

# Lexical Sense Labeling and Sentiment Potential Analysis using Corpus-Based Dependency Graph

---

**Ban Kirigin, Tajana; Bujačić Babić, Sanda; Perak, Benedikt**

Source / Izvornik: **Mathematics, 2021, 9**

**Journal article, Published version**

**Rad u časopisu, Objavljena verzija rada (izdavačev PDF)**

<https://doi.org/10.3390/math9121449>

Permanent link / Trajna poveznica: <https://urn.nsk.hr/urn:nbn:hr:186:136178>

Rights / Prava: [Attribution 4.0 International](#)/[Imenovanje 4.0 međunarodna](#)

Download date / Datum preuzimanja: **2024-07-10**






Repository / Repozitorij:

[Repository of the University of Rijeka, Faculty of Humanities and Social Sciences - FHSSRI Repository](#)



## Article

# Lexical Sense Labeling and Sentiment Potential Analysis Using Corpus-Based Dependency Graph

Tajana Ban Kirigin <sup>1,\*</sup> , Sanda Bujačić Babić <sup>1</sup>  and Benedikt Perak <sup>2</sup> <sup>1</sup> Department of Mathematics, University of Rijeka, R. Matejčić 2, 51000 Rijeka, Croatia; sbujacic@uniri.hr<sup>2</sup> Faculty of Humanities and Social Sciences, University of Rijeka, Sveučilišna Avenija 4, 51000 Rijeka, Croatia; bperak@uniri.hr

\* Correspondence: bank@uniri.hr

**Abstract:** This paper describes a graph method for labeling word senses and identifying lexical sentiment potential by integrating the corpus-based syntactic-semantic dependency graph layer, lexical semantic and sentiment dictionaries. The method, implemented as ConGraCNet application on different languages and corpora, projects a semantic function onto a particular syntactical dependency layer and constructs a seed lexeme graph with collocates of high conceptual similarity. The seed lexeme graph is clustered into subgraphs that reveal the polysemous semantic nature of a lexeme in a corpus. The construction of the WordNet hypernym graph provides a set of synset labels that generalize the senses for each lexical cluster. By integrating sentiment dictionaries, we introduce graph propagation methods for sentiment analysis. Original dictionary sentiment values are integrated into ConGraCNet lexical graph to compute sentiment values of node lexemes and lexical clusters, and identify the sentiment potential of lexemes with respect to a corpus. The method can be used to resolve sparseness of sentiment dictionaries and enrich the sentiment evaluation of lexical structures in sentiment dictionaries by revealing the relative sentiment potential of polysemous lexemes with respect to a specific corpus. The proposed approach has the potential to be used as a complementary method to other NLP resources and tasks, including word disambiguation, domain relatedness, sense structure, metaphoricity, as well as a cross- and intra-cultural discourse variations of prototypical conceptualization patterns and knowledge representations.

**Keywords:** lexical graph analysis; corpus; knowledge representation and reasoning; affective computing; sentiment analysis



**Citation:** Ban Kirigin, T.; Bujačić Babić, S.; Perak, B. Lexical Sense Labeling and Sentiment Potential Analysis Using Corpus-Based Dependency Graph. *Mathematics* **2021**, *9*, 1449. <https://doi.org/10.3390/math9121449>

Academic Editors: Jonatan Lerga, Ljubisa Stankovic, Nicoletta Saulig and Cornel Ioana

Received: 20 April 2021

Accepted: 12 June 2021

Published: 21 June 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The expression of feelings and moods in language is one of the foundations of social communication and interaction of personal and cultural values. Linguistic expressions activate feeling as an emergent cognitive interpretation of the components of the utterance: words and their syntactic organization. According to research in cognitive science [1] and linguistics [2], the process of affective evaluation of a symbolic code is an important component in emergent complex phenomena of creating a sense. Without recognizing, integrating and appraising an affective value in the linguistically articulated conceptual content, be it a rough grained positive vs. negative classification or a nuanced emotional categorization, there is no real comprehension of the text. For humans, the process of experiencing affective quality is evolutionary hardwired, sub-conscious trait that is activated by social interaction. However, humans also have difficulties objectively assessing the affective value of an utterance. On the other hand, for a sequence matching quantitative system, a computer, this is an even more difficult task.

Nonetheless, in recent years there has been a surge of natural language processing (NLP) techniques and resources that address affective and subjective phenomena in text analysis. Although reductive, the resources are becoming more extensive and versatile due to their quantitative nature. At the same time, graph theory, as a branch of discrete

mathematics that studies the properties of graphs, has developed applications in fields as diverse as, for instance, computer science, engineering, physics, sociology, biology, and so on. Graph theory has also had a strong impact on computational linguistics, providing representations of grammar formalism and lexical analysis, among other things. In our linguistic study, graph theory has provided a particularly powerful and useful method for modeling lexical phenomena with which we are concerned with.

This paper demonstrates the application of graph theory to the field of NLP resources and the processing of lexical affective dimensions in sentiment analysis research, and proves that it has the potential to push the quantitative nature of the research further in the qualitative direction.

Sentiment analysis aims to evaluate generalised feeling that people get from cognitive processing of an utterance without focusing on a specific class of emotions. It relies on a simplified system of classifying and/or assigning a normalized range of values for a specific affective dimension. The assignment of a value that can describe the feeling expressed in a language utterance is at the core of the process of sentiment analysis. Sentiment can be evaluated for words, concepts, multi-word phrases, sentences, paragraphs, or entire texts. However, the basic component of sentiment analysis is a word or a lexeme. Lexemes are symbolic representations of conceptual references to a class of things, psychological states and sociocultural constructs, their relations, processes, and characteristics. Some lexemes represent a concept that has a predominantly culturally associated positive feeling, such as: *joy, heart, flower, etc.*, while some lexemes represent concepts that are associated with negative feelings, such as *sorrow, death, war, etc.*

The basis for conducting a sentiment analysis is sentiment dictionaries, which contain information about the sentiment dimension expressed by words, phrases or concepts. These dictionaries are created by annotating the lexical entries based on the psychological evaluation of the words or by extending the already annotated dictionaries using various NLP techniques and resources. The two main shortcomings of the dictionary approach are (1) the lack of lexical entries and (2) problems in assigning sentiment values. The first problem could be easily solved with some labeling campaigns, but the second is a more difficult beast to struggle with, as it is burdened by the inherent subjectivity of evaluating the sentiment of a text, lexical ambiguity, and the domain and culturally specific word sense.

Nevertheless, NLP researchers have proposed general strategies for sentiment dictionary enrichment using a two-step procedure: (1) sentiment seed collection and (2) sentiment value propagation. In the first step, sentiment values of seed lexical structures are manually annotated or ascribed from existing dictionaries. The second step includes the propagation of sentiment values from the seeds to the remaining parts of the foundation graph, whether it is an existing word, phrase or concept graph [3–7].

In this paper, we approach the above problems by using both NLP and graph theory to develop graph methods for distinguishing word senses [8] and lexical sentiment potential enrichment. Our approach is based on the graph propagation algorithm, which uses a corpus-based syntactic dependency layer to compute a target sentiment value. Like other sentiment enrichment systems, we start from a sentiment dictionary and map the existing sentiment values to lexical nodes. Unlike standard bag-of-words abstraction, which lacks the sequential organization of a text, or neural network (skip-gram) sequence statistical analysis that does not distinguish the syntactic dependency categories, we use coordination syntactic dependency relation between lexemes in a corpus to construct a seed lexical graph and identify conceptual clusters.

This corpus-based graph method incorporates standard natural language processing techniques: tokenization, lemmatization, part-of-speech and syntactic dependency tagging, as well as the integration of an array of metadata about the lexical nodes from knowledge databases and sentiment dictionaries. The linguistic structures are transformed into a graph that can be used as a valuable resource for the detection of sense dynamics and the representation of lexical sentiment components.

The main contributions of this paper include:

- Graph method for assigning labels to semantically related lexical sense clusters of a seed lexeme using a directed WordNet-based hypernym graph layer;
- Graph algorithm for extending dictionary coverage by assigning sentiment values to the non-existent lexeme dictionary entries based on a corpus-specific coordination dependency graph layer;
- Graph algorithm for reevaluating sentiment values from a dictionary based on a corpus-specific coordination dependency graph layer;
- Graph metrics and representation of a sentiment distribution of a seed lexeme, called *sentiment potential*, based on a clustering of semantically related lexemes within a corpus-specific coordination dependency graph layer;
- Methodology for dynamic, transparent and corpus-specific sentiment value analysis and model creation.

These contributions introduce solutions to the problems of creating sentiment dictionaries that feature word sense discrimination, do not suffer from sparsity and can represent culturally specific sentiment values.

The paper is organized as follows. In Section 2, we present the method of assigning labels to subgraphs in the lexical dependency graph using the hypernym relation. Section 3, introduces the main contribution of the paper, namely the sentiment potential of lexemes. The presented assignment of sentiment values to lexemes is evaluated in Section 4, while the proposed methodology is discussed in Section 5. We conclude with Section 6, where we also propose avenues for future research.

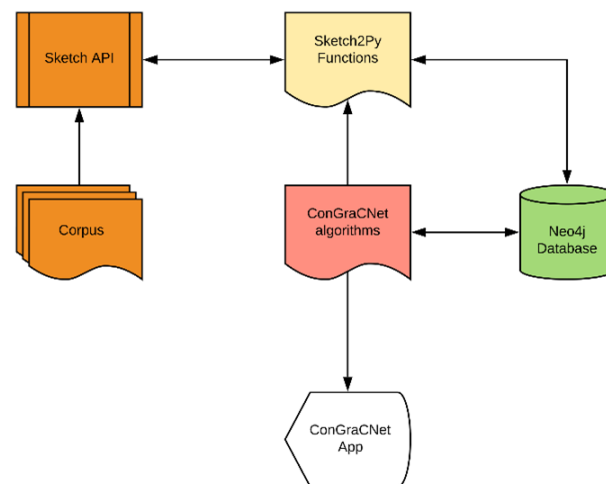
## 2. Labeling Lexical Graphs

As already mentioned, one of the main problems in the field of language processing is related to lexical ambiguity and semantic change. The same word may have multiple senses, or acquire new meanings, sometimes even semantically unrelated. For example, the noun lexeme *bass* may refer to a musical instrument or a species of fish.

This ambiguity problem is obviously related to the assignment of the sentiment value for a specific word, since each sense may have a different sentiment value. For this reason we develop a method for lexical sentiment analysis with distinguishing lexical associations for a seed word. This allows us to assign labels and compute the sentiment value for each associated sense, i.e., to assert the dynamic sentiment potential of a lexeme.

