty string "" (common to all languages); and "asteroid list" in both Russian (RU) and Armenian (HY): "список_астероидов" in (RU); "աստերոիդների_ցանկ" in (HY).

[4] https://github.com/VHRanger/nodevectors

[5] We keep the vector dimension the same.

[6] http://sensembert.org/resources/ares_embedding.tar.gz, ares_bert_base_multilingual.txt

[7] We must extract embeddings ourselves using the pretrained models (see A.1).

[8] https://pypi.org/project/fasttext/

[9] We use "average" rank mode from scipy.stats.rankdata()

[10] Since publication of the dataset, Arabic has been added.

	AR (838)	EN (1822)	ES (1728)	FI (1717)	FR (1798)	HE (1085)	PL (1176)	RU (972)	ZH (1563)	Mean ↑	Std. ↓
fastText	0.50	0.54	0.53	0.64	0.59	0.41	0.45	0.45	0.53	0.52	0.07
BERT	0.44	0.57	0.52	0.62	0.41	0.37	0.38	0.37	0.62	0.48	0.10
ARES	0.49	0.50	0.51	0.58	0.50	0.43	0.47	0.43	0.55	0.49	0.05
COLEX$_{cross}$	0.59	0.72	0.67	0.69	0.65	0.61	0.66	0.60	0.66	0.65	0.04
COLEX$_{maxsim}$	0.59	0.68	0.62	0.62	0.60	0.60	0.66	0.60	0.62	0.62	**0.03**
COLEX$_{mono}$	0.25	0.36	0.47	0.42	0.44	0.13	0.24	0.22	0.38	0.32	0.11
C+F	0.65	**0.76**	**0.71**	0.76	**0.73**	**0.64**	**0.68**	**0.64**	0.70	**0.70**	0.05
C+B	0.63	0.75	0.70	0.75	0.68	0.62	0.66	0.61	0.70	0.68	0.05
B+F	0.54	0.58	0.55	0.66	0.58	0.41	0.46	0.46	0.61	0.54	0.08
C+F+B	**0.66**	0.75	0.70	**0.78**	0.72	0.61	0.65	0.62	**0.72**	0.69	0.05

Table 1: Comparison of each word embedding technique on the LSIM task (Spearman's rank correlation) for nine different languages. Number of word pairs per language is given in parentheses below each language code. The '+' symbol indicates concatenation followed by PCA for fusion of two or more models ('C'=COLEX$_{cross}$, 'F'=fastText, 'B'=BERT). Best score per language is bolded; best non-fusion model score per language is underlined.

	AR	EN	ES	FI	FR	HE	PL	RU	ZH	All
Vocabulary size (# lexemes)	2.4M	6.1M	3.3M	2.6M	3.3M	2.1M	2.7M	3.2M	3.1M	953.4M
# Polysemous lexemes	41.2k	323k	163k	106k	159k	24.7k	106k	151k	86k	4.7M
# Synsets	2.1M	4.1M	2.8M	2.4M	2.9M	2.0M	2.5M	2.8M	2.6M	7.2M
# Colexified synset pairs	233k	1.3M	582k	474k	773k	54.6k	244k	1.0M	312k	8.5M
Mean # synsets per lexeme	1.03	1.08	1.08	1.07	1.08	1.02	1.05	1.08	1.05	1.007
Mean # syns. per polys. lexeme	2.97	2.56	2.58	2.67	2.66	2.46	2.3	2.61	2.65	2.43

Table 2: BabelNet statistics for evaluation languages given our experimental setup (using concept synsets only). "All" column indicates cumulative statistics over all 499 languages. Polysemous lexemes have two or more synsets.

study to 9 languages: Arabic (AR), Spanish (ES), English (EN), Finnish (FI), French (FR), Hebrew (HE), Polish (PL), Russian (RU), Mandarin Chinese (ZH). Thus, we can directly compare with other baselines that are only readily available for languages with large pretrained language models.

Cross-lingual lexical semantic similarity (CLSIM) is identical to LSIM, except the words in each word pair are from different languages. Multi-SimLex (Vulić et al., 2020a) also makes available these cross-lingual similarity scores, excluding AR.

Colexification Evaluation Setups. We evaluate monolingual and cross-lingual colexification-based synset embeddings[11] in three variants:

COLEX$_{mono}$: Construct the colexification graph using one language only. Build word embeddings from synset embeddings using Equation (1).

