Partial Colexifications Improve Concept Embeddings

Arne Rubehn and Johann-Mattis List
Chair for Multilingual Computational Linguistics
University of Passau
Passau, Germany
{arne.rubehn,mattis.list}@uni-passau.de

Abstract

While the embedding of words has revolutionized the field of Natural Language Processing, the embedding of concepts has received much less attention so far. A dense and meaningful representation of concepts, however, could prove useful for several tasks in computational linguistics, especially those involving cross-linguistic data or sparse data from low resource languages. First methods that have been proposed so far embed concepts from automatically constructed colexification networks. While these approaches depart from automatically inferred polysemies, attested across a larger number of languages, they are restricted to the word level, ignoring lexical relations that would only hold for parts of the words in a given language. Building on recently introduced methods for the inference of partial colexifications, we show how they can be used to improve concept embeddings in meaningful ways. The learned embeddings are evaluated against lexical similarity ratings, recorded instances of semantic shift, and word association data. We show that in all evaluation tasks, the inclusion of partial colexifications lead to improved concept representations and better results. Our results further show that the learned embeddings are able to capture and represent different semantic relationships between concepts.

1   Introduction

In structural linguistics, the linguistic sign is traditionally modeled as a bipartite structure, consisting of a *form* that signifies a *meaning* (de Saussure, 1916). Word embeddings computed from large text collections (Mikolov et al., 2013a) offer a convenient way to express semantic properties of words in a particular language in low-dimensional spaces. In such an approach, meaning is defined only structurally, in relation to other meanings. Whether word embeddings capture meaning in the form of a "relation between a linguistic form and communicative intent" (Bender and Koller, 2020), however, can at least be doubted.

While word embeddings provide information on semantic relations inside individual languages, *cross-linguistic colexifications* – instances of different concepts expressed by a single word form across multiple languages (compare German *böse* colexifying the concepts EVIL and ANGRY) – capture a different kind of semantic relation. Only small amounts of colexifications can be observed for any individual language in the world, yet taking the languages of the world together, we can infer dense concept networks from cross-linguistic colexification patterns that show fascinating semantic structures (List et al., 2013). In order to employ these patterns in machine learning approaches, previous studies have proposed to embed the underlying graphs, using graph embedding techniques by which nodes in a graph are projected on vector spaces that represent their relation to other nodes in the graph (Harvill et al., 2022). Given that colexification networks represent concepts defined independently of individual languages, concept embeddings can be learned from colexification networks in a similar way to how word embeddings are learned from distributional data.

All concept embeddings methods presented so far have exclusively relied on *full colexifications*. This means that the concept networks were derived from those cases in which two or more concepts are colexified by the *same* word form in a given language. *Partial colexifications* – colexification patterns in which words do not need to match entirely in order to colexify two or more concepts –, on the other hand, have not been tested so far.

Here, we present initial tests in which both full and partial colexification data are used to train concept embeddings with the help of different graph embedding techniques. We evaluate the quality of the learned embeddings on three different tasks,

including the modeling of semantic similarity, the prediction of semantic change, and the prediction of word associations. Our results show that enriching concept embeddings with partial colexification data greatly improves their performance on all three tasks, outperforming similarity metrics inferred from the graph topology directly, and surpassing static word embeddings in two out of three tasks.

2 Background

2.1 Word Embeddings

Based on the distributional hypothesis (Harris, 1954; Firth, 1957), word embeddings are dense, low-dimensional representations of words learned from their co-occurrence distributions. With their ability to capture different aspects of semantics, they have laid the groundwork for the success of state-of-the-art techniques in language processing. Static embedding models like Word2Vec (Mikolov et al., 2013a), GloVE (Pennington et al., 2014), and fastText (Bojanowski et al., 2017) have quickly gained popularity and are still widely used for modeling lexical semantics in a latent space, mainly because they are more transparent and interpretable than contextual embeddings employed in Transformer architectures (Vaswani et al., 2017) such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019). Speer et al. (2017) show that static word embeddings can be enhanced with information from knowledge graphs, leading to improved representations.

Word embeddings are inherently monolingual, since the embedded items are only meaningful with respect to a certain language. They can therefore not be used directly in cross-linguistic applications. In order to address this problem, multiple approaches for obtaining multilingual embeddings have been proposed. While some approaches map monolingual embeddings onto a shared space (Ammar et al., 2016; Artetxe et al., 2018; Lample et al., 2018), other approaches learn multilingual embeddings directly from parallel corpora (Dufter et al., 2018; Levy et al., 2016). To this end, Camacho-Collados et al. (2017) devised a benchmark dataset for evaluating multilingual and cross-lingual semantic word similarity.

2.2 Colexification Networks

Colexification, originally introduced by François (2008), is a cover term for both polysemy and ho-

mophony. Thinking of the word forms in a given language having different *senses* which refer to different *concepts*, two concepts are thought to be colexified when there is one word form that expresses both of them at the same time. While it would be difficult to assess *why* two words colexify, given that it is not a trivial task to distinguish polysemy from homophony (Tjuka, 2023), it is computationally very easy to *infer* colexifications from comparative wordlists where a list of concepts is translated into one or more languages (List, 2022). Collections of colexification across multiple languages can be modeled in the form of *colexification networks* (List et al., 2013), revealing interesting colexification patterns that may reflect typological or areal trends in lexical typology (List et al., 2018). The *Database of Cross-Linguistic Colexifications* (CLICS, Rzymski et al. 2020, https://clics.clld.org) offers a large collection of colexifications inferred from several thousand languages that can be used in computational approaches or interactively inspected.

Given that all semantic change takes polysemy as its starting point (Traugott and Dasher, 2002), colexification networks prove useful when investigating common pathways and tendencies of semantic change (Münch and Dellert, 2015; Dellert, 2024; Bocklage et al., 2024a). Given that they also reflect certain aspects of conceptual proximity, they have been used to study semantic similarity in different lexical domains, including emotion semantics (Jackson et al., 2019), body part terminology (Tjuka et al., 2024), and affective meaning (Di Natale et al., 2021).