### 2.1. ConGraCNet: Distinguishing Lexical Associations Using Syntactic Dependency Graph

The underlying graph-based computational method for the identifying lexical associations is implemented in the ConGraCNet application [8], developed as part of the EmoCNet project [9], for tagged corpora creation, data retrieval from digital corpora, modeling, storage, algorithmic processing, sentiment analysis, and visualization of semantic-syntactic structures. The ConGraCNet application is designed to integrate data from a number of NLP pipelines, lexical dictionaries and sentiment dictionaries. In this study, we used as textual input the pipeline that collects syntactic dependency data from the morpho-syntactically tagged corpora Sketch Engine API [10,11]. Specifically, we use a large English Web ententen13\_tt2\_1 (ententen13) [12] corpus containing 19 giga words, high frequency of lexical occurrence and various grammatical relation structures. The Sketch Engine API was used to extract a summary of various syntactic dependencies co-occurrence data for each lemma. The pipeline is shown in Figure 1.



**Figure 1.** EmoCnet project ConGraCNet application methodology pipeline.

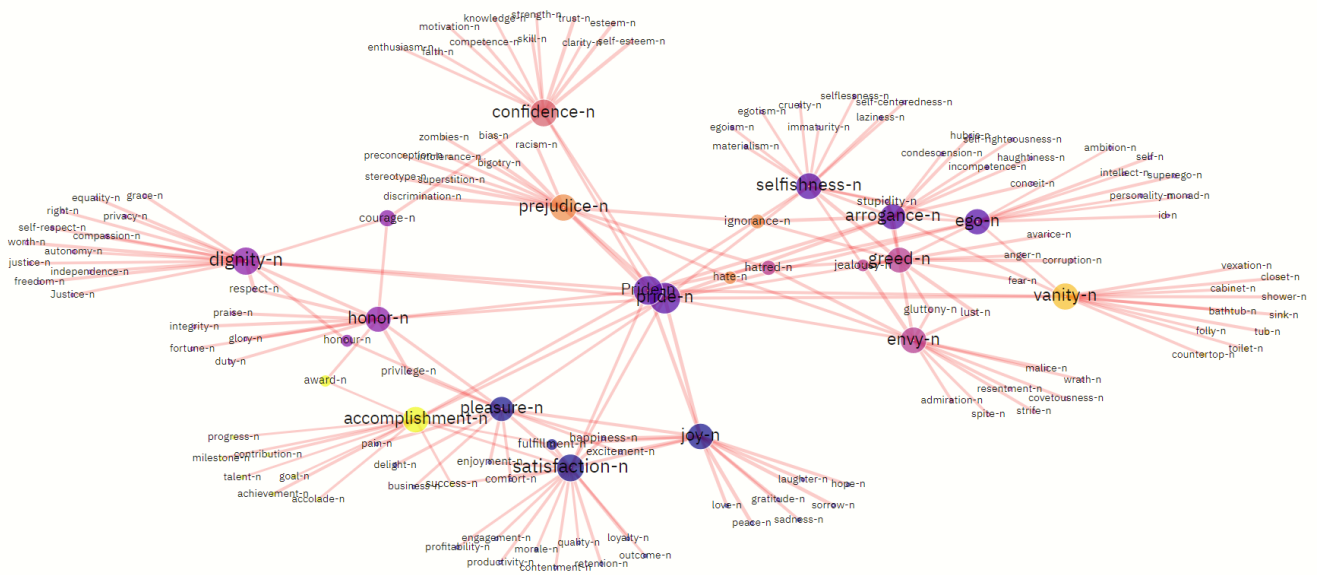
The tagged syntactic dependencies between lexemes in a corpus are used to construct a multilayer network of lexemes. Each layer is constructed from lexemes collocated in a syntactic dependency that can be harnessed for its semantic potential and function [13,14]. For example, *object of* the dependency yields a network of verbs and nouns representing processes that can be performed on entities, while *adjective modifier* represents the attributes of entities, etc. The association network layer constructed from collocated lexemes in the *and/or* syntactic dependency, also called *coordination construction*, typically associates two ontologically related entities, attributes and/or processes with an underlying tendency to be assigned to a category. Therefore, the highest-ranked co-occurrences of the seed lexeme in the coordinated construction [lexeme1 and | or lexeme2] are used to create and analyze a graph of syntactically collocated lexemes that form a kind of conceptually associated class. ConGraCNet coordination-type network construction consists of the following steps:

- (1) Constructing a weighted undirected first-order graph (seed-friend) from the constituent lexemes in the coordination syntactic-semantic construction;
- (2) Constructing a weighted undirected second-order friend-of-a-friend (FoF) graph from the collocated lexemes in the coordination syntactic-semantic construction;
- (3) Identifying prominent nodes in the graph using a centrality detection algorithm;
- (4) Centrality-based pruning of the graph;
- (5) Identifying subgraph communities of collocated lexemes using a community detection algorithm.

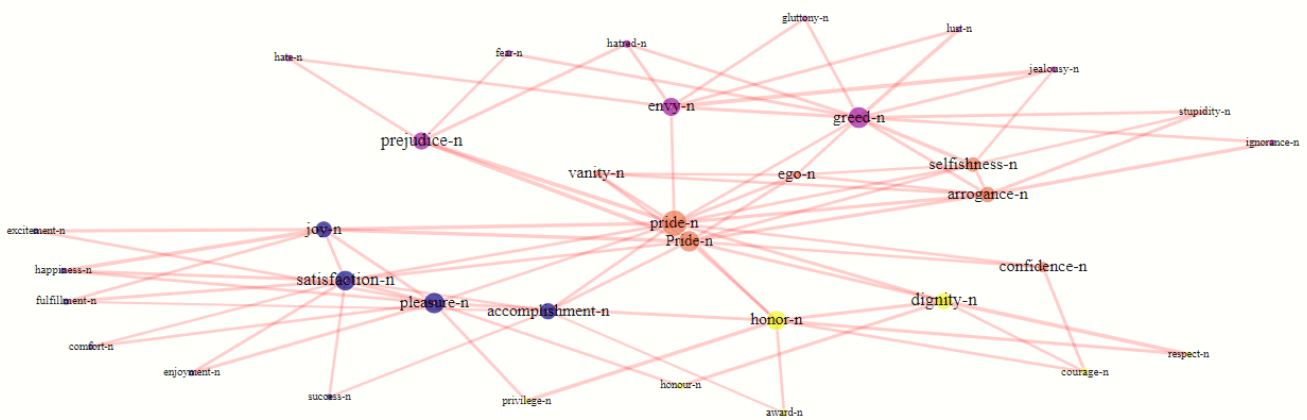
The whole procedure with a number of parameters for control over graph construction, pruning and cluster granularity is implemented in [8].

The second-degree coordination-based (pruned) graph is a representation of the conceptually associated lexemes of a seed lexeme, while the community algorithm clustering represents strongly associated senses. All lexical graphs in the research were constructed using the  $n = 15$  best-ranked coordination collocations in the first and second degree, and clustering was performed using the Leiden algorithm [15] and the mvp partition type.

The clustered second degree coordination dependency lexical graph of the seed lexeme *pride* with 143 nodes and 192 edges and 8 clusters is shown in Figure 2. Its pruned version with 35 nodes and 84 edges in Figure 3 is obtained by removing the nodes with *degree* less than 2. The pruning method allows for a more concise way of representation as well as faster computation, although some marginal information is lost, the most important semantic associations are preserved. The clustered lexical communities are listed in Table 1.



**Figure 2.** Representation of a second order clustered coordination based lexical graph of a seed lexeme *pride-n*: 143 nodes, 192 edges, 9 clusters; parameters: corpus = ententen13,  $n = 15$ , clustering = *leiden*, partition = *mvp*.



**Figure 3.** Representation of a second order clustered coordination based lexical graph of a seed lexeme *pride-n* pruned with  $degree \geq 2$ : 35 nodes, 84 edges, 4 clusters; parameters: corpus = ententen13,  $n = 15$ , clustering = *leiden*, partition = *mvp*.

**Table 1.** Labels of communities in lexical dependency graph of lexeme *pride-n* calculated on ententen13 corpus.

Community	Lexemes	Labels
1	<i>joy-n, accomplishment-n, satisfaction-n, pleasure-n, happiness-n, excitement-n, fulfillment-n, success-n, enjoyment-n, comfort-n</i>	<i>pleasure.n.01, activity.n.01</i>
2	<i>prejudice-n, envy-n, greed-n, hatred-n, hate-n, fear-n, ignorance-n, jealousy-n, lust-n, gluttony-n</i>	<i>mortal_sin.n.01, emotion.n.01</i>
3	<i>pride-n, arrogance-n, ego-n, vanity-n, selfishness-n, confidence-n, Pride-n, stupidity-n</i>	<i>pride.n.01, trait.n.01</i>
4	<i>dignity-n, honor-n, respect-n, honour-n, courage-n, award-n, privilege-n</i>	<i>honor.n.02, award.n.02</i>

The ConGraCNet methodology has shown perspective results, e.g., in the study of linguistic expressions of emotions and the conceptual analysis of cultural framing [16–21].



## 2.2. Labeling the Lexical Associations Using the Wordnet Hypernym Relation

Once the ConGraCNet algorithm yields the graph structure with subgraph communities for a seed lexeme, it is possible to assert labels for each subgraph community. For instance, the lexical community *cat, dog, mouse* could be labeled *domestic animals, mammals*, etc.

Labeling in this sense is a kind of generalization of community features using a more abstract concept. In other words, this labeling task aims to predict the most appropriate hypernym category for a subgraph of the lexical sense community.

The WordNet hypernym labeling method is based on constructing a hypernym graph from the constituents of the lexical subgraph using WordNet synset and hypernym relations in the following steps:

- (1) Each lexical node in a subgraph community is mapped to a lemma with corresponding WordNet synset;
- (2) Each synset is queried for a WordNet hypernym synset, creating:  $(lexeme) \rightarrow [has\_a\_synset] \rightarrow (synset) \rightarrow [has\_a\_hypernym] \rightarrow (hypernym)$  relation;
- (3) The resulting directed weighted  $(lexeme) \rightarrow (hypernym)$  graph is analysed with a centrality measure;
- (4) The most central nodes in the hypernym graph are selected to represent the label of the subgraph community.

The example of the WordNet hypernym labeling procedure for the community subgraphs of the seed word *pride-n* computed on the large web ententent13 corpus is shown in Table 1.

Each directed hypernym relation is uniformly weighted with a weight of 1. The ranking of synset candidates for the most appropriate hypernym label of a coordination-based lexical community is generated using the *degree* and the *PageRank* centrality measures. Intuitively, the *degree* measure yields the most connected nodes in the graph. The nodes with the same *degree* are further ranked by their *PageRank* values, if necessary. The *PageRank* values, on the other hand, indicate the nodes with higher association based on the directed nature of the hypernym relation, thus providing a fine-tuning ranking for nodes with similar *degree* values. The calculations are performed using the Python iGraph library [22]. Other centrality algorithms are also being developed aiming for a greater ranking granularity and enhanced mapping of the coordination-based source nodes importance on the hypernym target nodes.

