Towards Cross-Linguistic Semantic Grounding using Dictionary Graph Analysis

Anonymous ACL submission

Abstract

Previous work has explored the structure of dictionaries as directed graphs, with arcs between words when one word is used in the definition of another. We analyze the efficacy of these methodologies and explore the cross-linguistic patterns of the strongly connected components of multiple monolingual dictionaries. We find that the number of sources in the condensation graph of a directed dictionary graph is roughly stable across multiple different languages, and present future research directions.

1 Introduction

Explanatory dictionaries are an important tool for understanding lexical semantics. However, in order for lexical meaning to connect to real-world senses, not all meanings can be defined purely in terms of words; at least some words must be defined outside of the language in terms of sensorimotor experience. This observation is referred to as the symbol grounding problem (Harnad, 1990). Some theories, especially in cognitive semantics, solve this problem by choosing specific words or concepts that are considered fundamental (e.g., Semantic Primes (Wierzbicka, 1996)). One empirical approach towards this problem is to analyze dictionary structures, modeling them as directed graphs (e.g., Kostiuk et al. (2023)). As far as the authors are aware, there are two major approaches in the literature to analyzing these dictionary graphs for the purposes of semantic grounding.

The first approach considers Feedback Vertex Sets (FVS's). For a directed graph D, a Feedback Vertex Set is a set of vertices $F \subseteq V(D)$ such that $D \setminus F$ is acyclic. The Minimum Feedback Vertex Set Problem consists of finding an FVS that is minimum with respect to cardinality. In the context of semantic grounding, these sets have a convenient theoretical interpretation: if words from an FVS are removed, the dictionary becomes "grounded", i.e. there are no self-referential definitions.

The second approach considers the dictionary structure through strongly connected components, or SCCs (Vincent-Lamarre et al., 2016). For a directed graph D, a SCC is a maximal vertex set $S \subseteq V(D)$ such that there exists a directed path in D between every pair of vertices in S. The condensation of a graph can be thought of as the graph obtained by contracting each SCC into a single vertex. SCCs partition a directed graph into equivalence classes, and the corresponding condensation graph is acyclic. Thus this asks how many groups of equivalent words must be removed for the graph to be acyclic. Vincent-Lamarre et al. presented a taxonomy of the dictionary latent structure, with the sources in the condensation graph called the "core", and all other non-trivial SCCs referred to as "satellites". They also analyzed the psycholinguistic correlates of the words at various levels of the latent structure, finding words in the core to be more frequent, less concrete, and learned earlier than those in the satellites. Thus, the sources in the condensation graph occupy a fundamental role in the dictionary's structure.

While FVS's are more directly connected with the issue of grounding a dictionary, by removing self-referential definitions, there are major downsides. The minimum FVS Problem is NP-Hard (Karp, 1972), and the minimum sizes scale with the dictionary (Vincent-Lamarre et al., 2016). Since there are often different equivalent FVS's, it is difficult to compare structure cross-linguistically. By contrast, the SCCs of a digraph are unique and efficient to compute. They consider groups of self-referential words, and thereby remove arbitrary choice, facilitating cross-linguistic comparison.

With that in mind, this study explores the SCCs approach. In contrast to prior literature that focused only on English (Kostiuk et al. 2023, Vincent-Lamarre et al. 2016) or Spanish (Pichardo-Lagunas et al., 2017), we will analyze and compare English, French, German, Mandarin, Russian, and Spanish.

2 Methods

Monolingual dictionaries were acquired[1] from the Wiktionaries for English, French, German, Mandarin, Russian, and Spanish using Wiktextract (Ylonen, 2022). We limited analyses to content words by filtering for entries with a part of speech tag of noun, verb, adjective, or adverb, and with the python library `stopwordsiso` to remove function words. The definitions for all word senses for each entry were tokenized and lemmatized by STANZA (Qi et al., 2020).

The dictionaries were processed into directed graphs. Each headword was treated as a vertex, and an arc was added from vertex u to vertex v if the wordform u was included in at least one definition of v. For undefined words used within a definition, an arc from the lemma form was added, and if the lemma was not present, the word was excluded.

