


Modeling Semantic Relations from a Dependency-Based Graph: A Corpus-Based Network Analysis of Croatian Parliamentary Debates

Benedikt Perak 
University of Rijeka
Rijeka, Croatia

Abstract

The following paper examines the application of graph technologies to parliamentary data. Using the debates that took place in the Croatian Parliament between 2003 and 2017 as an example, it demonstrates how natural language processing (NLP) tools, graph databases, and network algorithms can be used to conduct corpus statistical, stylometric, and semantic analysis. Special attention will be paid to the structure of morpho-syntactically tagged corpora, which are embedded in a property graph database that enables the exploration of corpus-specific semantic relations. As will be shown, such a structure allows parliamentary data to be empirically analyzed with regard to the communication, conceptualization, and framing of social identities, interactions, institutions, and cultural models.

1 Introduction

This paper examines the application of graph technologies to parliamentary data by means of natural language processing (NLP) techniques, the graph



Creative Commons License Attribution 4.0 International (CC BY 4.0).

In: Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.): Graph Technologies in the Humanities - Proceedings 2020, published at <http://ceur-ws.org>.

database Neo4j, and a Python implementation of igraph algorithms. It describes the embedding of a dependency-tagged corpus in a graph database model, and presents an innovative approach to the empirical linguistic analysis of the social identities, interactions, institutions, and cultural practices recorded in parliamentary data. The goal of the project is to establish an extensive knowledge base with a clearly structured ontology, which in turn enables data integration, data enrichment, and quantitative-qualitative analysis.

In addition to the application of standard corpus statistical methods, semantic and stylometric analysis of the syntactic dependency structures was carried out using the Universal Dependencies NLP parser (Straka and Straková, 2017). Conceptual profiling of parliamentary discourse was undertaken based on the semantic features of the syntactic dependency lexical matrix with the aid of graph algorithms involving network centrality and community measures.

By way of a case study, this paper considers records detailing Croatian parliamentary debates between 2003 and 2017. The following section (Section 2) discusses parliamentary data in general, while Section 3 elaborates on the genesis of the Croatian Parliamentary Corpus (CroParl). Section 4 presents a number of specific examples drawn from the said corpus, whereas the fifth and final section contains concluding remarks and some thoughts on the work that lies ahead.

2 Parliamentary Data

Parliamentary data constitutes a rich and publicly available resource that is inherently endowed with politically and historically significant information concerning the discourse between the political representatives of democratic systems of government and the socio-cultural interactions in which they participate.

In recent years, the improved accessibility of parliamentary data has made the democratic process increasingly transparent (Janssen, 2011; Andrews and da Silva, 2013; Granickas, 2014). Digital records of parliamentary debates have become an indispensable resource for computational data analysis in the humanities and social sciences (Glavaš et al., 2019; Berntzen et al., 2019; Hofmann et al., 2020). Moreover, there are a number of reasons why companies, private citizens, and public organizations may wish to engage with parliamentary data, including but not limited to creating business value, enabling local citizen value, addressing global societal challenges, and advocating the open data agenda (Lassinantti et al., 2019).

Yet the proper archiving, structuring, synchronization, and visualization

of the treasure trove of multimodal data that is generated over the course of the socially and institutionally highly complex interactions that characterize parliamentary debates are not without their pitfalls. For one, the diversity of the data management approaches of national and regional parliaments with respect to issues such as language, rhetorical strategies, actor properties, social features, and political context poses a significant obstacle to systematic scientific inquiry.

Such standards for the processing of textual data and the creation of structured corpora of parliamentary and other governmental debates as currently exist – along with a broad variety of metadata formalizations – are the result of the efforts of a small number of scholars based out of various European research centers. At the European level, the Digital Corpus of the European Parliament (DCEP) (Hajlaoui et al., 2014) is perhaps the most significant digital collection, while the CLARIN ERIC consortium assembles the available resources from a number of European national parliaments.¹ Within the ParlaCLARIN initiative, Erjavec and Pancur have laid out Text Encoding Initiative (TEI) guidelines for corpora of parliamentary proceedings (Erjavec and Pancur, 2019), with recommendations for the structure of the corpus, the encoding of metadata (including, for example, the speakers and the political parties to which they belong), speeches, and notes, and guidelines for linguistic annotation and integration of multimedia content.² However, despite these efforts at standardization, the community of researchers working with parliamentary data remains fragmented and in need of a unified platform.

