# Partial Colexifications Improve Concept Embeddings

*Arne Rubehn & Johann-Mattis List*

Chair for Multilingual Computational Linguistics, University of Passau, Germany

# Introduction

***We learn meaningful, low-dimensional representations of concepts by applying graph embedding techniques on colexification networks.***

## Motivation

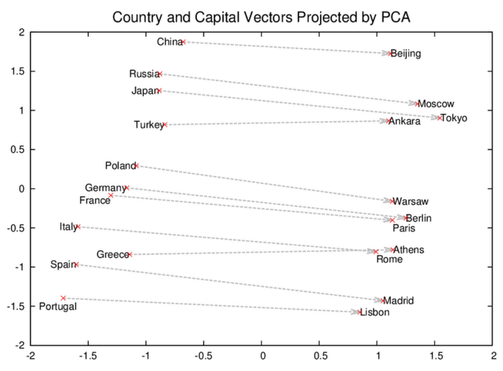
- State-of-the-art methods in computational historical linguistics are generally unable to model or reconstruct semantic change
- It is notoriously hard to quantify meaning cross-linguistically
- Modeling semantic relations between *concepts* reliably in a "computer-friendly" way could help bridging the gap

## Background

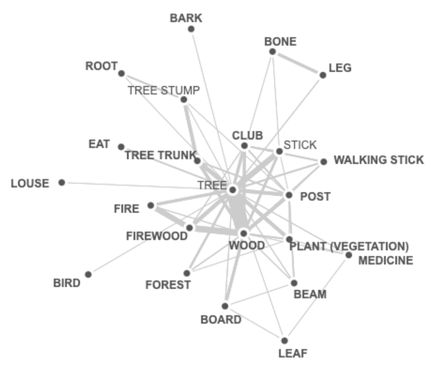
- **Word embeddings** have shown that various semantic relationships can be efficiently modeled in a low-dimensional vector space.
  - ...but since they embed words, they are not fit for wide cross-linguistic applications

- **Colexification networks** offer a cross-linguistic, concept-based approach on semantics, encoding common pathways of semantic shift
  - ...but their network structure can not readily be processed by downstream applications

**Using graph embedding techniques, we can learn embedded representations for concepts in colexification networks.**




(Mikolov et al., 2013)




(Rzymski et al., 2020)

# Materials and Methods

## Colexification Data

- Three types of colexification (List, 2023) inferred from the *Intercontinental Dictionary Series* (Key and Comrie, 2016):
  - full colexification
  - affix colexification
  - overlap colexification
- Colexification network for each type of colexification
- Concepts defined by the CLLD Concepticon (List et al., 2025)

## Graph Embedding Techniques

- ProNe (Zhang et al., 2019), Node2Vec (Grover and Leskovec, 2016) & SDNE (Wang et al., 2016)

## Training

- Train individual embeddings for each colexification network
- Combine embeddings from different colexification types as a post-processing step (concatenation + PCA)

# Experiments

## 1) Modeling Lexical Semantic Similarity

- Multilingual similarity ratings between word pairs, obtained from MultiSimLex (Vulić et al., 2020)
- Calculate cosine similarities between corresponding concept pairs
- Calculate Spearman's r between similarities

## 2) Predicting Semantic Change

- Obtain historically attested semantic changes from DatSemShift (Zalizniak et al., 2024)
- Negative sampling of random "changes"
- Train simple logistic regression classifier to tell apart true and false shifts, based on the respective concept similarities

## 3) Predicting Word Associations

- Obtain word association data from the Edinburgh Association Thesaurus (Kiss et al., 1973)
- Same experimental setup as previous task: Negative sampling, prediction with simple logistic regression classifier