COLEX$_{cross}$: Same procedure as COLEX$_{mono}$ except that we use all 499 languages to construct the colexification graph. The purpose is to capture the full complexity of the BabelNet data.

COLEX$_{maxsim}$: We use here the same synset embeddings from COLEX$_{cross}$. Specifically, following (Camacho-Collados and Pilehvar, 2018), for each evaluation word pair in the test set, we determine their similarity score by computing the maximum similarity among all pairs of synset embeddings. We perform PCA on the entire set of COLEX$_{cross}$ synset embeddings prior to similarity computation.

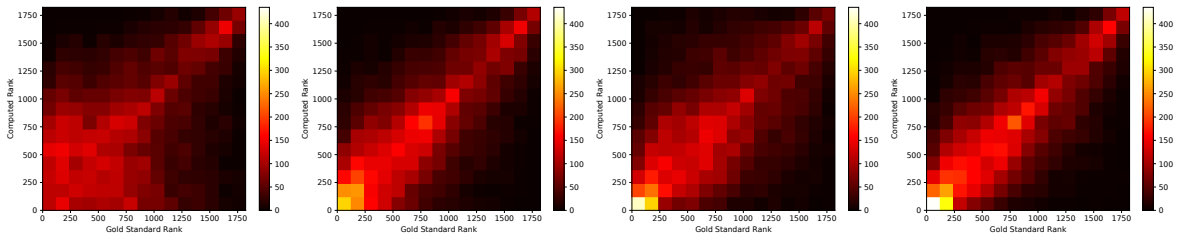
6 Results and Discussion

We introduce here various experiments to compare multiple methods. Since BERT and fastText can form representations from arbitrary string inputs, they have no out-of-vocabulary (OOV) words, while all methods relying on external resources (ARES and COLEX) require synset information for any word included in our experiments. For fair comparison, we limit our study to word pairs that include **both** words in the vocabulary of **all** approaches. Table 1 and Fig. 3 show the total number of word pairs (of the original 1,888) used for evaluation (relevant BabelNet stats in Table 2)[12].

We also use heatmaps (correlation plots) to give a clearer visual overview of the correlation between human judgments (i.e., gold standard) and our models' outputs, by word pair similarity rank (an ordinal measure) for all pairs in the test sets. Figure 1 shows the overall correlation

[11] All experiments were performed on an x86-64 server with a 32-core Intel(R) Xeon(R) Silver 4215R CPU and 754GB RAM. The node embedding process for the largest graph (COLEX$_{cross}$) took 20 minutes and 30GB of RAM. Our code is publicly available at https://github.com/jharvill23/Syn2Vec.

[12] OOV words breakdown by method: Table 5 (A.2)



(a) $\text{COLEX}_{\text{mono}}$ (b) $\text{COLEX}_{\text{cross}}$ (c) fastText (d) C+F Fusion

Figure 1: Heatmaps showing correlation of human rank judgments (X-coordinate) and overall computed ranks (Y-coordinate) for word pairs across all nine evaluation languages for four approaches: (a) $\text{COLEX}_{\text{mono}}$; (b) $\text{COLEX}_{\text{cross}}$, (c) fastText; (d) C+F Fusion (C: $\text{COLEX}_{\text{cross}}$ and F: fastText). The density of word pairs per square is represented by the square's color. Higher rank indicates that words in a given word pair are determined to be more similar, whereas low rank indicates dissimilarity.