Keeping these various factors in mind, the goal of this paper is to showcase a flexible information framework based on graph technologies that is capable of processing and integrating regional, national, and European parliamentary data. In so doing, I hope to demonstrate how current advances in natural language processing and data management can be harnessed to build a socio-linguistic parliamentary data network that is accessible not only to specialized scholars, but also to journalists, NGOs, and private citizens.

The working prototype for the web implementation of Croatian parliamentary data is available on the website hosted by the University of Rijeka's EmoCNet project.³

¹<https://www.clarin.eu>

²<https://github.com/clarin-eric/parla-clarin>

³<http://emocnet.uniri.hr/croparl>

3 Croatian Parliamentary Corpus

A significant portion of parliamentary data is made up of texts that detail political debates among representatives. This corpus of speech acts allows researchers, particularly linguists and political scientists, to engage in different kinds of semantic and pragmatic analysis.

Here I am concerned with the debates that took place in the Croatian Parliament between 2003 and 2017, during the fifth to ninth parliamentary assemblies. As can be seen in Figure 1, the process of corpus creation in this instance consisted of data gathering, NLP parsing, data integration, and data management in the property graph database (Perak and Rodik, 2018).

3.1 Data Gathering and Integration

The data was gathered using a Selenium scraper⁴ from the Croatian Parliament web repository.⁵ Data pertaining to the debates that took place during the fifth to the ninth parliamentary assemblies is to be found in two separate datasets: one containing values that relate to the assemblies, sessions, and topics, and another containing values that concern persons, debate transcripts, topic IDs, and announcement metadata.

3.2 NLP Parsing

Extracted from 390,000 transcripts, the various speech acts of the representatives in question were processed using the UDPipe Natural Language Processing Toolkit (Straka et al., 2016), which includes features such as tokenization, part-of-speech tagging, lemmatization, and dependency parsing. Parsing was done using the Universal Dependencies 2.0 Model (UDPipe repository)⁶ developed in Croatia (Agić and Ljubešić, 2015).

The NLP parsing created over 4 million sentences and 70 million word tokens in the CoNLL-U output format (Figure 2). The morpho-syntactic metadata appears in 10 tab-separated value fields:

1. ID: word index
2. FORM: word form or punctuation symbol
3. LEMMA: lemma or stem of word form
4. UPOS: universal part-of-speech tag
5. XPOS: language-specific part-of-speech tag
6. FEATS: list of morphological features from the universal feature inventory

⁴<https://github.com/ropensci/RSelenium>

⁵The data gathering process is published under the GitHub handle <https://github.com/rodik/Sabor>. The Croatia Parliament web repository can be accessed via <http://edoc.sabor.hr/>

⁶<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2364>

7. HEAD: head of the current word
8. DEPREL: universal dependency relation to the HEAD
9. DEPS: enhanced dependency graph
10. MISC: any other annotation

3.3 Data Integration

The data obtained from the Croatian Parliament was then integrated with the CoNLL-U formatted results Straka et al. (2016) in the property graph database Neo4j (Webber, 2012). Using the existing data categories of the Croatian parliamentary records in conjunction with parsed linguistic structures, an informational graph structure was created, in which labels, nodes, node properties, relationships, and relationship properties form an ontology of the various entities involved in the parliamentary debates (Figure 3). For example, labels represent instances of ontologically similar entities, while links represent their inter-structural connections. Labels also have as instances nodes that can store their relational properties in the form of key-value metadata structures, which is one of the major advantages of employing a property graph database.

Structurally complex entities are metonymically related to ontologically less complex entities: *Assembly* can be narrowed down to *Session*, which can be narrowed down to *Topic*, which can be narrowed down to *Utterance*, which can be narrowed down to *Sentence*, which can be narrowed down to *Token*. Utterances are related to a *Representative*, who is socially related to a *Party* and a *Parliamentary Club*. Some labels have inter-structural relationships: tokens, sentences, and utterances are *sequential* in nature, while tokens also have syntactic *dependency* relations.⁷

Arranging the data in question in a property graph has a number of advantages. First, it enables the integration of various datasets into a single framework. Second, labels, nodes, relations, and their properties can be easily updated or remodeled according to newly adopted standards, and the structure of the graph can be enriched with additional knowledge resources. Last but not least, the user-friendly graph representation of the information ontology enables digital humanities scholars to intuitively develop new approaches to their material based on the interrelation of data structures.⁸

⁷For details concerning the data integration and NLP parsing process, see Perak and Rodik (2018).