While most colexification studies only take *full* colexifications into account, where identical word forms colexify different concepts, Urban (2011) has shown that semantic relations derived from word parts – reflecting overt marking in lexical coding – may also reveal interesting structures that provide additional hints on the directionality of semantic change in particular and lexical motivation in general. List (2023) applies this idea to a smaller dataset of 329 languages taken from the *Intercontinental Dictionary Series* (Key and Comrie, 2016), showing that *partial colexifications* can be inferred conveniently in two flavors. *Affix colexifications* point to directional relations between two concepts where one of the words expressing the concepts is a prefix or a suffix – in the computational sense – of the other word. *Overlap colexifications* point to undirected relations between concepts that share a

substring of a given length. Overlap colexifications usually connect morphologically complex words that are derived or compounded using the same stem, while affix colexifications directly connect derived or compounded forms to their stems. Examples for full and partial colexifications according to this notion are given in Table 1.

2.3 Graph Embeddings

Inspired by the success of word embeddings and deep neural networks, graph embedding techniques have gained traction over the past decade. The task of embedding a graph is – similar to word embeddings – to learn low-dimensional representations for each node in a graph, such that properties of and relationships between nodes are preserved. Representing nodes in an embedded space has several advantages over the direct representation of a graph as a set of nodes and edges: The computational complexity decreases, downstream methods can be parallelized, and nodes can directly serve as inputs for machine learning models (Cui et al., 2019).

Most models for learning graph embeddings can be categorized according to three basic strategies, where we find models based on *matrix factorization* (Cao et al., 2015; Ou et al., 2016; Zhang et al., 2019), *random walks* (Perozzi et al., 2014; Grover and Leskovec, 2016), and *deep learning* (Wang et al., 2016; Kipf and Welling, 2016; Hamilton et al., 2017). Common applications of graph embeddings include node classification, link prediction, and network visualization (Cui et al., 2019; Goyal and Ferrara, 2018).

2.4 Concept Embeddings

Compared to the popularity of word embeddings, the embedding of concepts or word senses have received only little attention so far. For monolingual word senses, WordNet (Fellbaum, 1998) is often used as a reference catalogue whose synsets are embedded, usually for word sense disambiguation purposes (Trask et al., 2015; Rothe and Schütze, 2015; Jafarinejad, 2023). In the same spirit, scholars have trained embeddings for different senses of polysemous words (Camacho-Collados et al., 2015; Iacobacci et al., 2015; Colla et al., 2020; Pilehvar and Collier, 2016). While these works deal with the colexification of different senses, they lack a multilingual component, since the learned embeddings are inherently tied to a certain language.

Closest to our work are studies that learn concept embeddings from multilingual data. Harvill et al. (2022) learn concept embeddings from BabelNet (Navigli and Ponzetto, 2012), an automatically crawled multilingual extension of WordNet. Chen et al. (2023) combine BabelNet and CLICS to learn concept embeddings, which are aggregated to represent languages based on their colexification patterns. Conia and Navigli (2020) infer vector representations from BabelNet as well, however those are sparse and high-dimensional. Finally, Liu et al. (2023) train concept embeddings from colexification networks that are inferred automatically over parallel multilingual text corpora.

We intend to improve on recent work in two ways. First, our embeddings are trained on manually curated colexification data, which is less error-prone than automatically constructed colexification graphs. Second, we do not only take full colexification into account, but also consider partial colexifications. We believe that this adds complementary information of which the potential has not been explored yet.

3 Materials and Methods

3.1 Colexification Data

We use colexification data inferred by List (2023) from the *Intercontinental Dictionary Series* (IDS, Key and Comrie, 2016, https://ids.clld.org) spanning 329 language varieties and 1,310 concepts. The inventory of concepts is defined by Concepticon (List et al., 2025, https://concepticon.clld.org, Version 3.3), a reference catalog that defines more than 4,000 different concepts (referred to as *concept sets*), taken from questionnaires, wordlists, or concept lists that are typically used for language documentation. Being built upon resources designed by linguists to document languages all over the world, concept sets in Concepticon are defined independently of the languages in which they are expressed, even if the language used to gloss these concepts is typically English. This means that concept sets defined by Concepticon represent a cross-linguistic resource that attempts to define concepts that are expressed coherently across many of the world's languages.

The colexification data from List (2023) comprises not only *full* colexifications, but also *affix* and *overlap* colexifications, as described above (see § 2.2). From the colexifications identified in List (2023), it is straightforward to build three colexification networks, one for each colexification type. In each network, the concept sets form

Language	Word A	Concept A	Word B	Concept B	Colexification
Yaqui	dʒ u j a	TREE	dʒ u j a	FOREST	full
Guìlín (China)	çy ^{21}	TREE	çy ^{21} l i ŋ ^{21}	FOREST	affix
Fúzhōu (China)	tsʰ j eu ^{212} pʰ w oi^5	BARK	tsʰ j eu ^{212} l i ŋ ^{53}	FOREST	overlap

Table 1: Different colexification types, following List (2023) with examples across different languages (superscript letters code for tones in Sinitic examples).

the nodes that are connected by edges weighted by language families: If two concepts are connected with a weight of 5, this indicates that at least one instance of this colexification can be found in five different language families. Full and overlap colexifications are undirected by nature and thus form undirected graphs, whereas affix colexifications are inherently directed. However, since all downstream tasks only measure symmetric similarities between concepts, and most graph embedding techniques cannot account for directionality (Cui et al., 2019), the affix colexification network is transformed to an undirected network as well. An overview of the three resulting graphs is given in Table 2.

3.2 Concept Embeddings

We employ and compare three different graph embedding techniques to learn low-dimensional representations of the nodes (i.e. the concepts) in each graph; each method representing one of the three main paradigms for graph embedding techniques. SDNE (*Structural Deep Network Embedding*, Wang et al., 2016) is a deep autoencoder that optimizes embeddings towards retaining first- and second-degree neighborhoods. Node2Vec (Grover and Leskovec, 2016) is a well-known graph embedding technique that samples random walks from the graph, which then are used as training data for a simple Word2Vec model (Mikolov et al., 2013a). Due to the sampling of random walks, this technique is able to learn some more distant relationships between nodes, while still preserving close community structures. Finally, ProNE (Zhang et al., 2019) employs sparse matrix factor-

ization enhanced by spectral propagation, leading to extremely efficient training of high-quality embeddings.

