Dependency-based Labeling of Associative Lexical Communities

Benedikt Perak
Faculty of Humanities and Social Sciences,
University of Rijeka
Sveučilišna Avenija 4, 51000 Rijeka, Croatia
bperak@uniri.hr

Tajana Ban Kirigin
Department of Mathematics,
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
bank@math.uniri.hr

Abstract. *The article presents a graph-based method for automatically labeling senses associated with a polysemous lexeme. For a source lexeme, clusters of conceptually associated lexemes are formed, representing related senses and conceptual domains. The labeling task aims to generalize the sense features using a more abstract concept, with the label of a lexical sense community being the most appropriate hypernym category. Label abstraction is achieved by selecting hypernym candidates through syntactic patterns of the is_a relation. The most prominent is_a collocates of clustered lexemes are extracted from the corpus and selected from the constructed label-graph as labels for the sense community.*

Keywords. Lexical graph analysis; corpus; knowledge representation and reasoning; coordination; word sense; hypernym.

1 Introduction

This paper presents a graph-based method that automatically labels sense classes of a source lexeme based on syntactic-semantic collocations of the is_a relation in a corpus. The method, implemented in the ConGraCNet 2.x web application (available at www.emocnet.uniri.hr), projects a semantic function onto a particular syntactic relation and constructs a lexical dependency graph that can be queried for a range of lexical tasks and common-sense knowledge.

In particular, we study the structures of the dependency graph to clarify the labeling of lexemes that exhibit lexical ambiguity and semantic change. Namely, a word can have multiple senses or acquire new meanings that are often not even semantically related. For example, the noun lexeme *bass* may refer to a *musical instrument* or a type of *fish*. This ambiguity problem therefore implies a non-trivial assignment of hypernym labels.

To address the problem of lexical polysemy and disambiguation, we developed a dependency-based graph method for distinguishing lexical associations and label assignment. This method allows label assignment for polysemous lexemes using the best hypernym candidates for each associated sense. More specifically, the ConGraCNet method generates a lexical graph that is clustered into subgraphs that reveal the polysemous semantic nature of a lexeme in a corpus. A hypernym graph is then constructed to identify a set of synset labels for a lexical cluster in the ConGraCNet graph of the source lexeme. The labeling is based on a graph model representing the corpus-based *is_a* relation of iteratively computed local clusters of conceptually associated lexemes of a source word. The labeling extracts hierarchically abstract conceptual content that can be used to facilitate lexical understanding as well as to build corpus-based taxonomies. This work draws on data obtained from the Sketch Engine English corpora with preprocessed [word] is_a … dependency. This labeling task falls into the class of tacit knowledge encoding (Prince, 1978) and ontological representation, which is important for several subsequent natural language processing (NLP) tasks and applications (Hovy et al., 2013), such as intelligent personal assistants, question answering, information retrieval, etc.

The aim of this paper is to: (i) present a graph-based method for labeling hypernym semantic representations of associative concepts; (ii) present a semantic resource for web app that integrates manually annotated lexicons and semi-automatic corpus techniques. We provide the following contributions:

1. A graph-based approach to associative lexical cluster labeling using a combination of coordination and *is_a* syntactic-semantic lexical relations from Sketch Engine (Kilgarriff et al., 2014);
2. A web app implementation of the graph-based labeling algorithm.

We present examples from the English corpora and describe the construction of the graph, the candidate selection process using centrality measures, and present the ConGraCNet web app, available as a semantic resource at www.emocnet.uniri.hr, which integrates manually annotated lexicons and semi-automatic corpus resources and methods.

The paper begins with a description of related work, Section 2. The labeling method is described in Section 3 and then discussed in Section 4, where we also present potential future research directions. Conclusions are drawn in Section 5.

2 Related Work

Models for the computational representation of hypernym lexical-semantic knowledge have their origins in methods and resources that can be broadly divided into two categories.