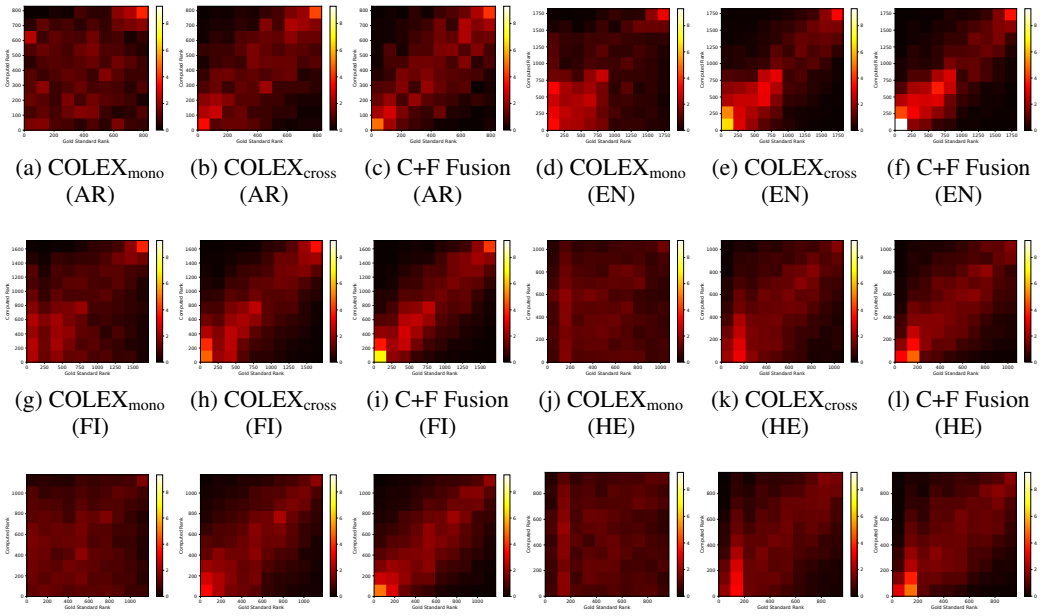
plots for all the test sets combined comparing our dictonary models ($\text{COLEX}_{\text{mono}}$, $\text{COLEX}_{\text{cross}}$), the best baseline (fastText), and our best fusion model ($\text{COLEX}_{\text{cross}}$+fastText). The color intensity of a square region corresponds to the number of word pairs in that region. We analyze the results of these different experiments next.

Cross-lingual vs. Monolingual Colexification. Cross-lingual colexification approaches $\text{COLEX}_{\text{cross}}$ and $\text{COLEX}_{\text{maxsim}}$ outperform the monolingual model $\text{COLEX}_{\text{mono}}$ by a large margin for every test language, and overall (Table 1). More specifically, the $\text{COLEX}_{\text{cross}}$ heatmap (Fig.1) shows significantly improved agreement on most dissimilar word pairs (i.e., the bottom left yellow squares), while more clearly converging on semantically similar instances (upper right squares), and reducing mis-ranked instances (away from the diagonal). This brings supporting empirical evidence for our main claim: *adding concepts (synsets) and edges in other languages captured as colexification patterns substantially contributes to the lexical similarity task.* For instance, unrelated words like 'aggressive'-'curved', 'airport'-'piece' are penalized by $\text{COLEX}_{\text{cross}}$, bringing the ranks closer to the gold standard. At the other end, the model better scores very similar pairs (like near synonyms): 'weird'-'strange', 'amazingly'-'fantastically', 'area' - 'region', 'capability'-'competence'. Various cases of colexification also bring to surface interesting lexico-semantic differences across languages – like 'charcoal'-'coal' or 'understand'-'know' which are not connected in English, but are colexified in other languages like Romanian – {'cărbune', 'jar', tăciune'} and {'a cunoaşte', 'a pricepe', 'a înţelege', 'a şti'}, respectively. Some languages have dedicated words that differentiate special instances of concepts, and thus are ranked as more dissimilar – i.e., 'toe'-'finger' or 'orteil-'doigt' (FR), while other languages (ES, RO: Romanian) colexify them, and perceive them as more similar: 'dedo del pie' – 'dedo' (ES) and 'deget de la picior' – 'deget' (RO) (translation: 'finger of/from foot' - 'finger').[13]

Mean vs. Max Similarity Representation. While both cross-lingual colexification approaches perform best on the LSIM task, there is still noticeable improvement of $\text{COLEX}_{\text{cross}}$ over $\text{COLEX}_{\text{maxsim}}$ for each language and overall across languages. In our experiments, comparing words by the average of their concepts proved more effective in modeling the semantic similarity for the evaluation of word pairs than making a comparison based on the most similar concepts from each word. A close comparison of the two models shows they differ in some specific cases, although these tendencies do not seem to generalize. On a few occasions, $\text{COLEX}_{\text{maxsim}}$ is better in penalizing hypernymy relations ('metal'-'aluminium', 'anatomy'-'biology'), as well as some syntagmatic relations ('breakfast'-'bacon', 'tsunami'- 'sea'). On the other hand, it seems to be over confident when it comes to nuanced concepts like 'stink'-'smell', 'mind'-'brain' whose interpretation varies more across languages, cultures, and philosophies.