For each of the three colexification networks and with each of the three embedding techniques, 128-dimensional embeddings are trained with fairly standard hyperparameters. In a postprocessing step, the embeddings trained on the full colexification network are combined with embeddings for either or both of affix and overlap colexifications. Following Harvill et al. (2022), this is achieved by concatenating the embeddings and then employing Principal Component Analysis (PCA) to reduce them back to the original dimensionality. Combining the embeddings learned on the different graphs as a postprocessing step has two major advantages over combining the graphs *before* training the embeddings: First, the graphs are not balanced in terms of edge weights. In particular, partial colexifications are more commonly found than full colexifications – simply aggregating the graphs would therefore overemphasize the role of partial colexifications over full colexifications, which are thought to be the strongest type of colexification. Second, this procedure ensures that as much information as possible is retained from all graphs – similar components of different graphs will lead to similarities along some dimensions between the respective embeddings. PCA reduces redundant information to singular dimensions, while keeping other pieces of information that are *complementary*, and thus not shared between the original graphs.

Finally, six sets of embeddings are obtained per embedding technique, combining information from different types of colexification: *full*, *affix*, *overlap*, *full/affix*, *full/overlap*, and *full/affix/overlap*.

3.3 Baselines

The quality of the trained embeddings is evaluated against four simple methods to infer similarities (or distances) between nodes from the graph directly. Baselines reported for multiple colexification types

Colexification Type	Nodes (Concept Sets)	Edges
full	1,246	4,008
affix	1,308	38,215
overlap	926	12,974

Table 2: Number of nodes (concept sets) and edges in the three colexification networks used in our study.

simply aggregate the edge weights of corresponding graphs.

The first baseline metric is the length of the *shortest path* between two nodes on a graph with inverted weights, where each weight w is transformed by the simple function $f(w) = w^{-1}$. This is necessary since the weights in the original graphs indicate similarities rather than distances, with higher weights between nodes indicating a stronger connection.

In the remaining three similarity metrics, we follow De Deyne et al. (2018) who compare *cosine similarity*, *positive pointwise mutual information (PPMI)*, and *random walk similarity* on word association networks. The first two metrics can be easily inferred from the graph's adjacency matrix by calculating the cosine similarity between the rows of individual concepts, or the PPMI respectively. The last metric simulates random walks from each node and measures how similar the average random walks for different start nodes are.

Finally, we map the embedded concepts to fastText embeddings in nine different languages (Grave et al., 2018) via parallel word pairs in Multi-SimLex (Vulić et al., 2020), which has been manually linked to Concepticon concept sets in the past (List, 2021). For concepts with multiple lexical realizations (e.g. the concept set CAR corresponds to Russian *avtomobil'* and *mashina*, two words that are both typically used to refer to a car in Russian), the fastText vector corresponding to that concept is the mean of the vectors of the respective words, weighted by their relative frequency. For a fair comparison, we reduce the dimensionality of the vectors from originally 300 to 128 components, the same size as the trained concept embeddings. We report the averaged and the best performance of these embeddings in all available languages and provide a direct comparison to the trained concept embeddings.

3.4 Experiments

We conduct three experiments, in which we test how well our concept embeddings align with different aspects of semantic similarity, concentrating on human similarity ratings for individual word pairs across multiple languages (*Modeling Lexical Semantic Similarity Task*, § 3.4.1), semantic change processes as documented in the linguistic literature (*Predicting Semantic Change Task*, § 3.4.2), and word associations derived from experimental data (*Predicting Word Associations Task*, § 3.4.3). With

these three tests, we intend to cover three scenarios in which concept embeddings might play an important role in future research. All three evaluations are based on similarities between concept pairs, inferred by calculating the cosine similarity between embedding vectors of individual concept sets.

3.4.1 Modeling Lexical Semantic Similarity

The *lexical semantic similarity* task (LSIM, Harvill et al., 2022) measures to which degree the similarities between embeddings correlate with human similarity judgements. We obtain these ratings from Multi-SimLex (Vulić et al., 2020), which encompasses similarity judgments for 1,888 word pairs in 12 typologically diverse languages. 546 of these pairs have been mapped to Concepticon (List, 2021), excluding word pairs for which no counter-part could be found in Concepticon. We further exclude near-synonyms linking to the same Concepticon concept sets, resulting in 538 concept pairs. The inferred embedding similarities are evaluated against this subsample. For each concept pair, we calculate the average similarity rating over all languages in Multi-SimLex in order to retrieve a cross-linguistic notion of semantic similarity that accounts for outliers in the data. The rating for the concept pair BANK vs. SEAT, for example, has a much higher similarity rating in Spanish than in any other language, due to the fact that Spanish *banco* is polysemous, referring to both BANK and BENCH. Following previous studies (Conia and Navigli, 2020; Harvill et al., 2022) we calculate Spearman's rank correlation (Spearman, 1904) between the averaged multilingual similarity ratings and the cosine similarities retrieved from the concept embeddings for different models and settings.