Top-down curated knowledge databases such as WordNet (Miller, 1995) and its counterparts in other languages (Bond and Foster, 2013) form the basis for computational lexicons and contextualization of paradigmatic hypernymy relations. WordNet lexical synsets, and later VerbNet (Schuler, 2005), PropBank (Kingsbury and Palmer, 2002), BabelNet (Navigli and Ponzetto, 2010) and VerbAtlas (Di Fabio et al., 2019) encode lexical semantic knowledge using word senses as units of meaning. A major problem with wordnets is the curated top-down structure of resource creation, which inevitably leads to a lower granularity and static nature of the inventories.

This class of resources also includes Common-Sense Knowledge (CSK) databases, which store descriptions of a set of common and generic facts or views of a set of concepts, including *is_a* relations. They describe the general information that people use to describe, differentiate, and reason about concepts. ConceptNet (Speer and Havasi, 2012; Speer et al., 2016) is one of the largest such resources, integrating data from the original MIT Open Mind Common Sense project. Unlike WordNet, which discriminates senses of a given lemma, terms in ConceptNet are not disambiguated, which can lead to confusion in the hypernym lexical-semantic relations for concepts denoted by ambiguous words (*e.g.*, bass as an instrument vs. a type of fish).

On the other hand, bottom-up approaches to semantic labeling rely on the extraction of semantic features from the syntagmatic idea that similar words are used in similar contexts (Harris, 1954) and the use of corpus-based syntactic pattern analysis (Hanks, 2004, 2013). The underlying idea is to analyze the prototypical syntagmatic patterns of words in use in large corpora and to attribute meaning on a contextual basis through prototypical sentence patterns. Automatic approaches such as those of Baroni et al. (2010), Navigli and Velardi (2010) and Boella and Di Caro (2013) use syntactic patterns for automatic extraction of concept descriptions.

A radically different bottom-up approach is based on vector space models of lexical representations that view concepts as geometric vectors whose dimensions are qualitative features (Gardenfors, 2004) and other similar methods such as Latent Semantic Analysis (Dumais, 2004), Latent Dirichlet Allocation (Blei et al., 2003), embeddings of words (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016) and word senses (Huang et al., 2012; Iacobacci et al., 2015; Scarlini et al., 2020). Most of the recent approaches have been modified with the introduction of the bidirectional open-source machine learning framework that uses the surrounding text to determine the context of words. These models allow direct similarity computation, but the knowledge does not explicitly define the concepts and the relations between vector representations are not ontologically organized. There are also some efforts to extract tacit human knowledge (Petroni et al. 2020; Kavumba et al. 2021; Weir et al. 2020; Roberts et al. 2020).

Finally, there are mixed methods, that use resources and methods from different approaches, such as the semagram-based knowledge model, which consists of 26 semantic relations and integrates features from different sources (Leone et al. 2020), or the multilingual label propagation scheme introduced by Barba et al. (2020), which leverages word embeddings and the multilingual information from a knowledge database.

In Croatian, computational word disambiguation and lexical labeling was investigated by Alagić and Šnajder (2016), who performed a work in their medium-sized lexical sample dataset Cro36WSD, constructed by multiple annotation. On the other hand, the graph method for distinguishing lexical senses and the task of labeling lexical sense using the WordNet hypernym graph was described by Ban et al. (2021) as part of the ConGraCNet application designed for integrating data from different NLP pipelines, lexical dictionaries, and sentiment dictionaries, and has shown perspective results, for example, in the study of linguistic expressions of emotion and the conceptual analysis of cultural framing (Perak 2017, 2019, 2020, Perak and Ban Kirigin 2020, Ban Kirigin et al. 2021).