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

The hypernym graph abstracts the categories of lexical communities with the knowledge from the WordNet dictionary relative to the data provided by a corpus. This yields a comparable corpus-based representation of lexical usage, with the same set of graph parameters. Table 2, for example, shows the labeling of communities based on the europarl7\_en English corpus of Euro Parliamentary discussions (europarl7).

**Table 2.** Labels of semantic communities in lexical dependency graph of lexeme *pride-n* calculated on europarl7.

Community	Lexemes	Labels
1	<i>pride-n, humility-n, arrogance-n, pleasure-n</i>	<i>feeling.n.01, trait.n.01</i>
2	<i>joy-n, emotion-n, satisfaction-n, enthusiasm-n</i>	<i>emotion.n.01, feeling.n.01</i>
3	<i>culture-n, responsibility-n, civilisation-n</i>	<i>society.n.01, culture.n.01</i>
4	<i>dignity-n, honour-n</i>	<i>dignity.n.01, pride.n.01</i>

In these examples, we can see that both corpora provide a set of lexical clusters that abstract *pride* in the sense of an affective and emotional state *feeling.n.01*, in particular as a feeling of self-esteem and personal worth *pride.n.01*, and satisfaction with achievements *pride.n.02*. However, *pride-n* in the web corpus ententen13 is associated with the unforgivable sin that entails a total loss of grace, while the European Parliamentary corpus abstracts community with the antonyms *humility-n*, *arrogance-n* as a distinguishing feature of one's personal nature, and provides more socially oriented labels that highlight an extended social group with a distinct cultural and economic organization at a particular time and place.

Another advantage of using WordNet is the ability to find corresponding hypernym structures in many languages via the Open Multilingual WordNet library [23]. For example, the results of a Croatian lexical concept *ponos* computed on the hrWac corpus [24] are given in Table 3. For example, the same concept *pride* in the Croatian corpus hrwac22 associates the *ponos* with the state of being honored, the state of being free and prosperous, and a feeling of gratitude and appreciation.

**Table 3.** Labels of semantic communities in lexical dependency graph of lexeme *ponos-n* calculated on hrWac22 corpus.

Community	Lexemes	Labels
1	<i>zadovoljstvo-n, radost-n, sreća-n, veselje-n, korist-n, uspjeh-n, mir-n, ispunjenje-n, smijeh-n, ljubav-n, uzbuđenje-n, zdravlje-n</i>	<i>emotional_state.n.01, joy.n.01</i>
2	<i>dika-n, slava-n, dostojanstvo-n, čast-n, Ponos-n, hvala-n, poštenje-n, ugled-n, sloboda-n, poštovanje-n</i>	<i>honor.n.01, freedom.n.01</i>
3	<i>ponos-n, predrasuda-n, prkos-n, tuga-n, ego-n, žalost-n, bijes-n, oholost-n, ljutnja-n</i>	<i>sadness.n.01, unhappiness.n.01</i>
4	<i>zahvalnost-n, priznanje-n</i>	<i>gratitude.n.01, feeling.n.01</i>

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

In the next section we demonstrate the use of sentiment analysis resources to map and assign sentiment value to the ConGraCNet lexical graph structures.

### 3. Assigning Sentiment Value in Lexical Graphs

Sentiment analysis aims to provide a simplified system for classifying of the expressed feeling as well as the assigning a normalized range of values for a specific affective dimension. Sentiment can be evaluated for words, concepts, multi-word phrases, sentences, paragraphs or whole texts. As mentioned earlier, one of the problems with lexical dictionaries is related to their sparseness and rather reductive nature.

The structure of the ConGraCNet lexical graph allows the assignment of sentiment values to a single lexical item or the computation of sentiment values for missing lexemes or whole lexical communities by propagating the values from an existing sentiment dictionary.

#### 3.1. Sentiment Dictionaries

For assigning sentiment values to lexeme nodes and their corresponding subgraph communities we rely on the existing sentiment dictionaries. Sentiment values are mapped to a lexical node using sentiment dictionaries, which contain sentiment data for words and concepts. Different dictionaries provide one or more sentiment dimensions, such as



polarity, positive score, negative score, etc. In this paper, we show the mapping results for three dictionaries: SenticNet, SentiWords and Senti WordNet.

One of the most extensive sentiment lexical dictionaries available is SenticNet6 [6,25]. Its commonsense knowledge base of 200 k lexical concepts is built by integrative top-down and bottom-up learning over an ensemble of symbolic and sub-symbolic AI tools. The upgrading of syntactic and logical features makes this approach very compatible with the ConGraCNet approach. On the other hand, due to the number of semantic dimensions, namely the ‘polarity\_score’, ‘sensitivity’, ‘attitude’, ‘temper’, and ‘introspection’, expressed in numerical values ranging from  $-1$  to  $1$ , we are able to calculate various numeric values representing sentiment of ConGraCNet lexical nodes and community sub-graphs. Throughout this paper, we mainly use the SenticNet 6 sentiment dictionary for our running examples.

SenticNet 6 was integrated into the ConGraCNet app using the Python library SenticNet 6 API [26] to retrieve sentiment dimensions and values. Of particular interest for non-English language corpora is the integration of the multilingual feature of the BabelNetSentic module, which provides sentiment values for 40+ languages [27].

Senti WordNet [28] is another standard sentiment dictionary integrated into the ConGraCNet application. It assigns sentiment values to WordNet synsets instead of words. Each Senti WordNet synset has a positivity and a negativity value normalized between 0 and 1. Since the combined scores never exceed 1, it is possible to calculate the neutral value by subtracting  $1 - (\text{positive} + \text{negative score})$ . This dictionary covers more than 150 k words, with inherited synset ID, part of speech tags and gloss from the WordNet dictionary, making it possible to extend the values to non-English WordNets. This dictionary maps a single lemma to more than one synset, with potentially different sentiment values. For example, the noun *fight* has positive 0 and negative 0 values for the synset glossed as ‘a boxing or wrestling match’, but positive 0 and negative 0.125 values for synset ‘an aggressive willingness to compete’. This structure that usage relies on the word sense disambiguation. This top-down approach to word sense disambiguation and sentiment assignment of the word related senses in the Senti WordNet dictionary complements our bottom-up graph approach to computing the lexical potential of the associated lexical communities.

Another sentiment analysis dictionary, SentiWords [29–31] is a high coverage resource containing about 155 k English words associated with a sentiment score ranging between  $-1$  and  $1$ . The words in this resource are in the form lemma#PoS and are matched against WordNet lists (which contain adjectives, nouns, verbs and adverbs). The scores are learned from Senti WordNet and represent the computation of words’ prior polarities (i.e., the polarity for non-disambiguated words) using Senti WordNet.

### 3.2. Calculating Sentiment Value of a Lexical Graph

This section explains graph procedures that can be used to mitigate the problems associated with sparsity and subjectivity in sentiment analysis. In particular, it demonstrates how to propagate the sentiment value of a lexeme based on a sentiment dictionary and how to evaluate the sentiment value of a lexeme sense subgraph.

In some of the existing sentiment dictionaries, sentiment values are expressed in quantitative dimensions, such as ‘PosScore’ or ‘NegScore’ in the Senti WordNet dictionary or ‘polarity\_value’, ‘attention\_value’, and ‘sensitivity\_value’ in the case of the SenticNet 6 dictionary. Our graph procedures refer to such numerical sentiment categories. For assigning and assessing the sentiment value of words of interest, and their associated lexical communities, we rely on ConGraCNet coordination dependency layer lexical graph with subgraph communities representing associated senses.

In the case of non-existing lexical dictionary entries, we can propagate the word sentiment value by computing the associated lexical graph sentiment value, this enriching the dictionary coverage. Moreover, the existing sentiment values for original dictionary entries can be reassessed using the same graph construction approach and calculating their sentiment value.

Given a lexical graph, we assign the sentiment value to a subgraph according to the dictionary valence values of the node lexemes and the corresponding graph structure. Node importance in the weighted undirected coordination dependency based ConGraCNet graph can be extracted using various node measures, such as *weighted degree*, *betweenness* [32] or *PageRank* [33]. For a chosen graph, a chosen sentiment dictionary and a sentiment category, we define the *Graph Sentiment Value (GSV)* measure.

Given a graph structure  $G$  with the lexemes as nodes, their dependency relations as edges, and a chosen centrality measure of node importance, let  $w(x)$  denote the measure of a node  $x$  in the graph  $G$ . Then, given a sentiment dictionary  $D$  and a sentiment category  $C$ , we extract the numerical sentiment values of the nodes in  $G$ ,  $v(x)$ , that appear in the dictionary  $D$ . Let  $V_G^D$  denote the set of nodes  $x \in G$  for which  $v(x)$  is known.

The graph sentiment value of graph  $G$  is defined for a nonempty set  $V_G^D$  as follows:

$$GSV(G) := \frac{\sum_{x \in V_G^D} v(x) \cdot w(x)}{\sum_{x \in V_G^D} w(x)}. \quad (1)$$

For an example of *GSV* values obtained from SenticNet 6 dictionary ‘polarity\_value’ for a ConGraCNet subgraph see Table 4. The *GSV* values shown were obtained using *weighted degree*, *betweenness* and *PageRank*. The subgraph node lexemes and the corresponding values of the various graph centrality measures are given along with SenticNet 6 ‘polarity\_values’. Note that the lexemes with the highest values of the centrality measure have the greatest impact on the *GSV* computed with respect to that centrality measure. For example, the lexeme *leader* is the only node in the subgraph with a negative ‘polarity\_value’, while the other values from the dictionary are positive or do not appear in the dictionary. Interestingly, many of the listed ‘polarity\_values’ are very high. In terms of graph properties, *betweenness* is the measure that has the highest relative differences between the values of the node measures among those listed. The highest *betweenness* values in the subgraph correspond to the lexeme nodes *Member* and *member*. These values are more than 11 times greater than the nearest *betweenness* value. At the same time, the majority of the nodes have the *betweenness* value of 0. This is not the case for the other measures presented. Consequently, the *GSV* value obtained with *betweenness* is strongly influenced by the negative polarity value of the lexeme *leader*, and only slightly corrected by the positive values of the remaining nodes with nonzero *betweenness* values. Therefore, the *GSV* value obtained with *betweenness* is negative, while the *GSV* values calculated with other centrality measures are positive. The lexeme *leader* does not have much impact on the propagation of *GSV* with respect to the measures *weighted degree* and *PageRank*.