3.4.2 Predicting Semantic Change

Given the importance of polysemy and colexification data for the handling of semantic change patterns, we are interested to know how well concept embeddings can distinguish likely patterns of semantic change from less likely ones. To test this, we extracted several hundred historically attested semantic shifts from the *Database of Semantic Shifts* (DatSemShift, Zalizniak et al., 2024, https://datsemshift.ru/), a database that collects and structures individual instances of semantic change from the literature. While the database offers various data on semantic change, ranging from colexifications, via partial colexifications, up to cognate words with diverging meanings, we extracted

547 concept pairs in which the shift took place from an attested ancestral language to an attested daughter language, thus guaranteeing that these shifts can be observed directly through the comparison of sources. For the extraction, we used a computer-readable version of the database (Bocklage et al., 2024b) that codes the data in Cross-Linguistic Data Formats (CLDF, Forkel et al., 2018, https://cldf.clld.org), a format specification for linguistic data that facilitates the use in computational applications. We frame semantic change prediction as a binary classification task, where the actual shifts are instances of the positive class, while the instances of the negative class are constructed by replacing one of the concepts in each pair by a randomly selected concept. This process is known as *negative sampling* and is often employed in training objectives to prevent overfitting of models (Smith and Eisner, 2005; Gutmann and Hyvärinen, 2010; Mikolov et al., 2013b). We train a simple logistic regression classifier on this sample with the embedding similarity of the respective concept pair as sole input feature. To even out coincidental biases in the negative sampling, we run the sampling procedure 50 times and fit a logistic regression on the sample. The negative samples are shared across all models and baselines to ensure a fair comparison. Since all samples are perfectly balanced (with the same number of items for the positive and the negative class), we report the prediction accuracy averaged over all samples for each model.

3.4.3 Predicting Word Associations

Link prediction is one of the central problems in network analytics, where it is mainly used to recover missing data in sparse networks or predict likely links that might arise in the future (e.g. predicting which authors are likely to collaborate in the future based on current co-authorship networks; Liben-Nowell and Kleinberg, 2007). For evaluation purposes, this task is usually framed in a way that links from a network are predicted using information from another network in the same domain (such that the nodes of both networks can easily mapped onto each other). Following this paradigm, we use embedding similarities in order to predict links from the *Edinburgh Association Thesaurus* (EAT, Kiss et al., 1973), a large-scale network constructed from data in word association studies. In the experiment, participants were given cue words to which they were to respond with the first word that came to their mind. As Tjuka et al. (2022)

show with their *Database of Norms, Ratings, and Relations* (NoRaRe, https://norare.clld.org), a network from the EAT data can be easily inferred by counting all cue-response pairs as weighted edges and linking them automatically to the concept sets in Concepticon. NoRaRE assigns 1,787 of the 55,732 nodes in the EAT to concept sets in Concepticon. In order to filter out noise, we remove edges with a weight of less than five. Reducing the association network further to concepts present in the colexification data by List (2023) – the concept space that is being embedded – results in a network of 746 concepts with 780 edges. We formalize the link prediction task as a binary classification task with negative sampling; conceptually identical to the prediction of semantic change. Again, we draw 50 negative samples and report the average accuracy.

3.5 Implementation

All training and evaluation procedures were implemented in Python 3.12. The training routine for SDNE and Node2Vec was implemented from scratch using PyTorch (Paszke et al., 2019) and Scikit-learn (Pedregosa et al., 2011). The ProNE models were trained using the NodeVectors package (Ranger, 2024). The shortest path baseline was computed using the NetworkX (Hagberg et al., 2008) implementation of Dijkstra's algorithm (Dijkstra, 1959). Spearman rank correlations were computed using SciPy (Virtanen et al., 2020). Evaluation data was retrieved using PyCLDF (Forkel et al., 2025), PyConcepticon (Forkel et al., 2024), and PyNoRaRe (List and Forkel, 2024). Visualizations were created using Matplotlib (Hunter, 2007) and adjustText (Flyamer, 2024) packages. The supplementary material offers access to all data and code needed to replicate the analyses described in this study, including a detailed account of the package versions that were used in our implementation.

4 Results

Table 3 summarizes the performance of the different models and baselines on the three evaluation tasks. Since fastText embeddings are trained independently from the different colexification networks, the reported metric does not relate to the different types of colexification. As indicated in the previous section, we report the performance of the best language and the average performance across languages in fastText.

Method	Colexification Type					
	full	*affix*	*overlap*	*full / affix*	*full / overlap*	*full / affix / overlap*
shortest path	-0.60	-0.49	-0.18	-0.51	-0.45	-0.47
cosine similarity	0.50	0.42	0.15	0.50	0.27	0.39
PPMI	0.53	0.33	0.21	0.31	0.34	0.23
random walks	0.43	0.34	0.10	0.38	0.22	0.25
SDNE	0.48	0.36	0.36	0.60	0.46	0.47
Node2Vec	0.64	0.58	0.26	<u>0.69</u>	0.61	0.66
ProNE	0.64	0.63	0.26	**0.72**	0.62	0.66
fastText-ZH (best)				0.44		
fastText (avg.)				0.36		

(a) Spearman rank correlations for Multi-SimLex similarity ratings and concept similarities.

Method	Colexification Type					
	full	*affix*	*overlap*	*full / affix*	*full / overlap*	*full / affix / overlap*
shortest path	0.79	0.73	0.69	0.78	0.79	0.79
cosine similarity	0.69	0.74	0.65	0.76	0.70	0.75
PPMI	0.71	0.73	0.62	0.78	0.73	0.79
random walks	0.71	0.72	0.65	0.74	0.73	0.74
SDNE	0.66	0.66	0.70	0.78	0.75	0.75
Node2Vec	0.79	0.75	0.72	**0.83**	0.82	**0.83**
ProNE	0.78	0.78	0.71	<u>0.82</u>	0.80	**0.83**
fastText-ET (best)				<u>0.82</u>		
fastText (avg.)				0.78		

(b) Avg. accuracies of predicting semantic change from DatSemShift.

Method	Colexification Type					
	full	*affix*	*overlap*	*full / affix*	*full / overlap*	*full / affix / overlap*
shortest path	0.71	0.76	0.70	0.77	0.67	0.78
cosine similarity	0.64	0.73	0.65	0.74	0.65	0.73
PPMI	0.65	0.76	0.68	0.77	0.67	0.78
random walks	0.65	0.70	0.65	0.71	0.67	0.71
SDNE	0.59	0.67	0.68	0.72	0.72	0.74
Node2Vec	0.71	0.75	0.72	0.78	0.77	0.79
ProNE	0.71	0.79	0.71	0.80	0.77	<u>0.81</u>
fastText-EN (best)				**0.87**		
fastText (avg.)				<u>0.81</u>		

(c) Avg. accuracies of link prediction in the EAT.