3 Graph-Based Associative Community Labeling

The hypernym labeling method proposed in this article does not rely on the commonsense knowledge bases, but is based on a dependency-based graph computational linguistic method for identifying lexical sense structure. The underlying method reveals the polysemous nature of lexical concepts through lexical associations. The ConGraCNet method relies on a set of syntactic relations used to construct dependency-based multilayer lexical networks. Each layer is constructed from lexemes collocated in a syntactic dependency that can be harnessed for its semantic potential and function (Ban Kirigin et al. 2015; Perak 2017). The method presented in this paper relies on two specific syntactic patterns: 1) associative syntactic coordination *and/or* and 2) hypernym *is_a* relation. The coordination network layer is constructed from

collocated lexemes in the *and/or* syntactic dependency, which typically associates two ontologically related entities, attributes and/or processes. The highest-ranked co-occurrences of the seed lexeme in the second-order coordinated construction $Lexeme_1$ *and|or* $Lexeme_2$ are used to construct and analyze a graph of syntactically collocated lexemes. They form conceptually associated local clusters or a second-order (friend-of-a-friend) lexical graph with subgraph communities of conceptually associated lexemes representing the conceptual domains related to the source lexeme. These semantically coherent lexical clusters form the basis for dependency-based hypernym labeling.

3.1. Dependency-Based Hypernym Labeling

The conceptual hypernym structure is expressed in language by nominal predicates and copula structures, for example: *car is a vehicle*. This construction allows us to formalize a syntactically based categorial labeling method that uses $Lexeme_x$ *is_a* $Lexeme_y$ patterns to form a lexical graph of semantic hyponym-hypernym relations. The *is_a* relation can be expressed by a copula relation, which is used to link a subject to a nonverbal predicate. The copula is often a verb, but nonverbal (pronominal) copulas are also common in many languages of the world. In this study, we rely on a set of predetermined syntactic patterns that use the *is_a* syntactic-semantic lexical relations defined by the Word Sketch grammars in English and other languages (Thomas 2016).

The construction procedure of the *is_a*-type label network associated to the ConGraCNet lexical cluster consists of the following steps:

- For each lexeme node c_i in the clustered coordination community $\{c_1 \ldots c_x\}$, a *k* number of the best ranked collocations in the $Lexeme_c$ *is_a* $Lexeme_h$ relation is identified;
- A weighted, first-order (source-friend) directed graph of lexical *is_a* collocates is constructed;
- The most prominent nodes in the *is_a* graph are identified using a centrality detection algorithm.

The obtained graph is used to predict the categorically abstract (hypernym) label of the corresponding lexical cluster from the list of most central lexical nodes.

The local nature of the graph computation implies parameterizing the top *k*-ranked hypernym collocates for each lexeme in the seed lexical graph, where the default value is *k*=25. The top *k* collocates are selected from the corpus using a standard collocation measure, e.g., *logDice* or frequency. We use *logDice* as the default corpus collocation measure and *PageRank* as the default graph centrality measure. Other measures can also be used or combined to select the most prominent nodes, e.g., *degree*, *weighted degree*, or *betweenness*. Graph calculations are performed using the Python iGraph library (Csardi et al. 2006). Other centrality algorithms are also developed, aiming at greater ranking granularity and enhanced mapping of the coordination-based source nodes importance on the hypernym target nodes.

For example, the *is_a* graph of the source lexeme *anger-n*, constructed using the large morpho-syntactically tagged English corpora enTenTen13 (19 giga-words), shows several associated lexical communities, listed in Table 1.

Table 1. Best ranked *is_a* hypernym labels and WordNet hypernym labels for associated communities of a source lexeme *anger-n* in enTenTen13.

	Associated Community	Labels ranked by PageRank	Labels ranked by weighted degree	WordNet Hypernym Labels
1	hatred-n, rage-n, bitterness-n, hate-n, jealousy-n, envy-n, bigotry-n, intolerance-n, prejudice-n, violence-n, revenge-n, fury-n	bigotry-n, disease-n, emotion-n, poison-n, sin-n	violence-n, jealousy, revenge-n, intolerance-n, hate-n	violence.n.03, intolerance.n.02, resentment.n.01
2	anger-n, frustration-n, sadness-n, grief-n, disappointment-n, depression-n, stress-n, anxiety-n, despair-n, sorrow-n	emotion-n, feeling-n, reaction-n response-n	anger-n, grief-n, depression-n, anxiety-n, sadness-n	sadness.n.01, emotion.n.01, disappointment.n.01
3	resentment-n, guilt-n, regret-n, hurt-n, blame-n	emotion-n, feeling-n, waste-n, anger-n, hurt-n, resentment-n	guilt-n, regret-n, resentment-n, hurt-n, emotion-n	pain.n.02, resentment.n.01, hostility.n.03
4	fear-n, insecurity-n, shame-n, greed-n	feeling-n, emotion-n, motivator-n, insecurity-n	fear-n, greed-n, shame-n, insecurity-n, emotion-n	emotion.n.01, anxiety.n.02, insecurity.n.01