For another example see Table 5. In this subgraph, the most influential node for the *betweenness GSV* value is the lexeme *lecturer* which has a fairly neutral sentiment value. The rest of the available lexeme nodes *ODV* values are all positive. Since the sentiment values of the *weighted degree* and *PageRank* are more evenly distributed, the obtained *weighted degree GSV* and *PageRank GSV* values are significantly more positive due to the influence of the strongly positive lexeme nodes.

In general, the number of node lexemes in a graph that do not have a sentiment value in a dictionary varies. The more values.

As the above examples show, the choice of the centrality measure is of great importance for the computation of the graph *GSV*. Since different measures reflect specific graph properties, the different obtained sentiment assessments of the lexical graph can be used according to a specific lexical analysis.

**Table 4.** GSV of community 2 of lexical dependency graph of lexeme *student-n* calculated with SenticNet 6 polarity\_value on ententen13 corpus.

SenticNet 6	<i>Student-n</i>	Community 2		
Lexeme	Weighted Degree	Betweenness	PageRank	ODV
<i>Member-n</i>	101.01	91.55	0.0653	nan
<i>member-n</i>	99.14	86.71	0.0653	nan
<i>leader-n</i>	22.96	7.81	0.0171	−0.90
<i>friend-n</i>	32.97	0.58	0.0209	0.298
<i>supporter-n</i>	20.84	0.25	0.0167	0.299
<i>volunteer-n</i>	24.55	0	0.0167	0.916
<i>guest-n</i>	15.96	0	0.0126	nan
<i>officer-n</i>	14.94	0	0.0126	0.9
<i>employee-n</i>	14.80	0	0.0126	0.917
<i>partner-n</i>	14.08	0	0.0126	0.45
GSV	0.3530	−0.7846	0.3636	

**Table 5.** GSV of community 3 of lexical dependency graph of lexeme *professor-n* calculated with SenticNet 6 polarity\_value on ententen13 corpus.

SenticNet 6	<i>Professor-n</i>	Community 3		
Lexeme	Weighted Degree	Betweenness	PageRank	ODV
<i>lecturer-n</i>	52.78	69.84	0.0593	0.03
<i>writer-n</i>	36.28	12.96	0.0326	0.883
<i>author-n</i>	30.08	10.66	0.0275	nan
<i>faculty-n</i>	21.80	6.16	0.0210	0.219
<i>clinician-n</i>	13.80	0.54	0.0157	0.823
<i>mentor-n</i>	12.45	0.54	0.0157	0.823
<i>instructor-n</i>	14.11	0	0.0156	nan
<i>speaker-n</i>	15.46	0	0.0159	nan
GSV	0.4376	0.1752	0.4227	

### 3.3. Calculating Sentiment Value of a Lexeme

It is typically the case that not all lexemes that appear in the ConGraCNet graph are labeled with sentiment values in a sentiment dictionary. Certainly, the more lexical nodes are found in the dictionary, the better the evaluation of the assigned sentiment values of the lexical graph. This leads to the idea of extending the coverage of the sentiment dictionary by sentiment value propagation.

In this dictionary enrichment task, the ConGraCNet coordination dependency graph layer identifies candidates without original sentiment values while providing a structure for sentiment value assignment.

Based on the sentiment values of the lexeme entries in the sentiment dictionary, called *Original Dictionary Values (ODV)*, we propagate the sentiment values to lexical nodes without *ODV* thus extending the coverage of the dictionary. We call the resulting sentiment value for lexeme *a* the *Assigned Dictionary Value (ADV)* of lexeme *a*.

Given a sentiment dictionary *D* and its sentiment dimension *C*, the assigned dictionary value *ADV* of the lexeme node *a*, which does not occur in *D*, is computed from the corpus-based coordination dependency lexical *FoF<sub>a</sub>* graph and the available sentiment values for its lexeme nodes. We extract the available valence values in category *C* from *D* for the nodes of *FoF<sub>a</sub>*, denoted by *v(x)*. Let  $V_a^D$  be the set of nodes  $x \in FoF_a$  for which *v(x)* appears in *D*.

The assigned valence value of node  $a$  in category  $C$  of dictionary  $D$ , denoted by  $ADV(a)$ , is defined as follows for a nonempty set  $V_a^D$ :

$$ADV(a) := \frac{\sum_{x \in V_a^D} v(x) \cdot b(x)}{\sum_{x \in V_a^D} b(x)}, \quad (2)$$

where  $b(x)$  is the *betweenness* measure of the node  $x$  in the  $FoF_a$  graph and  $v(x)$  is the valency value of lexeme  $x$  in category  $C$  of  $D$ .

Note that  $ADV(a)$ , as defined in (2), is actually  $GSV(FoF_a)$  with *betweenness* as the centrality measure. The chosen measure of  $ADV$  propagation is *betweenness* because it quantifies the role of the node as a bridge along the shortest path between two other nodes. It provides a way to determine the impact of a node on the connectivity of the graph. Given the  $FoF$  graph structure, the high values of the *betweenness* measure of the lexeme nodes in the ConGraCNet lexical graphs suggest that these node lexemes best represent the primary or dominant lexeme senses.

Some examples of lexemes and their computed  $ADV$  values are shown in Table 6. The absolute differences between the obtained  $ADV$  values and the extracted  $ODV$  values are significant in some cases. For some more examples of obtained  $ADV$  values and the comparison with the  $ODV$  values see Figure 8 in Section 4 where the subjective assessment is presented. For a discussion on the lexical implications of the question of differences between  $ODV$  and  $ADV$  values see Section 5.

**Table 6.** Examples of  $ODV$  and  $ADV$  calculated with SenticNet 6 polarity\_value on ententen13 corpus.

Lexeme	SenticNet 6 $ODV$	SenticNet 6 $ADV$	$ ODV - ADV $
<i>changemaker-n</i>	−0.93	0.128	1.058
<i>researcher-n</i>	−0.83	0.06	0.89
<i>continuance-n</i>	−0.91	−0.214	0.696
<i>hint-n</i>	0.982	0.45	0.532
<i>impatience-n</i>	−0.97	−0.535	0.435
<i>status-n</i>	−0.2	0.23	0.43
<i>desperation-n</i>	−1	−0.647	0.353
<i>hate-n</i>	−0.28	−0.468	0.188
<i>love-n</i>	0.83	0.673	0.157
<i>vacation-n</i>	0.442	0.411	0.031

### 3.4. Sentiment Potential: Graph Propagation Algorithm for Lexeme Sentiment Value

ConGraCNet lexical graph represents an information structure that can be enriched with sentiment values. Orthogonally, this graph can provide a basis for enriching the sentiment representation which addresses the reductive nature of quantifying lexically expressed feeling. We present a graph propagation algorithm that leads to an enriched, more complex sentiment representation for word sentiment, called *Sentiment Potential (SP)*.

For a chosen seed lexeme  $a$ , the  $SP$  propagation algorithm takes the following parameters: (1) a sentiment dictionary  $D$  and its numeric dictionary category  $C$ , (2) ConGraCNet graph parameters including the size of the computed  $FoF$  seed lexeme graph, the clustering algorithm, the centrality algorithm, and the choice of corpus.

For the chosen parameters, the  $SP$  propagation algorithm returns the  $SP(a)$  by performing the following steps:

- (1) Computing the ConGraCNet lexical graph  $G_a$  for the seed lexeme  $a$ ;
- (2) Mapping the  $ODV$  values for  $G_a$  lexeme nodes for which the sentiment values appear in category  $C$  in the sentiment dictionary  $D$ ;

- (3) Computing the *ADV* values for  $G_a$  lexeme nodes for which the sentiment values in category *C* do not appear in *D*;
- (4) Assigning labels for identified lexical subgraphs  $G_a^i$  of  $G_a$ ;
- (5) Computing *GSV* sentiment values for identified lexical subgraphs  $G_a^i$  of  $G_a$  representing the associated lexical senses of the seed lexeme *a*;
- (6) Calculating the average *SP* for the lexeme *a*.
- (7) Visual representation of  $SP(a)$ .

For a chosen lexeme *a* the ConGraCNet method provides the associated lexical graph  $G_a$ . The available sentiment values for the lexical nodes of  $G_a$  can be extracted from the chosen dictionary *D*. For the remaining lexical nodes of the  $G_a$  graph, *ADV* values are assigned, as defined in Section 3.3. These first steps of the *SP* algorithm, (1)–(3), yield the enriched lexical graph of lexeme nodes, each with its associated sentiment value.

Furthermore, in step (1), the subgraph communities of the graph are identified by the ConGraCNet method. Then in step (4), for each subgraph  $G_a^i$  of  $G_a$ , the abstracted category labels are computed using the WordNet hypernym graph, as presented in Section 2.

Next, in step (5), for each subgraph  $G_a^i$  of  $G_a$ , its sentiment value  $GSV(G_a^i)$  is computed using the *weighted degree* centrality measure, as defined in Section 3.2. Table 7 shows an example of a lexical subgraph, along with the *ADV* values assigned to the node lexemes that do not appear in the dictionary. The obtained *GSV* value of the subgraph community of the example lexeme *fight* is positive, since the lexical nodes with the highest rank are very positive, while the negative nodes have lower *weighted degree* values. The remaining subgraph communities of the lexeme *fight* and the *GSV* values for all of the subgraphs are shown in Table 8.

**Table 7.** Propagation of *GSV* of community 1 of lexical dependency graph of lexeme *fight-n* calculated with SenticNet 6 polarity\_value on ententen13 corpus.

SenticNet 6	<i>Fight-n</i>	Community 1	
Lexeme	Weighted Degree	ODV	ADV
<i>fight-n</i>	90.50	0.9	
<i>response-n</i>	32.55	−0.20	
<i>reaction-n</i>	26.42	−0.86	
<i>flight-n</i>	15.29	nan	0.226
<i>action-n</i>	13.69	nan	−0.146
<i>behavior-n</i>	11.98	−0.79	
<i>collision-n</i>	6.53	−0.83	
<i>fighter-n</i>	5.86	0.343	
<i>mode-n</i>	5.52	0.835	
<i>stand-n</i>	5.47	nan	0.093
<i>GSV</i>	0.23		

The obtained *GSV* sentiment values of all subgraphs of the lexical graph  $G_a$  provide a more detailed sentiment score for the lexeme *a*, where each of the  $GSV(G_a^i)$  values corresponds to one of the semantic domains related to the senses of the seed lexeme *a*. This spectre of sentiment values together with the corresponding subgraph communities and their labels, forms the sentiment potential of the seed word.

Finally, in step (6), the average representative sentiment value for the seed lexeme is computed. The *Average Sentiment Potential (ASP)* represents the average sentiment value of the seed lexeme over its semantic communities identified by the ConGraCNet method.