Comparison to Baselines. All three baseline models are fundamentally limited by the quality of the contextual representations learned from raw text during training. While ARES makes use of external annotations and knowledge bases to decide which

[13]Note that the addition 'of/from foot' (RO: 'de la picior' and ES: 'del pie') is one of the many context-rich ways to point to the right finger, and is not part of the concept itself in these languages.



(a) $\text{COLEX}_{\text{mono}}$ (AR) (b) $\text{COLEX}_{\text{cross}}$ (AR) (c) C+F Fusion (AR) (d) $\text{COLEX}_{\text{mono}}$ (EN) (e) $\text{COLEX}_{\text{cross}}$ (EN) (f) C+F Fusion (EN)

(g) $\text{COLEX}_{\text{mono}}$ (FI) (h) $\text{COLEX}_{\text{cross}}$ (FI) (i) C+F Fusion (FI) (j) $\text{COLEX}_{\text{mono}}$ (HE) (k) $\text{COLEX}_{\text{cross}}$ (HE) (l) C+F Fusion (HE)

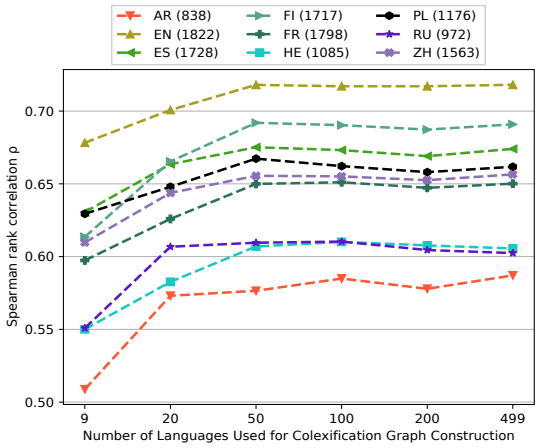
(m) $\text{COLEX}_{\text{mono}}$ (PL) (n) $\text{COLEX}_{\text{cross}}$ (PL) (o) C+F Fusion (PL) (p) $\text{COLEX}_{\text{mono}}$ (RU) (q) $\text{COLEX}_{\text{cross}}$ (RU) (r) C+F Fusion (RU)

Figure 2: Heatmaps representing correlation of gold standard (human) ranks and computed ranks for word pairs across six evaluation languages for $\text{COLEX}_{\text{mono}}$, $\text{COLEX}_{\text{cross}}$, and C+F Fusion. Due to each language having slightly different numbers of word pairs in our evaluation, we normalize each plot such that the density of each square represents the percentage (% out of 100) of the evaluation word pairs for the language. The density of word pairs per square is represented by the square's color. High rank indicates that the words in a given word pair are determined to be very similar, whereas a low rank indicates dissimilarity. Heatmaps for ES, FR, and ZH can be found in the Appendix (Fig. 4).

text should be used to represent synsets/senses, all text is passed through a BERT model for the final representation. To the best of our knowledge, distributional models (fastText, BERT) achieve the best performance published so far for Multi-SimLex (Vulić et al., 2020a,b). $\text{COLEX}_{\text{cross}}$ outperforms all baselines for all languages with a mean score >0.1 above the next-best baseline. This provides evidence that, for languages considered here, cross-lingual colexification-based word embeddings seem to capture word meaning more effectively compared to the baselines. The baseline scores correlate somewhat with one another, with lower scores for HE, PL, and RU, whereas this trend is not observed for cross-lingual colexification approaches. Additionally, cross-lingual colexification-based scores are more stable across evaluation languages with the lowest standard deviation of 0.03 for $\text{COLEX}_{\text{maxsim}}$, which is one big advantage of these approaches.

Embedding Fusion. We hypothesize that the baseline and cross-lingual colexification embeddings may contain rather different and possibly complementary semantic information due to the different paradigms for their construction (distributional hypothesis vs. knowledge-based), so we fuse these representations and evaluate on the LSIM task. Previous work has shown the simple concatenation of each method's word vector is rather unstable (Liu et al., 2020), leading to possibly worse results than each individual approach alone. However, by performing PCA on the resulting concatenated word vectors in the LSIM evaluation set, we see improved performance from fused methods for all languages (Table 1). These results favor our hypothesis: these combined representations align better with human perception than when evaluated individually.