The final dictionary directed graph was preprocessed. All leaves (vertices with no outgoing arcs) were removed recursively, since they were unused in definitions and not directly relevant for the analysis. This removed all trivial SCCs. We built the condensation graph of the directed dictionary graph using the built-in function from `networkx` (Hagberg et al., 2008), and finally extracted the sources from the condensation graph.

3 Results and Discussion

From each of the six monolingual dictionaries, we found the condensation graphs and sources within those graphs. The details on these experiments are outlined in Table 1, including the overall size of the dictionary graph for each language and the number of sources in the condensation graph.

Language	Order	Number of Sources
English	1053726	77
French	1849021	39
German	843506	65
Mandarin	25736	648
Russian	408173	134
Spanish	746297	29

Table 1: Number of wordforms in preprocessed dictionary graph, and number of sources in the condensation graph, for each language.

Observe that, overall, the number of sources in the condensation graphs are relatively close cross-linguistically. Mandarin appears to be an outlier, with 648 sources; however, it was the smallest dictionary by far with only 25736 words in total. Without Mandarin, the number of sources in the remaining 5 languages have a mean of 68.8 with a standard deviation of 36.9. Also note that as the size of the dictionary increases, the number of sources declines. Additionally, the rate at which the number of sources declines with respect to dictionary size is not constant. In fact, it appears to decrease, as illustrated in Figure 1.

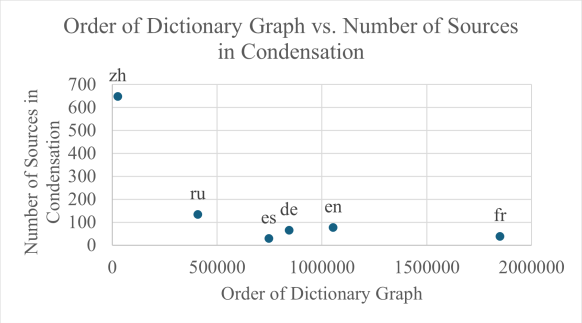

Figure 1: Scatter plot demonstrating the overall trend of fewer sources in the condensation given the order.

The results above suggest that for sufficiently large dictionaries, the number of sources in the condensation graph of a dictionary are somewhat consistent cross-linguistically. However, the data is by no means conclusive. Wiktionary was used because of their large dictionary sizes, unified data format for multiple different languages, and accessibility. However, professionally curated dictionaries would have higher data quality, which would make the results more conclusive. Additionally, the sizes of the dictionaries vary, with Mandarin about 1% the size of the English dictionary. More consistent dictionary sizes, or an approach to control for the affect of dictionary size, could provide a clearer analysis.

The methodology used can also be improved. Dictionary conversion is imperfect, ignoring undefined words in definitions, resulting in information loss. Additionally, words are connected regardless of which word sense is used. Thus, the definitional paths integral to this method could rely on two unrelated word senses, limiting the efficacy of SCC analysis. Finally, the approach ignores morphological complexity. Either the inflected wordform is present in the dictionary, or only the lemma is kept. Morphological parsing would prevent losing inflectional information, and help with consistency across typologically diverse languages. We explore these limitations in ongoing work.

[1] The dictionaries were accessed on 7/20/2024.

References

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 USA.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Richard M. Karp. 1972. *Reducibility among Combinatorial Problems*, pages 85–103. Springer US, Boston, MA.

Yevhen Kostiuk, Obdulia Pichardo-Lagunas, Anton Malandii, and Grigori Sidorov. 2023. Automatic detection of semantic primitives using optimization based on genetic algorithm. *PeerJ Comput Sci*, 9:e1282.

Obdulia Pichardo-Lagunas, Grigori Sidorov, Alexander Gelbukh, Nareli Cruz-Cortés, and Alicia Martínez-Rebollar. 2017. Automatic detection of semantic primitives with bio-inspired, multi-objective, weighting algorithms. *Acta Polytechnica Hungarica*, 14(3):113–128.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Philippe Vincent-Lamarre, Alexandre Blondin Massé, Marcos Lopes, Mélanie Lord, Odile Marcotte, and Stevan Harnad. 2016. The latent structure of dictionaries. *Topics in Cognitive Science*, 8(3):625–659.

Anna Wierzbicka. 1996. *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*, pages 367–377. Almquist & Wiksell.