Towards Cross-Linguistic Semantic Grounding using Dictionary Graph Analysis

Ethan Eschrich and **Zoey Liu**
Department of Linguistics
University of Florida
Gainesville, FL
{ethan.eschrich, liu.ying}@ufl.edu

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 for analyzing semantic grounding 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 languages, and present future research directions.

1 Introduction

Explanatory dictionaries are an important tool for lexical semantics. However, to connect lexical meaning to real-world senses, not all meanings can be defined in terms of words; some words must be defined outside of the language in terms of sensorimotor experience. This observation is the symbol grounding problem (Harnad, 1990). Some theories, especially in cognitive semantics, solve this problem by considering specific words or concepts as fundamental within a language and cross-linguistically (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)).

There are two major approaches for analyzing dictionary graphs. The first approach considers Feedback Vertex Sets (FVS's) (Kostiuk et al., 2023). 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. For 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 is 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, the condensation graph captures the structure between groups of "equivalent" words, and the sources (i.e., vertices with no incoming arcs) represent ungrounded groups. Vincent-Lamarre et al. presented a taxonomy of the dictionary latent structure in this manner, with the sources in the condensation graph called the "core" [1], 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 core occupies a fundamental role in the dictionary's structure.

While FVS's can be more directly interpreted as 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). FVS's are not unique, so we must arbitrarily choose one for comparison. 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.

This study utilizes the SCCs approach to identify common structure of monolingual dictionaries

[1] Vincent-Lamarre et al. described the taxonomy in alternate but equivalent terminology.

to lend credence to the cross-linguistic aims of cognitive semantics theories. 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 analyze and compare English, French, German, Mandarin, Russian, and Spanish.

2 Methods

We acquired monolingual dictionaries[2] from the Wiktionaries for English, French, German, Mandarin, Russian, and Spanish using Wiktextract (Ylonen, 2022), based on their availability of parsed data. We limited our analyses to content words by filtering for entries with a part of speech tag of either 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. Table 1 presents relevant descriptive statistics, including the overall size of the dictionary graph for each language and the number of sources in the condensation graph.

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 25,736 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.

Language	Order	Number of Sources
English	1,053,726	77
French	1,849,021	39
German	843,506	65
Mandarin	25,736	648
Russian	408,173	134
Spanish	746,297	29

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

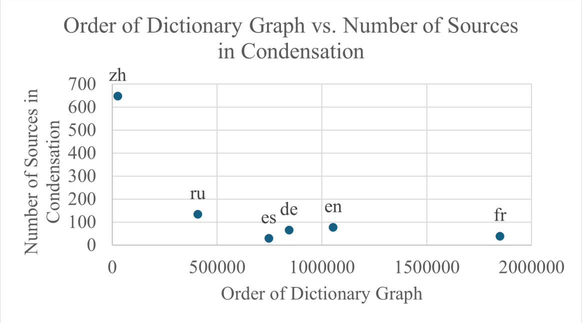

Figure 1: Scatter plot demonstrating the overall trend of fewer sources in the condensation given the order.

These results suggest that for sufficiently large dictionaries, the number of sources in the condensation graph are consistent cross-linguistically. Thus, the number of groups of "fundamental" words for grounding are similar, supporting Semantic Prime theory. While Wiktionary has large dictionary sizes, a unified format, varied selection, and accessibility, professionally curated dictionaries would provide more conclusive results. Additionally, the variation of dictionary size (Mandarin ~ 1% of English) could impact condensation graph structure; more consistent dictionary sizes, or an approach to control for the size, could improve results.

Dictionary conversion ignores undefined words and the differences of word senses, limiting both the number and reliability of connections. The conversion also ignores morphological complexity, using either the inflected wordform or solely the lemma. Morphological parsing would prevent losing inflectional information when not present within the dictionary, and help with consistency across typologically diverse languages.

[2] 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.