Comparison across Individual Languages. We also analyzed the results of the models in each and across individual languages (see Fig. 2). When comparing $\text{COLEX}_{\text{mono}}$ to $\text{COLEX}_{\text{cross}}$ to C+F Fusion, we notice a huge improvement in similarity rank correlations with human judgments across all individual languages from $\text{COLEX}_{\text{mono}}$ to $\text{COLEX}_{\text{cross}}$. According to the individual language heatmaps (Fig. 2), the languages that seem to benefit most from the cross-lingual colexifica-



(a)

# Langs	9	20	50	100	200	499
%	46.4	77.6	96.1	99.5	99.9	100.0

(b)

Figure 3: (a) Performance on LSIM task for each evaluation language using 9, 20, 50, 100, 200, or 499 input languages to build the $\text{COLEX}_{\text{cross}}$ graph. Number of word pairs per language is given in parentheses in the legend. (b) Percentage of total colexified synset pairs in BabelNet collected for each language scenario.

tion approach vs. the monolingual one are AR, HE, PL, RU (compare with Table 1), with HE, AR and PL having the smallest dedicated lexical resources (i.e., the smallest number of colexified synset pairs, see Table 2, A.2). From $\text{COLEX}_{\text{cross}}$ to C+F Fusion, however, we see significant boosts in number of instances ranked across the diagonal, especially in the lower-bottom squares, across all individual languages. An interesting case here is FI which improves consistently and uniformly across all three models: along the diagonal, but also in reducing the mis-ranked instances (i.e., away from the diagonal).

Effect of Language Inventory Size on Embedding Quality. As shown so far, the large raw number of colexified synset pairs has contributed, in part, to the boost in performance between monolingual and cross-lingual colexification methods. We also show here that, the more languages are added to the colexification graph, the more unique synset colexifications are gathered, leading to a richer semantic network, and thus, better correlation with human judgments. Due to the imbalance of resources per language in BabelNet, we first create an ordered list of languages and choose the first 9, 20, 50, 100, 200, or all 499 languages to build separate graphs. The language list is ordered as follows: (1) Put the nine evaluation languages at the front of the list; (2) Add remaining languages in decreasing order by number of colexified synset pairs. LSIM results for each graph are given in Figure 3 (a). We find noticeable improvements in performance as we move from 9 to 50 languages, after which it saturates. Figure 3 (b) provides the percentage of total colexified synset pairs available from concept synsets in BabelNet collected for the aforementioned numbers of languages. Performance correlates well with the percentage of colexified synsets collected, supporting the hypothesis that the number of synset-synset relationships aquired across languages is the main driver in performance for our colexification-based approach.

	EN	ES	FI	FR	HE	PL	RU	ZH
EN	-	0.56	0.63	0.72	0.56	0.54	0.55	0.58
ES	**0.70**	-	0.58	0.65	0.51	0.50	0.49	0.55
FI	**0.71**	**0.69**	-	0.66	0.53	0.56	0.53	0.61
FR	**0.75**	**0.68**	**0.69**	-	0.55	0.68	0.64	0.68
HE	**0.66**	**0.64**	**0.66**	**0.64**	-	0.49	0.47	0.56
PL	**0.69**	**0.67**	**0.69**	**0.71**	**0.65**	-	0.45	0.53
RU	**0.64**	**0.61**	**0.65**	**0.66**	**0.62**	**0.61**	-	0.50
ZH	**0.70**	**0.66**	**0.69**	**0.69**	**0.64**	**0.66**	**0.60**	-

Table 3: Comparison of $\text{COLEX}_{\text{cross}}$ and fastText on the CLSIM task for eight evaluation languages. The values reported below the diagonal are from $\text{COLEX}_{\text{cross}}$ while those above are from cross-lingual fastText embeddings created using the fully-supervised configuration of VECMAP (Artetxe et al., 2018). Best score for each language pair is bolded. (See Table 6 in the Appendix for information about number of cross-lingual word pairs per language pair.)

Cross-lingual Performance. We also compared $\text{COLEX}_{\text{cross}}$ to fastText (the best-performing baseline on the LSIM task), on CLSIM (Cross-lingual Semantic Similarity), a task identical to LSIM, except the words in each word pair are from different languages (Vulić et al., 2020a) (see Table 3). We rely on VECMAP (Artetxe et al., 2018) to map two monolingual fastText embedding spaces to a common bilingual space. Table 6 (see Appendix) shows the total number of word pairs used per language pair. For every language pair, $\text{COLEX}_{\text{cross}}$ outperforms fastText, often by a significant margin. We show these results to provide additional evidence of the universality of our synset embeddings. Words from different languages can be compared directly under our formulation with no language-specific adaptation while simultaneously outperforming a competitive baseline for this task.