It is computed as the mean of the propagated sentiment values of the identified lexical subgraphs as follows:

$$ASP(a) := \sum_{i=1}^m GSV(G_a^i) \cdot \frac{\sum_{x \in G_a^i} w(x)}{\sum_{x \in G_a} w(x)}, \quad (3)$$

where  $m$  is the number of subgraph communities  $G_a^i$  identified in the seed lexeme graph  $G_a$ ,  $GSV(G_a^i)$  is the  $GSV$  value of the subgraph  $G_a^i$ , and  $w(x)$  is the *weighted degree* of the node  $x$  in  $G_a$ .

For example, the lexeme *fight* has  $ASP -0.0452$ , which is an average sentiment value for this lexeme propagated using ententen13 corpus. Note that the corresponding  $ODV$  value is a strongly positive value of 0.9. The assigned  $ADV$  value is also positive, but significantly lower, due to the influence of the sentiment values of the associated nodes in the corpus-based lexical dependency graph layer.

As shown above, the propagation of the sentiment potential leads to a semantic description of the selected lexeme and an enriched representation of its sentiment value in the corpus-specific data.

As the results for the running example of the lexeme *fight*, presented in Table 8 show, the sentiment value of the word *fight* is positive when associated with collocating lexemes such as *response*, *action*, *reaction*, but negative when associated with collocating lexemes like *misunderstanding*, *quarrel*, *disagreement*. However, in the collocation subgraph containing the lexemes *battle*, *conflict*, *war*, *victory*, it is more neutral due to the mixed sentiment polarity of its community members.

The list of  $GSV$  values obtained represents multifaceted sentiment values of the seed lexeme specifically associated with each associated semantic community. As shown above, a lexical concept can convey very different sentiments in its different senses.

**Table 8.** Sentiment potential (SP): Labels and propagated  $GSV$  of lexical communities for seed lexeme *fight-n* calculated with SenticNet 6 polarity\_value on ententen13 corpus.

SenticNet 6 <i>Fight-n</i> , $ODV$ 0.9, $ADV$ 0.2619			
Community	Lexemes	Labels	$GSV$
1	<i>fight-n</i> , <i>flight-n</i> , <i>response-n</i> , <i>collision-n</i> , <i>fighter-n</i> , <i>mode-n</i> , <i>stand-n</i> , <i>reaction-n</i> , <i>action-n</i> , <i>behavior-n</i>	<i>manner.n.01</i> , <i>support.n.10</i>	0.23
2	<i>argument-n</i> , <i>quarrel-n</i> , <i>disagreement-n</i> , <i>debate-n</i> , <i>dispute-n</i> , <i>dissension-n</i> , <i>misunderstanding-n</i> , <i>contention-n</i>	<i>dispute.n.01</i> , <i>disagreement.n.03</i>	−0.55
3	<i>battle-n</i> , <i>struggle-n</i> , <i>conflict-n</i> , <i>war-n</i> , <i>victory-n</i> , <i>strife-n</i>	<i>conflict.n.01</i> , <i>military_action.n.01</i>	0.02
4	<i>chase-n</i> , <i>brawl-n</i> , <i>feud-n</i> , <i>gunfight-n</i>	<i>fight.n.02</i> , <i>pursuit.n.01</i>	−0.026
$ASP$	−0.0452		

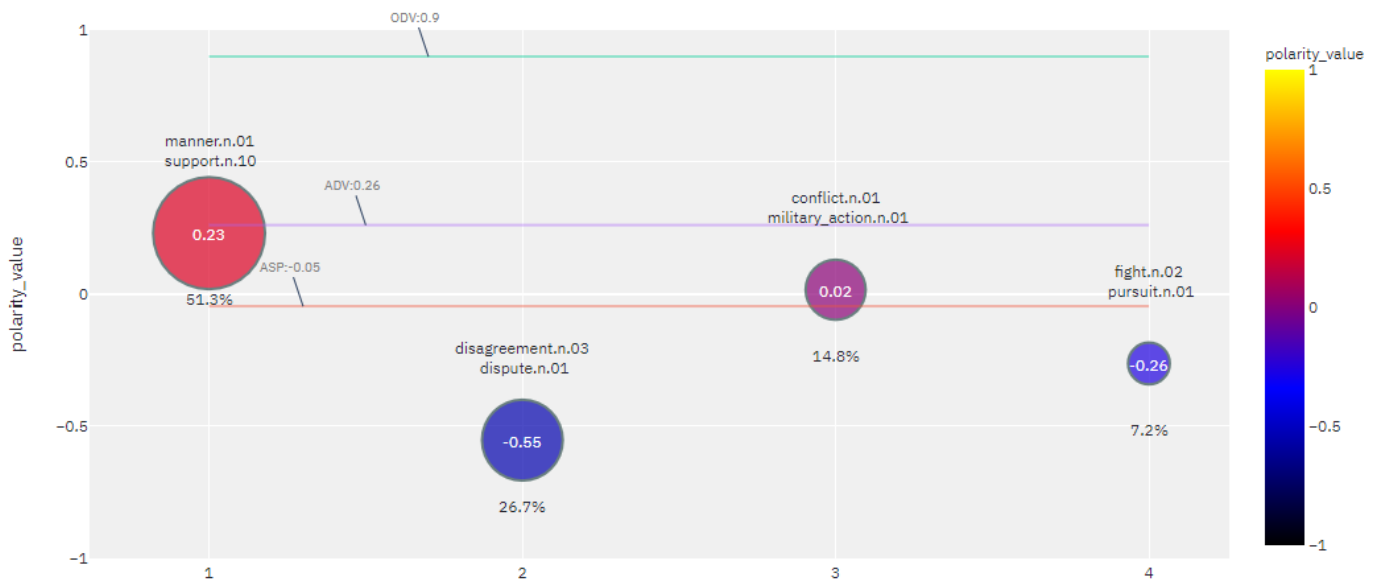
### 3.4.1. Visual Representation of Sentiment Potential

The lexical sentiment distribution contained in the sentiment potential is visualized using a graphical representation shown in Figure 4.

Each lexical community is marked as a circle of the corresponding size, calculated using the relative *weighted degree* proportion in the overall lexical dependency graph of the seed lexeme. The percentage and associated hypernym labels are attached to each community circle. The vertical placement of the community circles and the associated colors reflect the numerical sentiment value of the community, normalized in the range  $[-1, 1]$ . The lexical communities with negative  $GSV$  are positioned lower and darker, with black color representing the  $-1$  value, and vice versa, the positive communities are positioned higher and lighter, with yellow color representing the 1 value. The horizontal lines correspond to the original dictionary value  $ODV$ , the assigned dictionary value



ADV and the assigned sentiment potential *ASP* values obtained by dictionary mapping, computing the sentiment values of the associated lexemes, and the average sentiment value of their lexical clusters, respectively.



**Figure 4.** Sentiment potential (*SP*) of lexeme *fight-n* calculated with SenticNet 6 polarity-value; propagated on pruned graph clusters of 15 best ranked collocates in the first degree and 15 collocates in the second degree within ententen13 corpus.

### 3.5. Analyzing Sentiment Dynamics

In general, the obtained *SP* of a lexeme *a* is associated with some chosen parameters. By changing a parameter, e.g., choosing a different sentiment dictionary or changing a corpus, the comparison of the different *SP* values can give an insight into the dynamic sentiment potential of the chosen lexeme relative to the specific corpora in synchronic and diachronic dimensions.

For example, the *SP* values of the lexeme *fight* obtained for some other SenticNet 6 dimensions and the SentiWords dictionary are shown in Table 9, while the visualization of the *SP* values of the lexeme *fight* obtained using different corpora, English Plenary sessions of the EuroParliament (europarl\_plenary) and hrwac22, are shown in Figures 5 and 6.

**Table 9.** Sentiment potential (*SP*): Propagated SenticNet 6 dimensions: temper, attitude, sensitivity, and SentiWords *GSV* of lexical communities for seed lexeme *fight-n* calculated on ententen13 corpus.

<i>SP</i>	<i>Fight-n</i>	SenticNet 6 'Temper'	SenticNet 6 'Attitude'	SenticNet 6 'Sensitivity'	SentiWords
<i>ODV</i>		-0.0109	0.3891	-0.4184	-0.41836
1	manner.n.01, support.n.10	-0.021	0.067	0.2287	0.0426
2	dispute.n.01, disagreement.n.03	-0.057	-0.103	-0.3591	-0.3102
3	conflict.n.01, military_action.n.01	0	0.047	0.1066	-0.2365
4	fight.n.02, pursuit.n.01	-0.149	-0.343	0.1326	-0.0449
<i>ASP</i>		-0.0367	-0.0109	0.0468	-0.0992

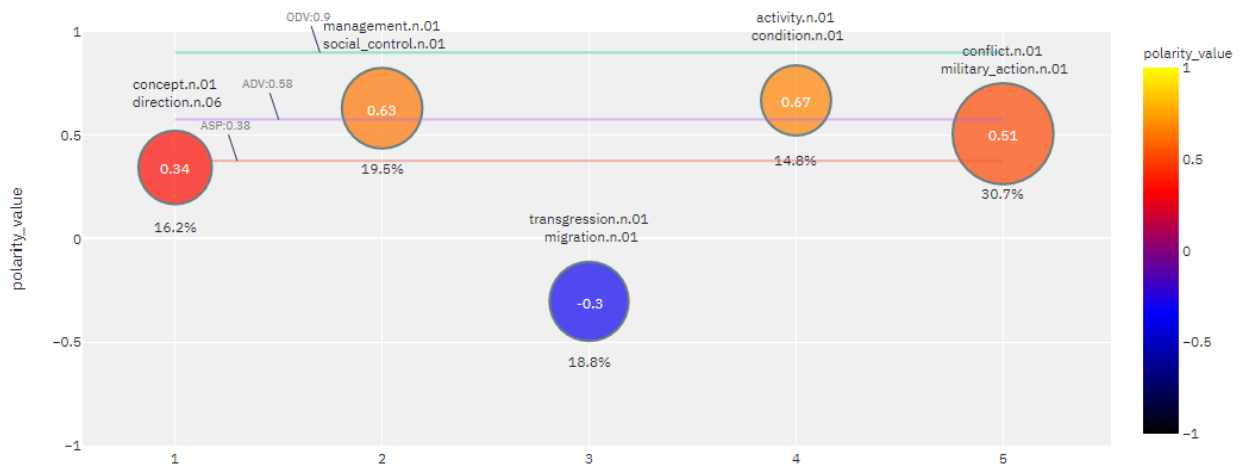


Figure 5. Sentiment potential (SP) of lexeme *fight-n* calculated with SenticNet 6 polarity-value; propagated on pruned graph clusters of 15 best ranked collocates in the first degree and 15 collocates in the second degree within europarl\_plenary