Table 3: Performance of concept embeddings in various flavors across our three evaluation tasks. Best models are marked in **bold** font. Second best models are marked by <u>underlining</u> the value.

Some general patterns can be observed across all three tasks. Concerning the baselines – similarities (or distances for the *shortest path* metric) inferred directly from the graph – simply measuring the length of the shortest path consistently is a better predictor than the three methods discussed in De Deyne et al. (2018). On the downside, path lengths can only be inferred between connected nodes; however, this problem can easily be solved by assigning a default value – the similarity between disconnected nodes in all other three baselines is always 0.

Regarding the models, it is clear that SDNE is not viable for embedding colexification graphs, as it performs substantially worse than the other two embedding methods, and is even beaten by simpler baselines in most experiments. This model therefore will not be discussed in detail, instead, we will focus on the more performant models, Node2Vec and ProNE. These two models perform almost on par, with only a few differences, the most notable being that ProNE consistently outperforms Node2Vec on affix colexifications. In this context, it is also important to mention that the training time for ProNE (38 ms) was substantially lower than the training time for Node2Vec (13:27 min), supporting the claim that ProNE can increase the training efficiency by several orders of magnitude compared

to other graph embedding techniques (Zhang et al., 2019).

The latter two tasks, semantic change prediction (Table 3b) and association link prediction (Table 3c) show the same trends and mostly mirror each other. Since the EAT is not a multilingual, but a monolingual evaluation resource in English, it is not surprising that the English word embeddings are the best predictor for English word associations. However, word embeddings from other languages are generally good at capturing associative links – even across languages – and on average perform similar to the best concept embedding model (ProNE, all colexifications). For the concept embeddings, it is clearly discernible that including partial colexifications improves the embedding quality in both prediction tasks. This holds true for both affix and overlap colexifications. Node2Vec and ProNE both prove themselves as reliable graph embedding techniques with very similar performances; Node2Vec has a slight edge over ProNE in predicting semantic shifts, while ProNE performs slightly better on the EAT data. Both models outperform fastText embeddings when trained on the entire colexification data.

The lexical semantic similarity (LSIM, Table 3a) task shows some different trends. Overlap colexifications correlate poorly with perceived lexical similarity, and the inclusion of overlap colexifications is consistently detrimental to the embeddings' performance on LSIM. On the other hand, the combination of full and affix colexifications is very fruitful: Across all embedding models, this set-up yields the best performance by far. Again, Node2Vec and ProNE behave very similarly in terms of overall performance, but the latter outperforms the former in embedding affix colexifications, which propagates through the combined embeddings.

Word embeddings perform notably poorly on LSIM, with significantly lower correlations than simple baseline metrics inferred directly from full colexification networks. The weak correlation with perceived lexical similarity and the high accuracies on predicting word associations provide further support for the claim that word embeddings capture semantic association rather than semantic similarity (Yang et al., 2022; Karidi et al., 2024).

The results show that concept embeddings generally benefit from partial colexification data. In both prediction tasks, embeddings learned on full colexifications alone perform significantly worse than those enhanced by either or both types of par-

tial colexification. However, overlap colexification patterns are rather detrimental than helpful for assessing semantic similarity. Integrating affix colexifications, on the other hand, is clearly beneficial for all three downstream tasks, including LSIM.

5 Discussion and Outlook

Our results underline the benefits of combining full and partial colexification patterns when inferring concept embeddings from cross-linguistic colexification data. The combined embeddings consistently outperform embeddings based on full colexifications in our three different downstream tasks. It seems that partial colexifications – in their two flavors of affix colexifications and overlap colexifications, as tested in our study – can capture conceptual relationships that are rarely expressed by full colexifications alone. As an example, consider Figure 1, where 44 concepts from Swadesh's list of 100 concepts (Swadesh, 1955) are plotted along two dimensions based on their embeddings learned from both full colexifications (left) and full colexifications combined with affix colexifications (right, created using t-SNE). While TREE and BARK are far apart in the embedded space learned from full colexifications alone, they appear very close to each other when affix colexifications are considered as well. This nicely reflects that words for BARK are often compounds that contain TREE as a member, indicating a clear conceptual relation, although the two concepts are rarely fully colexified.

It is also worthwhile to discuss the role of overlap colexifications in this context. While affix colexifications are often clearly motivated by transparent semantic relations (such as the meronymy above), the relationship between concepts sharing an overlap colexification is often more remote. In fact, overlap colexifications usually represent what would be a two-hop relationship in the affix colexification graph (e.g. *handbook ← hand → handful*), and the direct edges represent two forms being linked by a shared morpheme. The semantic relationships expressed by overlap colexification might therefore simply be too remote to accurately reflect narrow semantic relationships like perceived similarity. On the other hand, the inclusion of overlap colexifications still proved useful for the other two tasks, albeit only by a bit. They can also capture formal relationships between forms when the linking item is not found in the data itself – either because it is a lexical root that is not found in basic
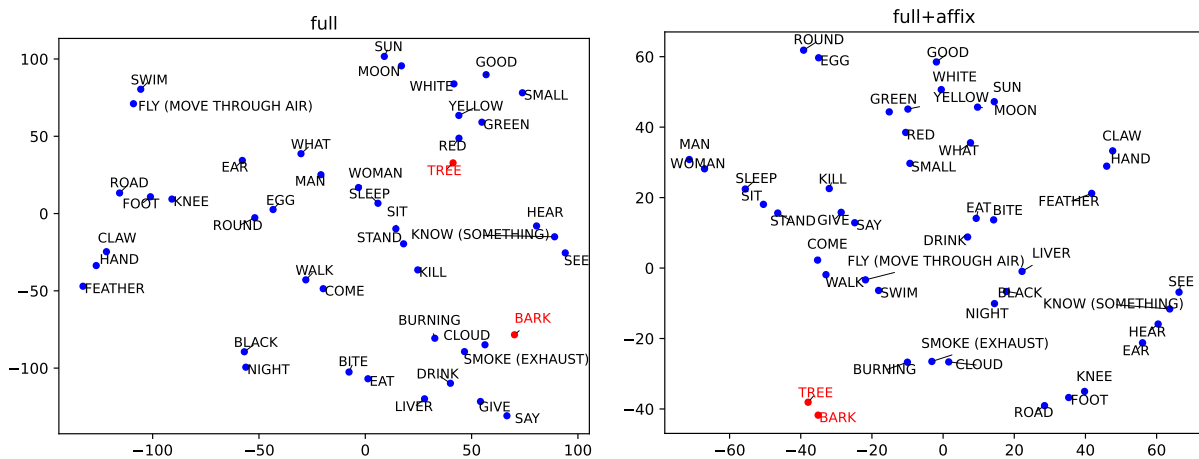
Figure 1: Two-dimensional visualizations of embeddings only learned on full colexification data (left) and embeddings learned on full and affix colexifications (right), created using t-SNE, for a small list of concepts taken from the Swadesh list of 100 items (Swadesh, 1955).