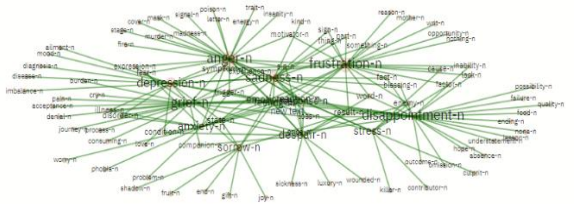

Figure 1. Lexical subgraph labeling: is_a hypernym graph for associated community (anger-n, frustration-n, sadness-n, grief-n, disappointment-n, depression-n, stress-n, anxiety-n, despair-n, sorrow-n) of a source lexeme *anger-n* in enTenTen13.

For textual input, we used the pipeline that collects syntactic dependency data from the morpho-syntactically tagged corpora Sketch Engine API (Kilgarriff et al., 2014; Sketch Engine). The Sketch Engine API was used to extract a summary of various syntactic dependencies co-occurrence data for each lemma. The coordination based second-degree graph is constructed using the default settings: enTenTen13 corpus, *n*=15 best ranked coordination collocations in the first- and second-degree networks, and with pruning removing nodes with *degree* less than 2. Clustering is performed using the *Leiden* community algorithm (Traag et al. 2018) with *mvp* partition type clustering. The lexeme *anger* is clustered in the coherent community: *anger-n, frustration-n, sadness-n, grief-n, disappointment-n, depression-n, stress-n, anxiety-n, despair-n, sorrow-n.* The subsequent projection of the *is_a* relation is shown in Figure 1.

The prediction of hypernym labels for the related senses of the source lexeme *anger-n*, obtained by *PageRank* and *weighted degree* centrality measures, are listed in Table 1. The corpus-based *is_a* label predictions are categorized into columns ranked by PageRank and weighted degree measures. The idea behind the *is_a* network is to identify hypernym candidates of associative lexemes as the central nodes of the network. Both measures provide representations that can be termed as hypernym summarizations. However, it seems that the PageRank ranking targets the more central and thus categorically abstract concepts, while the Weighted degree measure provides the most influential nodes in the network.

3.2 Comparison of the Dependency-Based Hypernym and WordNet Hypernym Labeling

The WordNet hypernym labeling method is based on the construction of a hypernym graph from the lexical subgraph constituents using WordNet synsets and hypernym relation (*lexeme*)-[*has_synset*]->(*synset*)-[*has_hypernym*]-> (*hypernym_synset*), as described in (Ban et al. 2021).

For the comparison of *is_a* hypernym labels and WordNet hypernym labels, see Table 1, which shows the prediction of hypernym labels for the related senses of the source lexeme *anger-n* for both methods of label propagation based on the same lexemes of the same corpus.

One of the main advantages of the WordNet hypernym graph algorithm is the symbolic categorical assignment of lexical nodes to a class within a structured taxonomy. This allows semantic enrichment of the associated lexical communities obtained by the unsupervised bottom-up graph classification method, and results in a set of synsets with well-defined and curated top-down knowledge relations.

The hypernym graph abstracts the categories of lexical communities using WordNet dictionary knowledge relative to the data provided by a large web corpus, such as enTenTen. This results in a comparable corpus-based representation of lexical usage given the same set of graph parameters.