7 Suggestions for Future Improvement

Future work can expand our large-scale study of constructing synset and word representations from cross-lingual colexification principles in a number of directions. First, our cross-lingual embedding models seem specifically useful at ranking highly similar words just by amassing a large number of colexified synset pairs from many of the world's languages. However, while some colexification patterns might show more universal tendencies, others are very specific to a geographic area or language family, while others are more unique, identifying isolated cases of homonymy or other non-similarity phenomena. One possible solution is to represent as edge weight the number of languages that have a colexification pattern between two given synsets. This might result in a stronger model to identify either generalizations or more specific areal patterns (like language contact) by zooming out or in various areas of the graph, depending on one's research interests. Our model of semantic similarity does not distinguish the degree of similarity captured by each colexified synset. Figuring out a way to remove semantic links between colexified synsets that are only weakly or historically related may lead to higher quality synset and word representations capturing universal semantic tendencies, and thus run less the risk of an ethnocentric bias in favour of a specific language/area. Since languages can be compared at various levels of linguistic organization, it would be interesting to empirically investigate how colexification patterns involving core vocabulary differ in their genealogical stability compared with patterns at the periphery of the lexicon (Gast and Koptjevskaja-Tamm, 2021).

Unlike fastText and BERT, which are fully unsupervised, our proposed approach relies on external resources (BabelNet) for lexeme and synset information. Moreover, BabelNet is rather skewed in geographical coverage, typological diversity, and size of vocabulary across languages. From a sociolinguistic perspective, most of the BabelNet coverage comes from socio-politically dominant modern languages, even heavily Anglocentric (i.e., very rich, fine-grained distinctions of English lexicalizations). It would, thus, be interesting to test the efficacy of our model on a more balanced set of languages (as well as number of lexemes and synsets) from a more diverse (sub)set of language families.

For our semantic similarity evaluation, we relied on Multi-SimLex whose perception ratings of wide coverage lexical words were determined in an out-of-context fashion via human subject questionnaires, and through translation from English. Norm-generating studies involving large number of words have become increasingly popular across the cognitive sciences particularly due to their ability to provide greater statistical power, reduce experimenter bias in item selection, and increase study reliability (Lynott et al., 2020). Thus, correlation plots which intend to capture the relative strength of different colexification patterns are, in fact, an exploratory method and do not represent an attempt at rigorous hypothesis testing (Georgakopoulos et al., 2021). A comparison of out-of-context vs. in-context judgments and of differences between universal vs. more culturally-specific types of knowledge would advance research in lexical semantics.

8 Conclusions

This paper contributes to the investigation of lexico-semantic structures in the lexicon as manifested by colexification patterns captured in large synset graphs across BabelNet's diverse set of 499 world languages. We introduced several approaches – monolingual and cross-lingual colexification graphs, popular distributional vector space models, as well as a fusion of such systems. We evaluated the extent to which these models correlate with human judgments on a semantic similarity task covering 9 typologically diverse languages.

A deep analysis of the semantic similarity – relatedness/association continuum is not only important for research in lexical semantics and typology, but can also benefit a range of language understanding tasks in NLP. Our large scale cross-lingual colexification graph investigations highlight an important contribution: our word representation approach relies on synset embeddings across languages as captured in colexification graphs, and thus, no adaptation of such word embeddings is necessary for cross-linguistic comparisons (i.e., there is no need for mapping monolingual embeddings to a shared bilingual vector space). We have also tested and validated our cross-lingual colexification models (Tables 3 and 6) on the CLSIM task (Vulić et al., 2020a). The findings of our exploratory investigation of large-scale collexification graphs could inspire significant advances in NLP across languages, especially for tasks involving languages which lack dedicated lexical resources, and can benefit from language transfer from multilingual repositories.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In Proceedings of the 11th Global Wordnet Conference, pages 1–7.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. Journal of Artificial Intelligence Research, 63:743–788.

Simone Conia and Roberto Navigli. 2020. Conception: Multilingually-enhanced, human-readable concept vector representations. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1370–1380.