⁸The tagged CroParl graph database's Neo4j dump is a case in point, <https://drive.google.com/file/d/1zRy3EmwPrb4r3vGM3J5N5bpeKwCLEaHF/view?usp=sharing>

4 Corpus Analysis

4.1 Statistical Summarization of Lexical Usage

The CroParl corpus contains data from five parliamentary assemblies and covers 5,599 topics, which were broached by 895 members of parliament belonging to 42 political parties. In the process of NLP parsing, 390,078 utterances were related to 4 million sentences and 70 million tokens.

4.1.1 Lexical Summarization

The graph-based lexical store thus enables standard lexical summarizations based on tokens, lemmas, or part-of-speech counts, and their respective proportions. For instance, Table 1 represents the lemma count for nouns in the corpus.

The corpus can also be used for the representation of a keyword in context (KWIC). The KWIC solution offered on the official site (<https://edoc.sabor.hr/>) relies on a string-based query search. This type of search is relatively easy to implement from a technical standpoint, and is suitable for morphologically lean languages. However, this is not the case for Croatian, which has a complex declination and inflection system. Consequently, a lemma-based query yields much more accurate results in the KWIC concordance. Table 2 presents an example of sentence level results for the noun *ljubav* ('love') using a lemma-based KWIC query. The results are based on the *lemma*='ljubav' and part-of-speech *pos*='noun' properties assigned to the tokens.

4.1.2 Stylometric Summarization

Stylometric summarization based on the linguistic behaviour of parliamentary representatives can provide us with valuable insights into their individual linguistic profiles, stylistic idiosyncrasies, and similarities compared to other speakers (Amancio, 2015). As I will demonstrate in this section, utterance analysis and word/concept counts are especially illuminating in this regard.

The number of utterances of a speaker, for example, corresponds fairly directly to their prominence within the parliament. Table 3 shows a list of the 15 representatives with the highest number of utterances in the CroParl corpus, with the overwhelming majority belonging to Vladimir Šeks, who served as President of the Croatian Parliament from 2003 to 2008. A quick examination of Table 3 reveals that the two runners-up, Bebić and Leko, also served in this role, which, by its formal nature, involves numerous speech acts with low token counts per utterance.

Prominence can also be measured by other lexical features, such as lemmas per representative, part-of-speech count per representative, etc., and

the data management structure enables various types of filtering (e.g. utterances/tokens/lemmas per date/assembly/session/topic).

Another application of corpus-based stylometric analysis is the frequency with which a certain lexeme is used by a specific representative. For instance, we might be interested in how often emotionally charged lexemes were used, and by whom. By way of example, Tables 4 and 5 show a portion of the data for two such terms: love and peace.

Here we see that the lexeme *mir* ('peace') was used almost five times more often than the lexeme *ljubav* ('love'). Column "F" represents the frequency with which the respective lexeme occurred in the representative's utterances, column "p in All" represents the overall proportion of the lemma in the corpus, column "Auth" the number of tokens per representative, and column "p in Auth" the proportion in which the lemma occurred in speeches by the representative. While the value "p in All" can be interpreted as an indicator for the representative's conceptual influence on the debates, the value "p in Auth" represents the importance assigned to a particular concept by a given speaker.

This type of analysis opens up a whole range of possibilities when it comes to further studies concerning the significance of specific concepts to individual representatives. Stylometric linguistic analysis could also be used to interrogate the socio-, pragma-, and cross-linguistic profiles of a given political party or parliamentary club based on how often and in which contexts lexemes such as 'love' and 'peace' are used by its members.

4.2 Dependency-Based Semantic Network Analysis

The NLP parsing of the texts in question created 74,968,809 syntactic dependency relations between tokens. The structure of the relations tagged by the Universal Dependencies parser is shown in Table 6.

The graph structure of these relations enables a dependency-based type of semantic analysis that is conducive to a number of NLP tasks, including the discovery of corpus-specific word senses.

4.2.1 Semantic Domain Induction Based on "Conj" Dependency

Word sense disambiguation and word sense identification have been important tasks from the earliest days of natural language processing research (Schütze, 1998; Ide and Véronis, 1998; Navigli, 2009). NLP studies have repeatedly demonstrated the usefulness of dependency-tagged corpora when it comes to the unsupervised assembling of semantic knowledge. The coordination structure labeled as *conj* in the UD framework, which expresses an asymmetrical dependency relation between two elements that are connec-

ted by a coordinating conjunction, such as *and*, *or*, etc., has proved particularly useful for distinguishing between associated classes of concepts (Widdows and Dorow, 2002; Cederberg and Widdows, 2003; Widdows, 2003).