vocabulary lists, or because it is a bound morpheme that naturally cannot occur on its own (compare the verbal infinitive suffixes in Romance languages). For now, it remains an open question what kind of conceptual relationships are actually represented by overlap colexifications, and to which degree they can be used to train richer and more accurate concept embeddings. There is no doubt, however, that the inclusion partial colexifications in general is highly beneficial for learning concept embeddings.

So far, related studies have only considered full colexifications that have been inferred from multilingual corpus data (Liu et al., 2023; Harvill et al., 2022; Chen et al., 2023). Our findings imply that corpus-based approaches could benefit from looking into partial colexifications in order to improve the representation of semantic relations across multiple languages. However, inferring morphological structures from corpus data still poses a challenge to which a widely applicable solution has not yet been found (Manova et al., 2020).

Since our embeddings are based on colexifications attested in multilingual wordlists rather than being automatically inferred from texts they suffer less from noise introduced during colexification inference, resulting in higher-quality concept embeddings. The drawback of this approach is, however, a lower coverage resulting from smaller networks. With growing amounts of standardized multilingual wordlists (List et al., 2022; Blum et al., 2025), we are confident that the conceptual space covered by colexification networks – as well as by the resulting embeddings – will increase in the future. Being able to model meaning independently of individ-

ual languages in a continuous space could advance computational approaches to historical linguistics and linguistic typology in many ways, providing new possibilities for automatic cognate detection (Wu et al., 2020) and the reconstruction of semantic change scenarios (Urban, 2015).

Supplementary Material

Acknowledgments

Funding

Limitations

The major and most obvious limitation of the present approach is that it currently only covers around 1,000 core concepts, due to the fact that the underlying colexification networks are mostly inferred from comparative wordlists for basic vocabulary. While the restriction to basic concepts is a common practice in historical linguistics and linguistic typology, it poses an obstacle to direct applications on multilingual tasks in the context of NLP applications. While we plan to expand our approach to full and partial colexifications derived from larger datasets, such as Lexibank 2 (Blum et al., 2025) and CLICS 4 (Tjuka et al., 2025), these can only add a couple of hundred more concepts to the sample. As of now, the application range of our approach is therefore restricted to the fields of computational historical linguistics and computational linguistic typology.

The graph embedding techniques discussed in this paper rely on the graph topology alone: disconnected nodes (concepts with no attested colexifications), therefore, cannot be embedded with this approach. To circumvent this problem of missing data in downstream applications, future approaches might benefit from using node features (e.g. part-of-speech) on top of the graph topology. Recent advances in Graph Neural Networks have sucessfully shown the potential of combining these two pieces of information (e.g. Hamilton et al., 2017).

Finally, while our results clearly show that affix colexifications provide crucial information complementary to what is encoded in a full colexification network, the role of overlap colexifications remains unclear. Further research is required to better understand what kind of semantic and conceptual relations different types of colexification represent.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

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. Association for Computational Linguistics.

Emily Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 5185–5198, Online. Association for Computational Linguistics.

Frederic Blum, Johann-Mattis List, Carlos Barrientos, Johannes Englisch, Robert Forkel, Simon J. Greenhill, and Christoph Rzymski. 2025. Lexibank 2: precomputed features for large-scale lexical data [version 1; peer review: awaiting peer review]. *Open Research Europe*, 5(126):1–19.

Katja Bocklage, Anna Di Natale, Annika Tjuka, and Johann-Mattis List. 2024a. Directional tendencies in semantic change. *Humanities Commons*, 2024(2):1–28.

Katja Bocklage, Anna Di Natale, Annika Tjuka, and Johann-Mattis List. 2024b. Representing the Database of Semantic Shifts by Zalizniak et al. from 2024 in Cross-Linguistic Data Formats. *Computer-Assisted Language Comparison in Practice*, 7(4):25–35.

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

José Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, BC. Association for Computational Linguistics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. NASARI: a novel approach to a semantically-aware representation of items. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 567–577, Denver, CO. Association for Computational Linguistics.

Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2015. GraRep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 891–900, New York, NY. Association for Computing Machinery.

Yiyi Chen, Russa Biswas, and Johannes Bjerva. 2023. Colex2Lang: Language embeddings from semantic typology. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 673–684, Tórshavn. University of Tartu Library.

Davide Colla, Enrico Mensa, and Daniele P. Radicioni. 2020. LessLex: Linking multilingual embeddings to senSe representations of LEXical items. *Computational Linguistics*, 46(2):289–333.

Simone Conia and Roberto Navigli. 2020. Conception: Multilingually-enhanced, human-readable concept vector representations. In *Proceedings of the 28th International Conference of Computational Linguistics*, pages 3268–3284, Barcelona (Online). International Committee on Computational Linguistics.

Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. 2019. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):833–852.

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

Ferdinand de Saussure. 1916. *Cours de linguistique générale [Course on general linguistics]*. Payot, Lausanne.

Johannes Dellert. 2024. Causal inference of diachronic semantic maps from cross-linguistic synchronic polysemy data. *Frontiers in Communication*, 8(1288196):1–15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN. Association for Computational Linguistics.