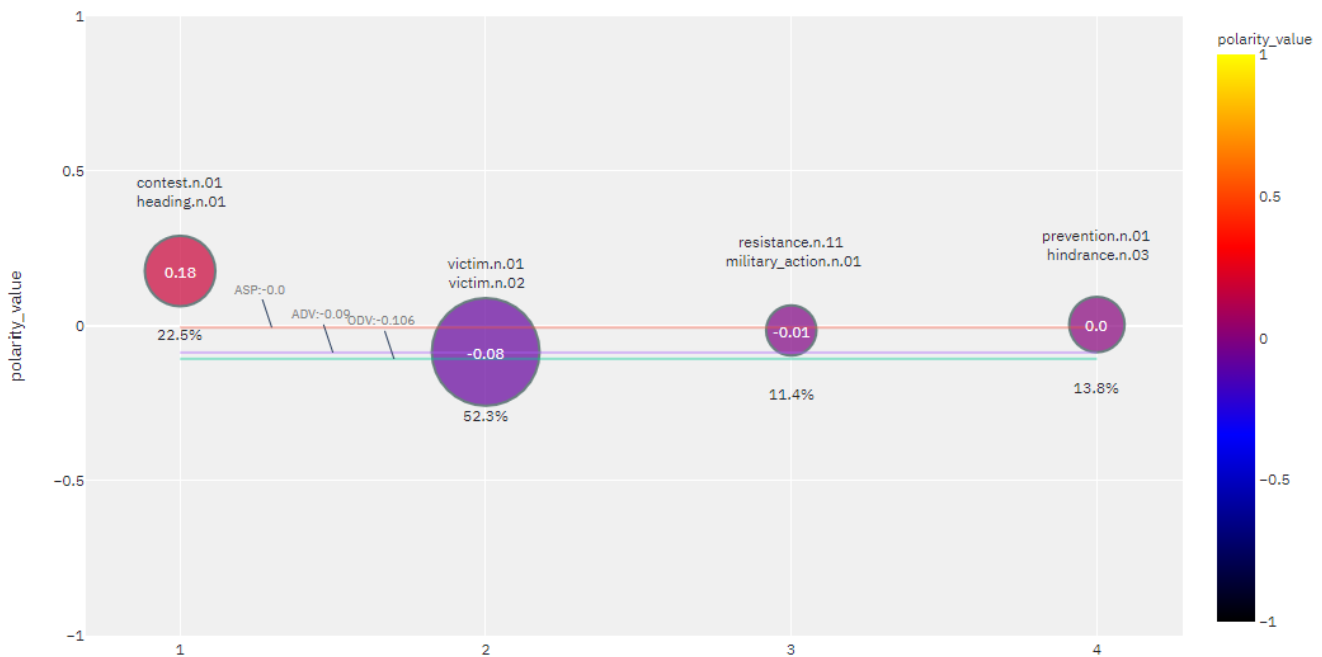


Figure 6. Sentiment potential (SP) of lexeme *borba* (eng. *fight-n*) calculated with SenticNet 6 polarity-value; propagated on pruned graph clusters of 15 best ranked collocates in the first degree and 15 collocates in the second degree within hrwac22 corpus.

The resulting *ASP* value of the lexeme *fight* in the europarl\_plenary corpus is significantly more positive than that obtained in the general ententen13 corpus. The obtained labels for lexical clusters in the europarl\_plenary indicate a domain specific political conceptualization of *fight* as a *concept* with *direction* describing how something should be done or managed.

Moreover, with different language corpora, a cross-language comparison of sentiment representations for the same seed concept can be achieved. In this way, the proposed sentiment potential provides the method for multilingual and cross-cultural linguistic studies.

#### 4. Evaluation

Representing word sentiment with numerical values, or even classifying sentiment as positive, neutral or negative is not a straightforward task. In fact, different sentiment values have been assigned to the same lexeme in different sentiment dictionaries. For example, the lexeme *cream* has the SentiWords polarity score of 0.4479, the SenticNet 6 sentiment polarity value of  $-0.95$ , and a neutral valence in the Senti WordNet sentiment dictionary.

The criteria by which these values were assigned are not always clear and are necessarily subjective. What should be less subjective is the relative difference in sentiment value between words with similar senses and between closely related words in a chosen sentiment dictionary. This is not always the case. For example, the Senti WordNet sentiment dictionary marks the lexeme *researcher* as positive, while *research* is neither positive nor negative.

In other sentiment dictionaries, their sentiment values differ as well. In SentiWords, *researcher* has the positive ‘polarity\_score’ of 0.1756 and the lexeme *research* has a neutral ‘polarity\_score’ 0. In SenticNet 6 the ‘polarity\_value’ of the lexeme *researcher* has a quite negative value of  $-0.83$ , while the lexeme *research* has the opposite ‘polarity\_value’ of 0.883. For comparison, our methodology yields an *ADV* value of 0.0621 for the lexeme *researcher* and *ADV* value of 0.7587 for lexeme *research*.

Although it is difficult to imagine an objective evaluation of sentiment values, we attempted to measure the subjective evaluation of sentiment values from a dictionary in two separate surveys. Survey responses were anonymized. The surveys were conducted anonymously among elementary school and university students, both male and female.

The first survey tested subjective ratings of sentiment polarity scores from the SenticNet 6 dictionary for approximately 30 selected noun lexemes. The selected lexemes had different sentiment values, ranging from strongly negative to neutral to strongly positive. A group of 27 subjects were first presented with word examples and their semantic values to familiarize themselves with the word sentiment. In the survey, the original sentiment dictionary value was presented for each lexeme, e.g., *chocolate*, 0.003. Subjects were asked to indicate their personal perception of the presented sentiment value on a scale of 1 to 5, where 1 marked strong disagreement and 5 strong agreement with the presented value.

The results of the evaluation are shown in Figure 7. In addition, Figure 7 shows examples of lexeme ratings with divergent results. For example, when rating the lexeme *chocolate* the vast majority of subjects had an opinion, either agreeing or disagreeing, whereas for the lexeme *hunger* opinions were not as clear. For the lexeme *happiness* the vast majority of subjects strongly agreed with the polarity value presented.

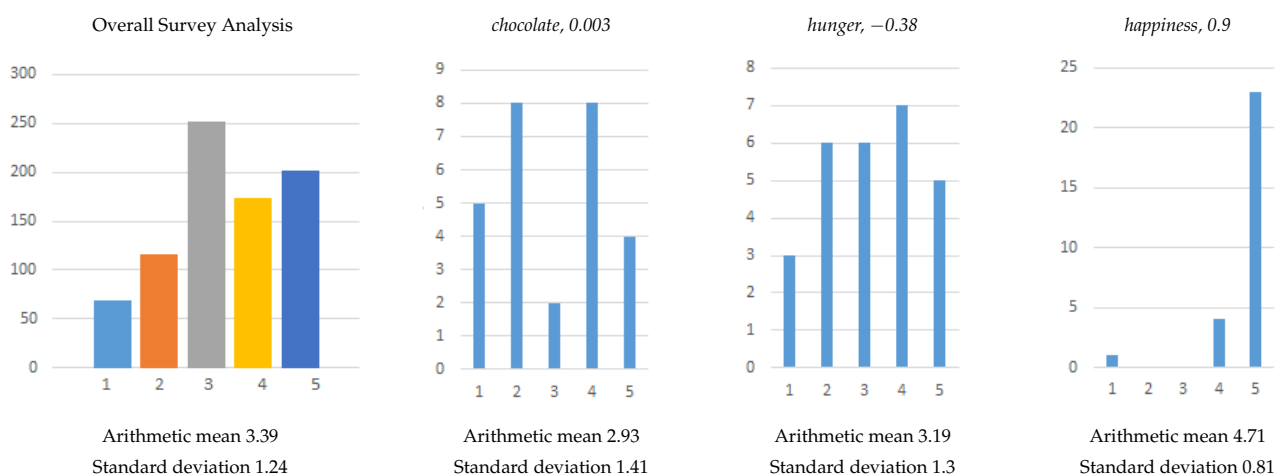


Figure 7. Survey Analysis on SenticNet 6 polarity-value ODV evaluation.

The obtained results clearly show that the evaluation and sentiment assessment is highly subjective. We suspect that the range of results would be similar for another sentiment dictionary. This volatile assessment is difficult to capture accurately due to the polysemous nature of the words.

In addition, we surveyed a group of 66 subjects who evaluated the original SenticNet 6 against the assigned dictionary ‘polarity\_value’ obtained by our methodology on sample of lexemes. The criterion for selecting the survey lexemes was the indication of a noticeable difference in sentiment value between the original (ODV) and the computed numerical value (ADV). The absolute difference between the corresponding ODV and ADV values ranged from 0.21 to 0.92, with an average of 0.65. In some cases, the polarity\_values of the presented lexemes had opposite signs, i.e., some lexemes have a positive valence calculated by our methodology and a negative valence calculated by the SenticNet 6 dictionary, or vice versa. For each of the selected lexemes both ODV and ADV numerical values were presented for assessment. Subjects were asked to select which of the presented values was closest to their own perception. The numerical values were not labeled, and their order was mixed, so subjects were unable to identify the origin of the values.

The overall results on lexeme evaluations are shown in Figure 8. Overall, for a set of 23 lexemes for which both ‘polarity\_values’ were given and anonymized in the sense that the ODV and ADV labels were not visible, 76% of subjects chose the ADV over the ODV value. Figure 8 on the right also shows the results on the ratings of the example lexemes, from a large majority for ADV values, to no clear majority to a majority for ODV values. The cases with the large majority of ADV preference may indicate an error in dictionary input, antonymy, or a lexeme with clear polysemous nature. The results may also be influenced by the lack of POS labeling in the SenticNet 6 sentiment dictionary. For a discussion, see Section 5.

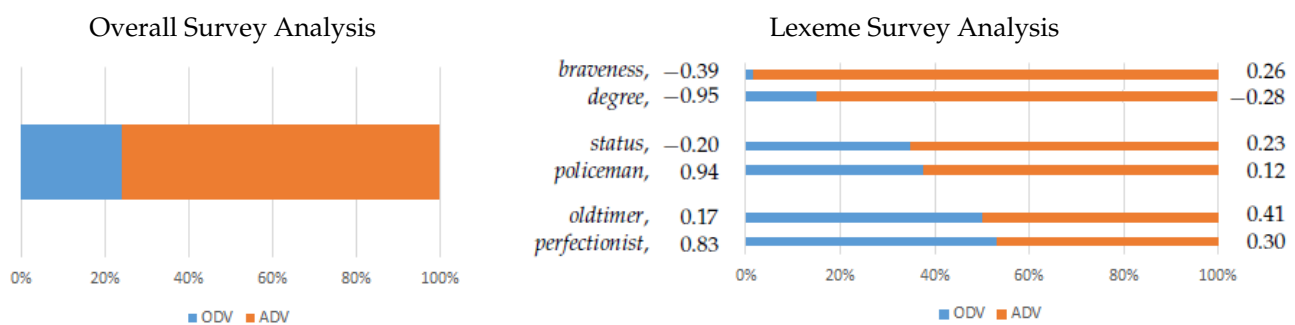


Figure 8. Survey Analysis on SenticNet 6 polarity\_value ODV vs. ADV on example lexemes evaluation.