For example, Table 2. shows the labeling of communities based on the English Timestamped newsfeed 2014-2019 corpus. We can see that both corpora yield a set of sense clusters that abstract *anger* in a comparatively similar sense of distinct strong feeling *emotion.n.01*, associated in particular with *violence, insecurity, intolerance, resentment, sadness.*

Table 2. Best ranked WordNet hypernym labels for associated communities of a source lexeme *anger-n* in English Timestamped newsfeed 2014-2019 corpus

	Associated Community	WordNet Hypernym Labels
1	disappointment-n, disgust-n, shock-n, surprise-n, dismay-n, outrage-n, disbelief-n, horror-n	disgust.n.01, surprise.n.02, fear.n.01
2	anger-n, frustration-n, fear-n, confusion-n, anxiety-n, panic-n, uncertainty-n, doubt-n	emotion.n.01, cognitive_state.n.01, anxiety.n.01, uncertainty.n.01
3	hatred-n, hate-n, bigotry-n, intolerance-n, contempt-n, prejudice-n, racism-n	intolerance.n.02, bias.n.01, emotion.n.01
4	resentment-n, bitterness-n, regret-n, jealousy-n, envy-n	resentment.n.01 envy.n.01 hostility.n.03, bitterness.n.02
5	sadness-n, grief-n, rage-n, despair-n, sorrow-n	sadness.n.01, feeling.n.01, sorrow.n.01, sadness.n.02

Table 3. Best ranked WordNet hypernym labels for associated communities of a source lexeme *ljutnja-n* in Croatian hrWac 2013 corpus

	Associated Community	WordNet Hypernym Labels
1	ljutnja-n, frustracija-n, nezadovoljstvo-n, nervoza-n, stres-n, strah-n, napetost-n, nemir-n	fear.n.01 nervousness.n.03 condition.n.01
2	bijes-n, srdžba-n, ljubomora-n, mržnja-n, zavist-n, osveta-n, nesigurnost-n	anger.n.01 envy.n.01 Retaliation.n.01 hate.n.01
3	gorčina-n, tuga-n, očaj-n, cinizam-n, jad-n, jal-n, bol-n	feeling.n.01 sadness.n.01 sorrow.n.01 unhappiness.n.02
4	razočaranje-n, ogorčenje-n, gnjev-n, neuspjeh-n, nevjerica-n, revolt-n, gnušanje-n	anger.n.01 nonaccomplishment.n.01 frustration.n.03
	ogorčenost-n, povrijeđenost-n, nemoć-n, osvetoljubivost-n, zamjeranje-n, zgražanje-n, zabrinutost-n	powerlessness.n.01 quality.n.01 weakness.n.03

Another advantage of using WordNet is the ability to find corresponding hypernym structures in many languages via the Open Multilingual WordNet library (Bond and Foster 2013). The results of the comparative translation equivalent of the concept *anger* in Croatian, concept *ljutnja*, calculated on the basis of the hrWac corpus (*hrWac*), are presented in Table 3.

The cross-linguistic comparison yields a commensurable and yet culturally specific insight into the associative conceptual matrix of the lexeme *ljutnja*, which indicates a somewhat different picture than its English translation equivalent. The lexeme *ljutnja* in hrWac also displays the aggressive features of anger, such as retaliation, but is more abstractly associated with states and emotions experienced when one is not well or does not achieve a desired goal, as well as with the quality of having no strength or power.

In this way, corpus-based graph structures highlight usage-based and cultural differences in the semantic processing of the same lexical concept. These features provide a transparent and consistent approach to intra- and cross-cultural analysis of associative semantic lexical potential for a given source word.

As a drawback of the method, it should be noted that the lexical sparseness of the WordNet hypernym relations hinders the full scope of the mapping.

Nevertheless, the structure of the coordination layer subgraphs can be compensated to some extent by the association of more frequent noun lexemes, which provide a more conventional abstract categorical label for an associated sense of a source lexeme.