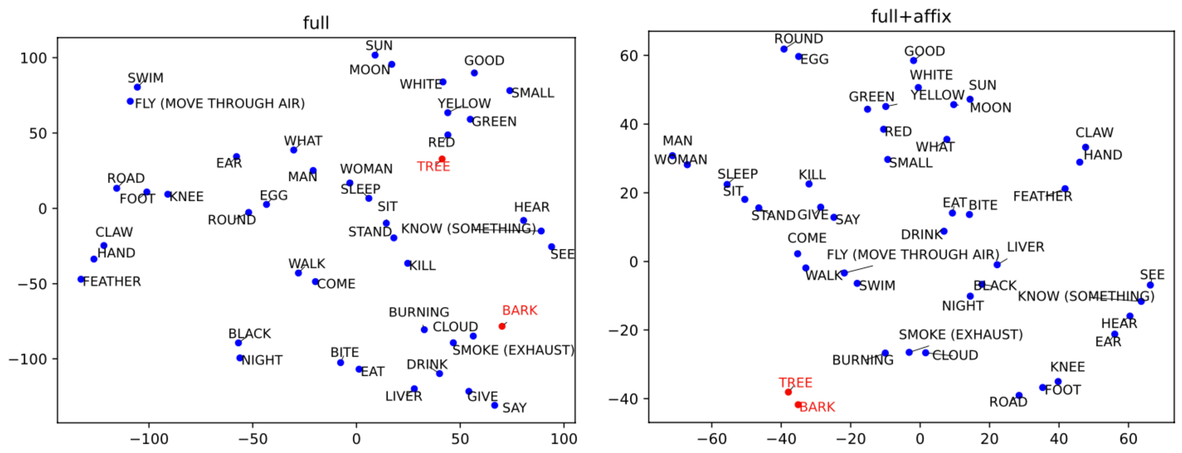
## Baselines

- Various similarity metrics inferred from the graph directly
- Multilingual fastText vectors (Grave et al., 2018)

# Results & Discussion

## Results

- Embeddings learned with ProNE perform the best on average, closely followed by Node2Vec; SDNE does not seem viable
- Concept embeddings outperform graph-based baselines in all three tasks, and fastText embeddings in 2 of 3 tasks
- Almost identical patterns between tasks 2 and 3
- Affix colexifications lead to better embeddings on all three tasks
- Overlap colexifications are beneficial for predicting semantic change and word association, but detrimental for modeling lexical similarity (the same pattern holds for fastText embeddings!)




## Discussion

- Partial colexifications can capture semantic relations that are rarely expressed by full colexification
- Affix colexification seems to capture a more direct relationship between concepts than overlap colexification

**Enriching embeddings with partial colexification data always leads to better results than relying on full colexification data alone!**

**References:** Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. · Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. · Key, M. R., & Comrie, B. (2016). *The Intercontinental Dictionary Series [Dataset]*. URL: https://ids.clld.org · Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). *An associative thesaurus of English and its computer analysis*. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), The computer and literary studies (pp. 153–165). Edinburgh: Edinburgh University Press. · List, J.-M. (2023). Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology*, 14(1156540), 1–10. · List, J.-M., Tjuka, A., Blum, F., Kučerová, A., Ugarte, C. B., Rzymski, C., … Forkel, R. (Eds.). (2025). *CLLD Concepticon 3.3.0*. URL: https://concepticon.clld.org/ · Mikolov, T., Yih, W.-T., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. · Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., … Others. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(1), 13. · Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., … Korhonen, A. (2020). Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4), 847–897. · Wang, D., Cui, P., & Zhu, W. (2016). Structural Deep Network Embedding. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1225–1234. · Zalizniak, A., Smirnitskaya, A., Russo (Rousseau), M., Gruntov, I., Maisak, T., Ganenkov, D., … Voronov, M. (2024). *Database of Semantic Shifts [Dataset, Version 3.0]*. URL: https://datsemshift.ru · Zhang, J., Dong, Y., Wang, Y., Tang, J., & Ding, M. (2019). ProNE: Fast and Scalable Network Representation Learning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 4278–4284. **Acknowledgements:** This study was supported by the ERC Consolidator Grant ProduSemy (Grant No. 101044282).

European Research Council
Established by the European Commission

arne.rubehn.com

arne.rubehn@uni-passau.de