## 5. Discussion

As shown in the previous sections, sentiment analysis is a complex area of linguistic research. In addition to the psychological subjectivity inherent in emotional phenomena, it deals with the ambiguity and polysemous nature of the symbolic code, and with reductive attempts to assign the normalized token count to a dimension of the emotion invoked by an utterance. However, the cognitive structure of linguistic-syntactic dependencies revealed in a corpora gives us the opportunity to develop a more consistent and transparent domain-dependent enrichment of symbolic tokens with numerical sentiment values [34].

Instead of considering a text as a bundle of words, or only superficially considering the structure of syntactic information, the ConGraCNet graph machine learning methodology uses a corpus-based syntactic dependency graph layer to learn the domain-specific cognitive structures that can be more consistently enriched with sentiment values. The coordination dependency lexical graph layer reveals the commonsense associative structure of a single lexical node, which enables more transparent and human-like assignment of a normalized sentiment value and representation of its polysemous sentiment potential within a domain of texts.

The polysemous nature of the lexeme in a corpus is identified using a clustering algorithm of a seed lexeme. The hypernym graph label assignment described in Section 2 provides labels for each of these subgraph communities. The assigned labels emerge as the most central nodes in the hypernym graph obtained using WordNet hypernym and synset relations. Among the different implemented approaches to labeling data in ConGraCNet, we used a dictionary-based approach for assigning labels to a subgraph community using WordNet synsets hypernym relations. The other approaches could involve other knowledge databases, such as ConceptNet.io [35], and DBpedia, or a combination with corpus-based *is\_a* syntactic dependency methods.

Particularly important for building sentiment analysis resources, this graph method can be used to check the consistency of the initial sentiment dictionary values to identify, reassess and possibly reassign some values with respect to the associated lexical clusters. In our research, we created a blueprint of the procedure on a small sample of SenticNet 6 dictionary values, which includes (1) reassignment of sentiment values for each node, (2) comparison of *ODV* with *ADV*, (3) ranking of nodes with largest difference, (4) assessment of sentiment values by human subjects. In our future work, we will incorporate this iterative process with the goal of generating representative and consistent sentiment dictionaries from a set of sparse sentiment dictionary entries, reassigning values in accordance with human evaluation, or evaluating a new sentiment dimension.

Of course, the structural dependency of the syntactic graph on a specific corpus should be taken into account. The larger the corpus, the more generalized, robust and rich the semantic relations represented in the associative layer will be. In general, the (re)assignment of sentiment dictionary is more representative if it is built on a larger corpus. However, if the sentiment values in a word dictionary are (re)evaluated reasonably consistently, the corpus can be considered as the dependent variable. The ConGraCNet procedure can then be used to reveal the cultural, domain specific, stylistic or idiomatic variations of sentiment potential, as is described in Section 3.

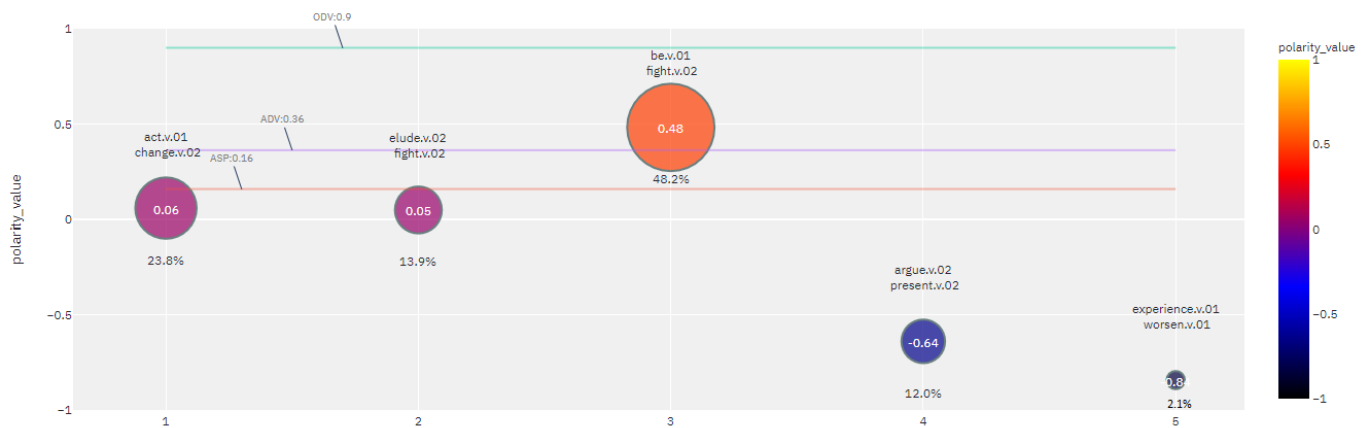
Moreover, the sentiment potential representation *SP* has the ability to enrich lexical semantics [36] by adding a dynamic and polysemous dimensions of sentiment values that, to our knowledge, are not yet part of the existing semantic research. For instance, by using time-stamped data, one can measure not only diachronic semantic change at the level of lexical relations, but also the underlying salience of the sentiment potential that occurs as a result of this change. A somewhat vivid example of this diachronic change is the lexeme *corona*, which has a rather positive sense potential *ASP* 0.246, with *ODV* 0.27 using a SentiNet 6 dictionary and the WordNet hypernym labels related to beer, aureole and landmarks. However, the recent COVID-19 crisis has imparted new layer to the sentiment potential of the lexeme *corona*. Some recent corpora, like Timestamped web corpus of English 2014–2020 [37], identify this different sentiment potential of the lexeme *corona* with *ASP* 0.08. When compared to the previous dominant trends, there appears to be a distinct lexical cluster labeled as *influenza* and *contagious disease* with graph sentiment value *GSV*  $-0.373$  comprising 16 percent of the seed lexeme *corona-n* graph structure of 50 best ranked collocates in the first degree and 5 in the second degree. This type of analysis opens up a whole new dimension in the semasiological change research.

Another interesting lexical semantics feature is the presence of antonymous lexemes and lexical communities within a seed lexeme graph, as illustrated in the seed graph of the lexeme *pride*. *Pride* is associated with pleasure, but also with mortal sin. It arouses positive sentiments labeled as *pleasure*, such as *joy*, *happiness*, but also negative ones, such as *greed*, *envy*, *jealousy*, labeled as *mortal sin*. This phenomenon is part of embodied cognitive structures that have a long history of research [38,39], but sentiment potential and lexical clustering representation can add new data to a somewhat rigid one-dimensional conceptual view of antonymy phenomena. One of our future research tasks will be to further investigate the sentiment-enriched graph representation to find domain specific lexemes and lexical structures with antonymous features.

From the data processing perspective, the method requires that the corpus be tagged with grammatical and syntactic features in order to construct a coordination network. This process requires the extensive use of NLP morpho-syntactic tagging and a database to store and extract the occurrence of lexical nodes and the co-occurrence of the nodes. However, resulting coordination layer can then be used to identify conceptual associations between nouns as entities, adjectives as attributes of entities, verbs as processes and adverbs as attributes of processes. Other parts of speech are necessarily dependent on these grammatical categories and therefore have no inherent prototypical sentiment. For instance, it would be odd to assign a definite sentiment value to a particle *a*, *the*, logical connectors *and*, *or*, or a pronoun *you*, etc., since they acquire their sentiment values in syntactic-semantic constructions.

After constructing a syntactic dependency network embedding, the ConGraCNet approach to lexical sentiment analysis is actually language agnostic. Equipped with a relevant sentiment dictionary with enough entries to boot-start graph propagation procedure, the method can be applied to a lexical coordination graph of any language.

On this topic, it is important to mention the morpho-syntactic format of the sentiment dictionaries. The two out of three sentiment dictionaries used here, SentiWords and SentiWordNet have the Part of Speech (POS) tags. POS grammatical metadata gives a more accurate mapping of the *ODV* to a lexical node. SenticNet, due to the nature of symbolic data extraction, does not provide the POS tags, which can result in a somewhat overlapping mapping of the same sentiment value to an orthographically identical lexeme but grammatically completely different category, as in the case of the lexeme *fight*, which symbolically can be both noun *fight-n* and verb *fight-v*. In addition to the different semantic subgraphs and corresponding labels, the average sentiment potential value of these lexemes is also different, namely,  $ASP(fight-v) = 0.16$  and  $ASP(fight-n) = -0.05$ , see Figures 4 and 9. In this case, the calculation of *ADV* can be used to assess the adequacy and consistency of *ODV*.



**Figure 9.** Sentiment potential (*SP*) of lexeme *fight-v* calculated with SenticNet 6 polarity-value; propagated on pruned graph clusters of 15 best ranked collocates in the first degree and 15 collocates in the second degree within ententent13 corpus.

The computation of the values depends on the choice of the sentiment dictionary, while the assignment of the new values and re-computation of the old ones additionally depends on the choice of the parameters of the graph-based ConGraCNet methodology, such as the size of the computed seed graph, the clustering and centrality algorithms, and the choice of the corpus.

This syntactic dependency lexical approach can be useful for a number of downstream applications dealing with sentiment analysis of larger linguistic structures, such as constructions, phrases, sentences, texts. The *ADV* values assigned to the lexemes and their *GSV* values together with their abstracted labels can be used as components for contextual sentiment text summary or dynamic compositional sentiment analysis [40].



## 6. Conclusions and Future Work

The paper addresses the problems of data integration, processing and enrichment in the context of the graph method used for labeling lexical clusters of a seed lexeme and identifying its sentiment potential using the syntactic dependency layer of a morpho-syntactically tagged corpus, implemented in the ConGraCNet application. According to the processing procedure, after the construction of the coordination graph layer for a seed word and its lexical clusters, we first implement the cluster labeling algorithm. This provides the conceptual abstraction for semantically related lexical communities of a seed lexeme. Labels for each community are propagated by constructing a WordNet hypernym graph from the clustered lexemes in a coordination graph layer and selected based on a set of centrality measures. The candidate labels abstract the central theme of a particular cluster and provide a way to distinguish it from other clusters in an efficient manner.

Although the results of WordNet labeling are promising, in our future work we plan to integrate other knowledge databases with similar semantic relations, such as Conceptnet *is\_a*, and compare the results with corpus-based *word is\_a category* and *category is\_a word* syntactic dependencies.