Anna Di Natale, Max Pellert, and David Garcia. 2021. Colexification networks encode affective meaning. Affective Science, 2(2):99–111.

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

Volker Gast and Maria Koptjevskaja-Tamm. 2021. Patterns of persistence and diffusibility in the european lexicon. Linguistic Typology.

Thanasis Georgakopoulos, Eitan Grossman, Dmitry Nikolaev, and Stéphane Polis. 2021. Universal and macro-areal patterns in the lexicon. Linguistic Typology.

Sahar Ghannay, Benoit Favre, Yannick Esteve, and Nathalie Camelin. 2016. Word embedding evaluation and combination. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 300–305.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pages 855–864.

Martin Haspelmath. 2003. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. Wikireading: A novel large-scale language understanding task over wikipedia. arXiv preprint arXiv:1608.03542.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 897–907.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. Science, 366(6472):1517–1522.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Päivi Juvonen and Maria Koptjevskaja-Tamm. 2016. The lexical typology of semantic shifts, volume 58. Walter de Gruyter GmbH & Co KG.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french.

Johann-Mattis List, Simon J Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. Clics2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. Linguistic Typology, 22(2):277–306.

Qianchu Liu, Diana McCarthy, and Anna Korhonen. 2020. Towards better context-aware lexical semantics:adjusting contextualized representations through static anchors. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4066–4075, Online. Association for Computational Linguistics.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. Behavior Research Methods, 3:1271 – 1291.

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In Proceedings of the 48th annual meeting of the association for computational linguistics, pages 216–225.

Vladimir Pericliev et al. 2015. On colexification among basic vocabulary. Journal of Universal Language, 16(2):63–93.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 701–710.

Christoph Rzymski, 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. 2020a. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. Scientific data, 7(1):1–12.

Christoph Rzymski, 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. 2020b. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. Scientific data, 7(1):1–12.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3528–3539.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020a. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. Computational Linguistics, 46(4):847–897.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020b. Probing pretrained language models for lexical semantics. arXiv preprint arXiv:2010.05731.

Yang Xu, Khang Duong, Barbara C. Malt, Serena Jiang, and Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. Cognition, 201:104280.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. Proceedings of the National Academy of Sciences, 113(7):1766–1771.

Jie Zhang, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding. 2019. Prone: Fast and scalable network representation learning. In IJCAI, volume 19, pages 4278–4284.

A Appendix

Appendix includes additional statistical information on experiments performed in this paper. Tables and Figures included: Tables 4, 5, 6; Fig. 4.

A.1 Context Examples and Pretrained BERT Models

We collect example sentences containing the evaluation words for each language from 2018 Wikipedia dumps[14]. We use the Perl script[15] from linguatools to convert xml format to raw text, excluding paragraph and heading mark-ups, and math and table tags. From raw text, we collect context sentences containing the evaluation words. Due to relatively insignificant differences between using 10 or 100 context examples for embedding extraction (Vulić et al., 2020b), we use 10 context examples for speed in running experiments. We choose the $L \leq 8$ layer setting and all other optimal settings from the original paper (Vulić et al., 2020b). We find pretrained BERT models for all languages except FR, for which we use a similar model called FlauBERT (Le et al., 2020). Pretrained models used in our re-implementation are given in Table 4. Note that the FR and RU models are cased, which may slightly affect the results. These were the only models we could find for these languages.

A.2 OOV Words

A detailed breakdown of OOV words by method is given in Table 5.

A.3 Heatmaps per Language

Fig. 4 shows heatmaps for languages missing from Fig. 2 for $\text{COLEX}_{\text{mono}}$, $\text{COLEX}_{\text{cross}}$, C+F Fusion[16].

[14] https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

[15] https://www.dropbox.com/s/p3ta9spzfviovk0/xml2txt.pl?dl=0

[16] The discussion of Fig. 2 is focused around low-resource languages.