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

Edsger W. Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.

Philipp Dufter, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze. 2018. Embedding learning through multilingual concept induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1520–1530, Melbourne. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An electronical lexical database*. MIT Press, Cambridge, MA.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford. Reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

Ilya Flyamer. 2024. adjustText - automatic label placement for matplotlib. Python package, version 1.3.0.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205):1–10.

Robert Forkel, Christoph Rzymski, and Sebastian Bank. 2025. *PyCLDF [Software, Version 1.40.4])*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Robert Forkel, Christoph Rzymski, and Johann-Mattis List. 2024. *PyConcepticon [Python library, Version 3.1.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Alexandre François. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Martine Vanhove, editor, *From polysemy to semantic change*, pages 163–215. Benjamins, Amsterdam.

Palash Goyal and Emilio Ferrara. 2018. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, San Francisco, CA. Association for Computing Machinery.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort. JMLR Workshop and Conference Proceedings.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pages 11–15, Pasadena, CA.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1025–1035, Red Hook, NY. Curran Associates Inc.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

John Harvill, Roxana Girju, and Mark Hasegawa-Johnson. 2022. Syn2vec: Synset colexification graphs for lexical semantic similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5270, Seattle, WA. Association for Computational Linguistics.

John D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing. Association for Computational Linguistics.

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

Fatemeh Jafarinejad. 2023. Synset2node: A new synset embedding based upon graph embeddings. *Intelligent Systems with Applications*, 17(200159):1–9.

Taelin Karidi, Eitan Grossman, and Omri Abend. 2024. Aligning alignments: Do colexification and distributional similarity align as measures of cross-lingual lexical alignment? In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 327–341, Miami, FL. Association for Computational Linguistics.

Marie Ritchie Key and Bernard Comrie. 2016. *The Intercontinental Dictionary Series [Dataset]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Thomas N. Kipf and Max Welling. 2016. Variational graph auto-encoders. In *Bayesian Deep Learning Workshop*, NIPS 2016, Barcelona.

George R. Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of English and its computer analysis. In A.J. Aitken, R.W. Bailey, and N. Hamilton-Smith, editors, *The computer and literary studies*, pages 153–165. Edinburgh University Press, Edinburgh.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC. ICLR.

Omer Levy, Anders Søgaard, and Yoav Goldberg. 2016. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 765–774, Valencia. Association for Computational Linguistics.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031.

Johann-Mattis List. 2021. Mapping Multi-SimLex to Concepticon. *Computer-Assisted Language Comparison in Practice*, 4(3):1–8.

Johann-Mattis List. 2022. How to compute colexifications with CL Toolkit (How to do X in Linguistics 10). *Computer-Assisted Language Comparison in Practice*, 5(6):1–6.

Johann-Mattis List. 2023. Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology*, 14(1156540):1–10.

Johann-Mattis List and Robert Forkel. 2024. *PyNoRaRe [Python library, Version 1.1.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):1–16.

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 help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306.

Johann-Mattis List, Anselm Terhalle, and Matthias Urban. 2013. Using network approaches to enhance the analysis of cross-linguistic polysemies. In *Proceedings of the 10th International Conference on Computational Semantics - Short Papers*, pages 347–353, Potsdam. Association for Computational Linguistics.

Johann-Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill, and Robert Forkel, editors. 2025. *CLLD Concepticon 3.3.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Renhao Pei, and Hinrich Schütze. 2023. Crosslingual transfer learning for low-resource languages based on multilingual colexification graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8376–8401, Singapore. Association for Computational Linguistics.

Stela Manova, Harald Hammarström, Itamar Kastner, and Yining Nie. 2020. What is in a morpheme? Theoretical, experimental and computational approaches to the relation of meaning and form in morphology. *Word Structure*, 13(1):1–21.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Alla Münch and Johannes Dellert. 2015. Evaluating the potential of a large-scale polysemy network as a model of plausible semantic shifts. In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*, Tübingen. Universität Tübingen.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. 2016. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1105–1114, New York, NY. Association for Computing Machinery.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., Vancouver, BC.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha. Association for Computational Linguistics.

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*, KDD '14, page 701–710, New York, NY. Association for Computing Machinery.

Mohammad Taher Pilehvar and Nigel Collier. 2016. Deconflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1680–1690, Austin, TX. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Matt Ranger. 2024. nodevectors. Python package, version 0.2.0.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1793–1803, Beijing. Association for Computational Linguistics.

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. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):13.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, Ann Arbor, MI. Association for Computational Linguistics.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, volume 1, San Francisco, CA. AAAI Press.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

Annika Tjuka. 2023. The origins of cross-linguistic colexifications. *Computer-Assisted Language Comparison in Practice*, 6(1):1–9.

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2022. Linking norms, ratings, and relations of words and concepts across multiple language varieties. *Behavior Research Methods*, 54(2):864–884.

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2024. Universal and cultural factors shape body part vocabularies. *Scientific Reports*, 14(10486):1–12.

Annika Tjuka, Robert Forkel, Christoph Rzymski, and Johann-Mattis List. 2025. Advancing the Database of Cross-Linguistic Colexifications with new workflows and data. *arXiv*, 2503(11377):1–14.

Andrew Trask, Phil Michalak, and John Liu. 2015. sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388*.

Elizabeth Closs Traugott and Richard B. Dasher. 2002. *Regularity in semantic change*. Cambridge University Press, Cambridge.

Matthias Urban. 2011. Asymmetries in overt marking and directionality in semantic change. *Journal of Historical Linguistics*, 1(1):3–47.

Matthias Urban. 2015. Lexical semantic change and semantic reconstruction. In Claire Bowern and Bethwyn Evans, editors, *The Routledge Handbook of Historical Linguistics*, pages 374–392. Routledge, London.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY. Curran Associates Inc.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, Christoph J. Carey, İbrahim Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897.

Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1225–1234, New York, NY. Association for Computing Machinery.

Mei-Shin Wu, Nathanael E. Schweikhard, Tim Bodt, Nathan W. Hill, and Johann-Mattis List. 2020. Computer-assisted language comparison: State of the art. *Journal of Open Humanities Data*, 6(2):1–14.