The main contribution of the paper is related to the description of the graph algorithms that can be used to calculate new lexical sentiment values and extend the range of sentiment dictionaries that currently suffer from sparseness and lack of culture-specific sentiment values; to re-evaluate existing sentiment values from a sentiment dictionary based on a corpus-specific coordination dependency graph layer; to introduce polysemous sentiment graph metrics and distribution of a seed lexeme, called *sentiment potential*, with description of the sentiment potential algorithm, a graph procedure for mapping, identification and reassignment of the sentiment values from a number of sentiment dictionaries.

Starting from SenticNet, SentiWords, SentiWordsNet dictionary based sentiment dimensions and values, associated to lexemes in a coordination graph layer, we have demonstrated the efficiency of the graph algorithm for assigning the sentiment values to the lexical sense communities, to missing lexemes or for reassignment of the sentiment values based on a specific corpus. This methodology for dynamic, transparent and corpus-specific sentiment value analysis and model creation allows us to create sentiment lexical dictionaries for various languages and specific corpora.

We performed an evaluation of the reliability of the sentiment values on a sample of 30 lexemes, as well as the subjective evaluation of the original dictionary values vs. graph assigned values on a set of lexemes with slightly larger difference via questionnaire on 66 subjects. The distribution of *ODV* evaluation have shown relative agreement with the original values. Subjective ratings of the appropriateness of the *ADV* were relatively high with 76% of subjects choosing the *ADV* value over the *ODV* value. These promising results motivate us to further experiment further with the corpus-based graph algorithm to extend the dictionary coverage by assigning a sentiment value to non-existing lexeme dictionary deal with. In our future work we will implement the algorithms to develop dense lexical sentiment dictionaries and sentiment potential models calculated from various large corpora of major languages that could be used in standard procedures for sentiment analysis. The other line of applications will address sentiment dictionary consistency analysis.

Our future work will also extend the syntactically based analysis to a other syntactic dependency layers to create sentiment potential for various multi-word expressions.

In general, the proposed approach has the potential to be used as a complementary method to other NLP contemporary resources for the enrichment of various semantic tasks including word disambiguation, domain relatedness, sense structure, synonymy, antonymy and metaphoricality, as well as establish a cross-and intra-cultural discourse variations of prototypical conceptualization patterns and knowledge representations.

**Author Contributions:** Conceptualization, B.P. and T.B.K.; methodology, T.B.K., B.P., S.B.B.; software, B.P. and S.B.B.; validation, T.B.K., S.B.B. and B.P.; formal analysis, T.B.K., S.B.B. and B.P.; investigation, T.B.K., S.B.B. and B.P.; resources, B.P.; data curation, B.P. and S.B.B.; writing—original draft preparation, T.B.K., S.B.B. and B.P.; writing—review and editing, T.B.K., S.B.B. and B.P.; visual-

ization, B.P., S.B.B. and T.B.K.; supervision, B.P. and T.B.K.; project administration, B.P. and T.B.K.; funding acquisition, B.P. and T.B.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** 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.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. The data analysis and representation can be found here: ConGraCNet webapp <http://emocnet.uniri.hr/congracnet2/> (accessed on 16 June 2021) A list of Sketch Engine corpora and their availability: <https://www.sketchengine.eu/corpora-and-languages/corpus-list/> (accessed on 16 June 2021). SenticNet datasets can be found at: <https://sentic.net/downloads/> (accessed on 16 June 2021); WordNet synsets: <http://globalwordnet.org/resources/wordnets-in-the-world/> (accessed on 16 June 2021); WordNet domain and SentiWords 1: <https://wdomains.fbk.eu/download.html> (accessed on 16 June 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Okon-Singer, H.; Stout, D.; Stockbridge, M.; Gamer, M.; Fox, A.; Shackman, A. The interplay of emotion and cognition. In *The Nature of Emotion: Fundamental Questions*; Fox, A.S., Lapate, R.C., Shackman, A.J., Davidson, R.J., Eds.; Oxford University Press: New York, NY, USA, 2017; pp. 181–185.
- Alba-Juez, L. Emotion and appraisal processes in language. *Constr. Discourse Verbal Interact.* **2018**, *296*, 227.
- Tsai, A.C.R.; Wu, C.E.; Tsai, R.T.H.; Hsu, J.Y.J. Building a concept-level sentiment dictionary based on commonsense knowledge. *IEEE Intell. Syst.* **2013**, *28*, 22–30.
- Cambria, E.; Fu, J.; Bisio, F.; Poria, S. AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–29 January 2015; Volume 29.
- Cambria, E.; Poria, S.; Hazarika, D.; Kwok, K. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
- Cambria, E.; Li, Y.; Xing, F.Z.; Poria, S.; Kwok, K. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Galway, Ireland, 19–23 October 2020; pp. 105–114.
- Ahmed, M.; Chen, Q.; Li, Z. Constructing domain-dependent sentiment dictionary for sentiment analysis. *Neural Comput. Appl.* **2020**, *32*, 14719–14732.
- ConGraCNet Application. Available online: <https://github.com/bperak/ConGraCNet> (accessed on 6 June 2021).
- EmoCNet Project. Available online: [emocnet.uniri.hr](http://emocnet.uniri.hr) (accessed on 6 June 2021).
- Kilgarrieff, A.; Baisa, V.; Bušta, J.; Jakubiček, M.; Kovář, V.; Michelfeit, J.; Rychlý, P.; Suchomel, V. The Sketch Engine: Ten years on. *Lexicography* **2014**, *1*, 7–36.
- Sketch Engine. Available online: <https://www.sketchengine.eu/> (accessed on 6 June 2021).
- EnTenTen. Available online: [https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen13\\_tt2\\_1](https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen13_tt2_1) (accessed on 6 June 2021).
- Ban Kirigin, T.; Meštrović, A.; Martinčić-Ipšić, S. Towards a formal model of language networks. In *International Conference on Information and Software Technologies*; Springer International Publishing: Cham, Switzerland, 2015; pp. 469–479.
- Perak, B. 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: Newcastle upon Tyne, UK, 2017; pp. 299–319.
- Traag, V.; Waltman, L.; van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *arXiv* **2018**, arXiv:1810.08473.
- Perak, B. Emocije u korpusima: Konstrukcijska gramatika i graf metode analize izražavanja emotivnih kategorija. In *Emocije u hrvatskome jeziku, književnosti i kulturi. Zbornik radova 48. seminara Zagrebačke slavističke škole (Emotions in Croatian Language, Literature and Culture. Proceedings of the 48th Seminar of the Zagreb School of Slavic Studies)*; Filozofski fakultet Sveučilišta u Zagrebu i Zagrebačka slavistička škola—Hrvatski Seminar za Strane Slaviste: Zagreb, Croatia, 2020; pp. 100–120.
- Perak, B.; Ban Kirigin, T. Corpus-Based Syntactic-Semantic Graph Analysis: Semantic Domains of the Concept Feeling. *Raspr. Časopis Inst. Hrvat. Jez. Jezikoslovlje* **2020**, *46*, 493–532.
- Perak, B.; Damčević, K.; Milošević, J. O sranju i drugim neprimjerenim stvarima: Kognitivno-lingvistička analiza psovki u hrvatskome. In *Jezik i Njegovi učinci: Zbornik Radova s međunarodnoga Znanstvenog Skupa Hrvatskoga Društva za Primijenjenu*

- Lingvistiku Održanoga od 4. do 6. Svibnja 2017. Godine u Rijeci*; Diana, S., Vlastelić, A., Eds.; Hrvatsko Društvo za Primijenjenu Lingvistiku: Zagreb, Croatia, 2018; pp. 245–270.
19. Perak, B. 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., Pauković, D., Eds.; Routledge: London, UK, 2019; pp. 63–100.
  20. Perak, B. Developing the ontological model for research and representation of Commemoration Speeches in Croatia using a graph property database. In *Digital Humanities: Empowering Visibility of Croatian Cultural Heritage*; Cambridge University Press: Cambridge, UK, 2020; pp. 88–111.
  21. Sardelić, M.; Perak, B. Jealousy vs. Envy: European Cultural Background and Croatian Linguistic Examples. *Coll. Antropol.* **2021**, *45*, 55–66.
  22. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst.* **2006**, *1695*, 1–9.
  23. Bond, F.; Foster, R. Linking and extending an open multilingual wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, 4–9 August 2013; pp. 1352–1362.
  24. hrWac22. Available online: [https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fhrwac22\\_ws](https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fhrwac22_ws). (accessed on 6 June 2021).
  25. Sentic. Available online: <https://sentic.net/> (accessed on 6 June 2021).
  26. Sentic API. Available online: <https://github.com/yurimalheiros/senticnetapi> (accessed on 6 June 2021).
  27. Vilares, D.; Peng, H.; Satapathy, R.; Cambria, E. BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1292–1298.
  28. Baccianella, S.; Esuli, A.; Sebastiani, F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*; European Language Resources Association (ELRA): Valletta, Malta, 2010.
  29. Guerini, M.; Gatti, L.; Turchi, M. Sentiment analysis: How to derive prior polarities from SentiWordNet. *arXiv* **2013**, arXiv:1309.5843.
  30. Gatti, L.; Guerini, M.; Turchi, M. SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Trans. Affect. Comput.* **2015**, *7*, 409–421.
  31. Warriner, A.B.; Kuperman, V.; Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **2013**, *45*, 1191–1207.
  32. Menczer, F.; Fortunato, S.; Davis, C.A. *A First Course in Network Science*; Cambridge University Press: Cambridge, UK, 2020.
  33. Brin, S.; Page, L. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **1998**, *30*, 107–117.
  34. Cambria, E.; Poria, S.; Gelbukh, A.; Thelwall, M. Sentiment analysis is a big suitcase. *IEEE Intell. Syst.* **2017**, *32*, 74–80.
  35. ConceptNet. Available online: [conceptnet.io](http://conceptnet.io) (accessed on 6 June 2021).
  36. Geeraerts, D. *Theories of Lexical Semantics*; Oxford University Press: Oxford, UK, 2010.
  37. Timestamped Web Corpus of English 2014–2020. Available online: [https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Feng\\_jsi\\_newsfeed\\_virt](https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Feng_jsi_newsfeed_virt) (accessed on 6 June 2021).
  38. Panther, K.U.; Thornburg, L. Antonymy in language structure and use. In *Cognitive Linguistics Between Universality and Variation*; Mario, B., Ida, R., Milena Žic, F., Eds.; Cambridge Scholars Publishing: Newcastle upon Tyne, UK, 2012; pp. 159–186.
  39. Čulig Suknaić, J. Antonimija Kao Pojmovna Kategorija Značenjske Suprotnosti u Engleskome i Hrvatskome Jeziku. Ph.D. Thesis, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia, 2020.
  40. Kim, T.; Choi, J.; Edmiston, D.; Bae, S.; Lee, S.G. Dynamic compositionality in recursive neural networks with structure-aware tag representations. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6594–6601.