4 Discussion and Future Work

The dependency-based hypernym labeling procedure can be used to construct a taxonomic structure of lexical semantics. We have shown that the result depends on the centrality measures used for ranking, with *PageRank* possibly prevailing over *weighted degree* with respect to hypernym structures. This measure-dependent feature needs to be validated in our future work. The informational advantage of corpus-based hypernym labeling is the representation of corpus-specific abstract hypernym structures for a set of associated lexemes. However, this can also be a disadvantage for smaller corpora with a sparse set of syntactic patterns forming an *is_a* representation.

Moreover, the labels do not always match exactly the ontological sense of the hypernym concepts, and some of them express the conventional metaphorical patterns of the conversation. For example, one of the candidates for the hypernym of community 1 (*hatred-n*, *rage-n*, *bitterness-n*, *hate-n*, *jealousy-n*, *envy-n*, *bigotry-n*..) in Table 1 is *poison-n.* To say that hate is a poison would obviously be a metaphorical way of saying that *hate* poisons or clouds judgement, where the underlying conceptual metaphorical extension can be formalized as HATE is not POISON BUT maps the pejorative features of POISON onto HATE (Brdar et al 2019). This procedure, coupled with an additional investigation of ontological similarity, can be used to extract the metaphorical collocations with the *is_a* constructions. This apparent anomaly indicates one of the future directions of the work.

Moreover, we plan to implement the syntactic *is_a* relation for other languages in the future, in particular we are developing a Universal Dependency based tagging construction adapted for the Croatian grammatical expression of the *is_a* relation. In particular, we will extract the *is_a* dependency from the well-known hrWac web corpus, but also from the hr-engRi corpus (Bogunović et al. 2021), using a construction [$\text{NOUN}_{\text{head}}$ - nsubject - $\text{NOUN}_{\text{dependency}}$]. The corpus consists of texts collected from the most popular news portals in Croatia in the period from 2014 to 2018: Direktno, Dnevno, Net Hr, Hrt, Index_Hr, Jutarnji, Novilist, Rtl, SlobodnaDalmacija, Večernji, Tportal, Dnevnik. Linguistic processing of the corpus was performed using the CLASSLA package (https://pypi.org/project/classla/) at the levels of tokenization, sentence splitting, morphosyntactic tagging, lemmatization, dependency parsing and named entity recognition.

In terms of extending the knowledge-base approach to labeling, although the WordNet labelling results

show perspective results, in our future work we plan to integrate other knowledge databases with similar semantic relations, such as Conceptnet and Wikipedia, and compare the results with corpus-based *word is_a category* and *category is_a word* syntax dependency.

Finally, in light of the new state of the art methodologies, we plan to introduce the described dependency layers within a Graph-to-graph Transformer (Mohammadshahi and Henderson 2019), which has shown possibilities for integration of dependency layers into sequence-based language modeling, as well as experiment with building graph neural network language models (Wu et al. 2020).

5 Conclusion

The article describes the procedure of integrating and processing the labeling of lexical clusters of a source lexeme using the *is_a* syntactic dependency layer of a morpho-syntactically tagged corpus, implemented in the ConGraCNet application. The idea is that given corpus evidence for the sense potential of a source word, we can assign the label that can be used to decide which sense is referred to in each context.

In accordance with the processing procedure, after constructing the coordination graph layer for a source word and its lexical clusters, the labeling graph is constructed for each cluster, which provides the conceptual abstraction of the sense clusters associated with the source lexeme. The candidate labels, extracted from the corpus through the is_a collocation relation and identified using centrality measures, abstract the central theme of a particular cluster and provide a means of differentiating abstract meaning in a conceptually rich, ontologically transparent, and computationally efficient manner.

Acknowledgments

This work has been supported in part by the Croatian Science Foundation under the project UIP-05-2017-9219 and the University of Rijeka under the project UNIRI-human-18-243.

References

Alagić, D. and Šnajder, J. (2016) Cro36WSD: A lexical sample for Croatian word sense disambiguation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* pp. 1689-1694.