AR	https://huggingface.co/asafaya/bert-base-arabic
EN	https://huggingface.co/bert-base-uncased
ES	https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
FI	https://huggingface.co/TurkuNLP/bert-base-finnish-uncased-v1
FR	https://huggingface.co/flaubert/flaubert_base_cased
HE	https://huggingface.co/onlplab/alephbert-base
PL	https://huggingface.co/dkleczek/bert-base-polish-uncased-v1
RU	https://huggingface.co/DeepPavlov/rubert-base-cased
ZH	https://huggingface.co/bert-base-chinese

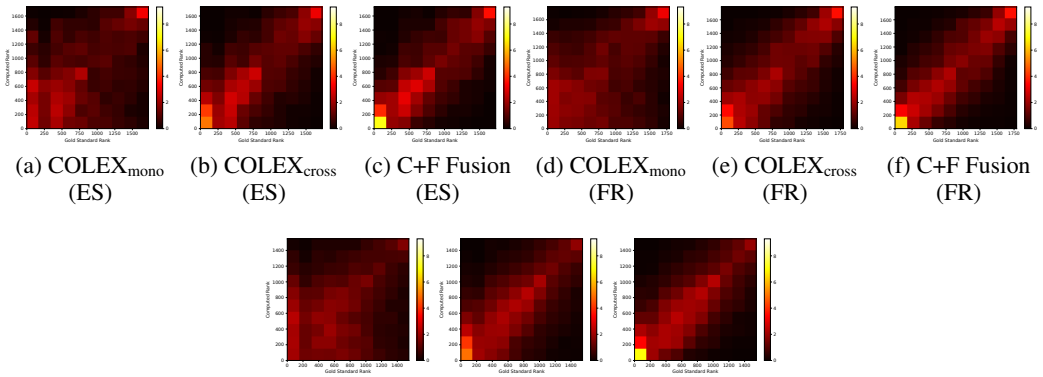
Table 4: Links to pretrained BERT models for each language

	AR	EN	ES	FI	FR	HE	PL	RU	ZH
fastText	0	0	0	0	0	0	0	0	0
BERT	0	0	0	0	0	0	0	0	0
BabelNet	709	0	84	72	33	374	342	479	209
ARES	824	10	99	84	47	469	372	628	250
$\text{COLEX}_{\text{mono}}$	860	45	117	140	68	623	547	710	295
$\text{COLEX}_{\text{maxsim}}$	755	0	88	73	34	402	350	526	226
$\text{COLEX}_{\text{cross}}$	755	0	88	73	34	402	350	526	226

Table 5: OOV words from Multi-SimLex for each approach. We provide OOV words for each language when querying BabelNet, because all COLEX approaches and ARES rely on BabelNet synset annotations. Any further OOV words for COLEX and ARES approaches beyond those not in BabelNet are due to not having at least one synset embedding for an evaluation word. For ARES, we are restricted to the pretrained embeddings provided at http://sensembert.org/resources/ares_embedding.tar.gz. For COLEX, a synset must be colexified at least once to have a vector representation.

	EN	ES	FI	FR	HE	PL	RU	ZH
EN	-	3222/3318	3275/3352	2257/2284	2794/3274	2798/3274	2551/3222	2913/3151
ES	3222/3318	-	3084/3256	2544/2645	2704/3256	2681/3250	2428/3189	2805/3116
FI	3275/3352	3084/3256	-	2595/2682	2718/3243	2756/3294	2502/3257	2850/3137
FR	2257/2284	2544/2645	2595/2682	-	2462/2903	1972/2379	1696/2219	2041/2243
HE	2794/3274	2704/3256	2718/3243	2462/2903	-	2391/3201	2242/3226	2453/3056
PL	2798/3274	2681/3250	2756/3294	1972/2379	2391/3201	-	2262/3209	2419/3009
RU	2551/3222	2428/3189	2502/3257	1696/2219	2242/3226	2262/3209	-	2244/3032
ZH	2913/3151	2805/3116	2850/3137	2041/2243	2453/3056	2419/3009	2244/3032	-

Table 6: Ratio of cross-lingual word pairs used for each language pair. Numerator represents number of word pairs used and denominator represents total word pairs provided in Multi-SimLex for each language pair.



(a) $\text{COLEX}_{\text{mono}}$ (ES) (b) $\text{COLEX}_{\text{cross}}$ (ES) (c) C+F Fusion (ES) (d) $\text{COLEX}_{\text{mono}}$ (FR) (e) $\text{COLEX}_{\text{cross}}$ (FR) (f) C+F Fusion (FR)

(g) $\text{COLEX}_{\text{mono}}$ (ZH) (h) $\text{COLEX}_{\text{cross}}$ (ZH) (i) C+F Fusion (ZH)

Figure 4: Heatmaps for ES, FR, and ZH. See Fig. 2 from main paper.