Dongqiang Yang, Ning Li, Li Zou, and Hongwei Ma. 2022. Lexical semantics enhanced neural word embeddings. *Knowledge-Based Systems*, 252(109298):1–13.

Anna Zalizniak, Anna Smirnitskaya, Maksim Russo (Rousseau), Ilya Gruntov, Timur Maisak, Dmitry Ganenkov, Maria Bulakh, Maria Orlova, Marina Bobrik-Fremke, Oksana Dereza, Tatiana Mikhailova, Maria Bibaeva, and Mikhail Voronov. 2024. *Database of Semantic Shifts [Dataset, Version 3.0]*. Institute of Linguistics, Russian Academy of Sciences, Moscow.

Jie Zhang, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding. 2019. ProNE: Fast and Scalable Network Representation Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 4278–4284, Macao. International Joint Conferences on Artificial Intelligence Organization.

A Visualizations

Figure 2 shows shows two-dimensional visualizations of all different combinations of full and partial colexifications tested in our study for a small list oc concepts taken from the Swadesh list of 100 items (Swadesh, 1955).
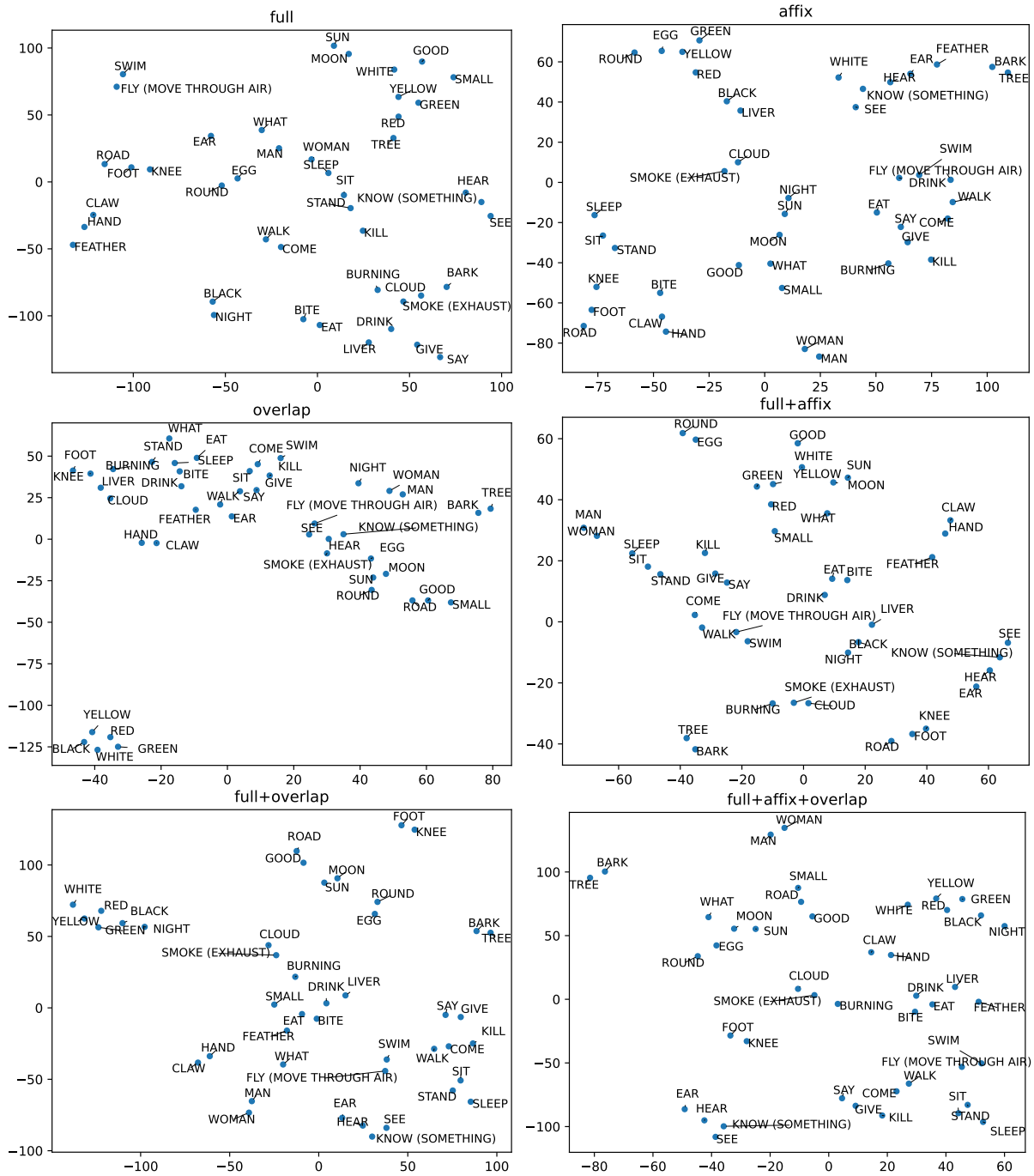

Figure 2: Two-dimensional visualization of embeddings for concepts from the Swadesh-100 list (Swadesh 1955) that are present in all three colexification networks, created using t-SNE.

B Hyperparameter Choice

The models discussed in the paper were trained on an Apple M2 Pro CPU with the following hyperparameters:

Model	Hyperparameters	Training Time
SDNE	hidden sizes $= (256, 128), \alpha = 0.2, \beta = 10$, learning rate $= 0.001, \text{epochs} = 10000$.	02:00 min
Node2Vec	n(number of random walks per start node)=5, w(length of random walks) $= 10, p = 1, q = 1$, c(context window size) $= 2, \text{test split} = 0.2$, w(length of random walks) $= 10, p = 1, q = 1$, learning rate $= 0.001, \text{epochs} = 1500$. Underlying Word2Vec model: SkipGram, no negative sampling.	13:27 min
ProNE	step $= 10, \mu = 0.2, \theta = 0.5, \text{exponent} = 0.75$.	38 ms