Ban Kirigin, T., Meštrović, A., & Martinčić-Ipšić, S. (2015) Towards a formal model of language networks. In *International Conference on Information and Software Technologies.* Springer, Cham, pp. 469-479.

Ban Kirigin, T., Bujačić Babić, S., & Perak, B. (2021) Lexical Sense Labeling and Sentiment Potential Analysis Using Corpus-Based Dependency Graph. *Mathematics*, 9(12), 1449.

Barba, E., Procopio, L., Campolungo, N., Pasini, T. and Navigli, R. (2020). MuLaN: Multilingual Label propagation for Word Sense Disambiguation. In *IJCAI* (pp. 3837-3844).

Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010) Strudel: A corpus-based semantic model based on properties and types. *Cognitive science*, 34(2):222–254.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003) Latent Dirichlet allocation. *Journal of machine Learning research*, 3:993–1022

Boella, G. and Di Caro, L. (2013) Extracting definitions and hypernym relations relying on syntactic dependencies and support vector machines. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*, 2013, Sofia, Bulgaria, Volume 2, pp 532–537.

Bogunović, I., Kučić, M., Ljubešić, N., and Erjavec, T. (2021) Corpus of Croatian news portals ENGRI (2014-2018), *Slovenian language resource repository* CLARIN.SI, http://hdl.handle.net/11356/1416.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016) Enriching word vectors with subword information. *arXiv preprint* arXiv:1607.04606.

Bond, F., & Foster, R. (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)* (pp. 1352-1362).

Brdar, M., Brdar-Szabó, R., & Perak, B. (2019) Metaphor repositories and cross-linguistic comparison: Ontological eggs and chickens. *Metaphor and Metonymy in the Digital Age: Theory and methods for building repositories of figurative language*, 225-252.

ConGraCNet application. https://github.com/bperak/ConGraCNet.

Csardi, G., & Nepusz, T. (2006) The igraph software package for complex network research. *InterJournal, complex systems*, *1695*(5), 1-9.

Di Fabio, A., Conia, S., and Navigli, R. (2019) VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing, pp 627–637

Dorow, B., & Widdows, D. (2003) Discovering corpus-specific word senses. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Dumais, S. T. (2004) Latent semantic analysis. *Annual review of information science and technology*, 38(1):188– 230.

EmoCNet project. Emocnet.uniri.hr.

Enderton, H. B. (1977) *Elements of set theory*. Academic press.

EnTenTen. https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen13_tt2_1.

Gardenfors, P. (2004) *Conceptual spaces: The geometry of thought*. MIT press.

Hanks, P. (2004) Corpus pattern analysis. In *Euralex* Proceedings, volume 1, pages 87–98.

Hanks, P. (2013) *Lexical analysis: Norms and exploitations*. Mit Press.

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

Hovy, E. H., Navigli, R., and Ponzetto, S. P. (2013) Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

hrWac22. https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fhrwac22_ws

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012) Improving word representations via global context and multiple word prototypes. In Proc. of *ACL*, pp 873–882.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2015) SensEmbed: learning sense embeddings for word and relational similarity. In Proceedings of *ACL*, pp 95– 105.

Kavumba, P., Heinzerling, B., Brassard, A. and Inui, K. (2021) Learning to Learn to be Right for the Right Reasons. *arXiv preprint* arXiv:2104.11514.

Kingsbury, P. and Palmer, M. (2002) From treebank to propbank. In LREC, pages 1989–1993. Citeseer

Leone, V., Siragusa, G., Di Caro, L., & Navigli, R. (2020) Building Semantic Grams of Human Knowledge. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 2991-3000).

Lin, D. (1998) Automatic retrieval and clustering of similar words. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2* (pp. 768-774).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., P. Rychlỳ, & Suchomel, V. (2014) The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013) Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp 3111– 3119.

Miller, G. A. (1995) WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Mohammadshahi, A. and Henderson, J., 2019. Graph-to-graph transformer for transition-based dependency parsing. *arXiv preprint* arXiv:1911.03561.

Moschovakis, Y. (2006) Notes on set theory. Springer, New York

Navigli, R. and Ponzetto, S. P. (2010) BabelNet: Building a very large multilingual semantic network. In Proc. of *Association for Computational Linguistics*, pp 216–225.

Navigli, R., & Velardi, P. (2005) Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, *27*(7), 1075-1086.

Navigli, R. and Velardi, P. (2010) Learning word-class lattices for definition and hypernym extraction. In Jan Hajic, et al., editors, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL 2010), Uppsala, Sweden, pap 1318–1327.

Pantel, P., & Lin, D. (2002) Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 613-619).

Pennington, J., Socher, R., and Manning, C. D. (2014) Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Perak, B. (2017) Conceptualisation of the Emotion Terms: Structuring, Categorisation, Metonymic and Metaphoric Processes within Multi-layered Graph Representation of the Syntactic and Semantic Analysis of Corpus Data. In Cognitive Modelling in Language and Discourse across Cultures; Cambridge Scholars Publishing, pp. 299–319

Perak, B. (2019) An ontological and constructional approach to the discourse analysis of commemorative speeches in Croatia. In *Framing the Nation and Collective Identities Political*

Rituals and Cultural Memory of the Twentieth-Century Traumas in Croatia; Pavlaković, V. and Pauković, D., (Eds).; Memory Studies: Global Constellations, Routledge, pp. 63–100.

Perak, B. (2020) Emocije u korpusima: Konstrukcijska gramatika i graf metode analize izražavanja emotivnih kategorija. In: *Zagrebačka slavistička škola-48. hrvatski seminar za strane slaviste*, 2020, pp. 100–120.

Perak, B., Ban Kirigin, T. (2020) Corpus-Based Syntactic-Semantic Graph Analysis: Semantic Domains of the Concept Feeling. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46, 493–532.

Perak, B.; Damčević, K.; Milošević, J. (2018) O sranju i drugim neprimjerenim stvarima: Kognitivno-lingvistička analiza psovki u hrvatskome. In: *Jezik i njegovi učinci*, pp. 245–270.

Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A.H. and Riedel, S. (2020) How Context Affects Language Models' Factual Predictions. *arXiv preprint* arXiv:2005.04611.

Prince, E. F. (1978) On the function of existential presupposition in discourse. In *Chicago Linguistic Society*,*miko* Vol. 14, pp. 362–376.

Roberts, A., Raffel, C. and Shazeer, N. (2020). How Much Knowledge Can You Pack Into the Parameters of a Language Model?. *arXiv preprint* arXiv:2002.08910.

Scarlini, B., Pasini, T., and Navigli, R. (2020) SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. In Proceedings of the *34th Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence*

Schuler, K. K. (2005) Verbnet: A broad-coverage, comprehensive verb lexicon.

Schütze, H. (1998) Automatic word sense discrimination. *Computational linguistics*, *24*(1), 97-123.

Schütze, H., & Pedersen, J. O. (1995) Information retrieval based on word senses. In Proceedings of SDAIR'95, pages 161–175. Citeseer

Sketch Engine. https://www.sketchengine.eu/.

Speer, R. and Havasi, C. (2012) Representing general relational knowledge in ConceptNet 5. In *LREC*, pages 3679–3686.

Speer, R., Chin, J., and Havasi, C. (2016) Conceptnet 5.5: An open multilingual graph of general knowledge. *arXiv preprint arXiv*:1612.03975

Thomas, J. (2016) *Discovering English with Sketch Engine Workbook*. Lulu. Com.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1-12.

Velardi, P., Faralli, S., & Navigli, R. (2013) Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, *39*(3), 665-707.

Weir, N., Poliak, A., and Van Durme, B. (2020) Probing neural language models for human tacit assumptions. *arXiv preprint* arXiv:2004.04877.

Widdows, D., & Dorow, B. (2002) A graph model for unsupervised lexical acquisition. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

WordNet synsets. http://globalwordnet.org/resources/wordnets-in-the-world/

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C. and Philip, S.Y. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1), pp.4-24.