In the method being outlined here, the iterating graph algorithm is used to produce an undirected graph from all the nouns collocated with the construction dependency, operating as a kind of lexical embedding that represents the semantic structure of the concepts found in the Croatian Parliamentary corpus. The coordination dependency graph can be used to identify ontologically similar lexemes and related semantic domains for a given source lexeme:

- choose a source lexeme and identify n number of the most frequently collocated target lexemes
- construct a lexical network with associated lexemes as nodes and coordination-based weighted relations
- detect lexically coherent communities with sub-graphs as a representation of the semantic domains, and use parametrizable granularity as a measure of the level of categorical consistency vs continuity and abstraction
- analyze the distribution of prototypical association patterns as an indication of cross-cultural framing and cross-linguistic variations

Graph analysis, meanwhile, is performed with the help of the Python implementation of *igraph*, which is used to identify, measure, and visualize the conceptual framing. The procedure starts with extracting collocations of an arbitrary source lexeme. For instance, the first 50 most common collocates for the lexeme *mir* ('peace') are represented in Figure 4.

For the source lexeme *mir* with $n = 50$ first order lexemes, the structural function of a friend coordination network is enhanced by identifying $n=50$ second order lexemes in a friend-of-a-friend (FoF) pattern. By analyzing the subgraphs created by the second order coordinated dependency noun collocates we can distinguish the semantic clusters that indicate the sense association of the source lexeme.

The FoF network with 945 nodes was pruned with regards to the node degree measure $degree > 4$ in order to filter out the less interconnected lexemes from the semantic network. As can be seen in Figure 5, the resulting FoF network for *mir* contained 120 nodes, which were then clustered using the Leiden clustering algorithm (Traag et al., 2019). The resulting network with six clusters reveals the corpus-specific structure of the semantically associated concepts and domains, as can be seen in Table 7. If needed, modifications of the resolution parameter can render more fine-grained or more robust communities.

Finally, the prominence of the associative conceptual profiling of the source lexeme *mir* can be discerned from a speaker's weighted contribution to the coordination dependency graph. The weighted graph was constructed using the collocates that are in a coordination dependency with the source lexeme *mir* used by each representative.

The resulting graph with 461 nodes was pruned to highlight the more prominent associations using weighted degree > 4 settings. The weighted degree represents the sum of weights assigned to the node's connections. By filtering out the nodes by weighted degree, we were able to focus on the lexemes and representatives that saliently contribute to the association matrix of the lexeme *mir* by their occurrence. As is shown in Figure 6, the pruned associative network with 97 elements was then clustered using the cpm resolution 0.34, which yielded 50 clusters. The 12 most prominent association clusters of the lexeme *mir* can be found in Table 8.

Clearly, then, graph measures and network representation of socio-lexical phenomenon can be used to glean an empirical, yet intuitive, understanding of the complex weighted structural relations between lexical content and the agents who generate this content. Of particular interest here is the question of which lexemes exhibit the highest number of associations, and who introduced these associations. Figure 6 shows that *stabilnost* ('stability'), *sigurnost* ('safety'), and *red* ('order') are central for understanding the associative conceptual content of the lexeme *mir*. The clusters, represented with different colors in Figure 6 and transcribed in Table 8, reveal the agents who most frequently made use of the lexemes in question. Interestingly, most of the agents connected with the concepts 'safety' and 'order' have served as members of institutions responsible for the nation's defense. For instance, both Ivan Šantek and Tomislav Čuljak were members of the committee on Internal Policy and National Security. Šantek was also a member of the Defense Committee from 2013, while Berislav Rončević served as its vice chairman.

Conversely, the concept 'peace' has a different connotation when coupled with the lexemes 'stability' and 'cooperation' – the agents in this cluster seem to be more concerned with international relations. Jozo Radoš, for example, was vice-chairman of the Committee on European Affairs and a member of the Committee on European Integration, while Davor Ivo Stier was a member of the Committee on Interparliamentary Cooperation, the Foreign Policy Committee, and the Committee on European Affairs. The same holds true for Stier's fellow part member, Andrej Plenković, who would go on to become prime minister. Similarly, the aspect of 'prosperity' in connection with 'peace' was promoted with particular vigor by the deputy chairman

of the Committee on Croats outside the Republic of Croatia, Boro Grubišić, and by a member of the Committee on Regional Development and European Union Funds, Petar Baranović. The antonym ‘war,’ meanwhile, was prominently associated with ‘peace’ by the president of the Delegation of the Croatian Parliament to the NATO Parliamentary Assembly, Krešimir Čosić.

Although these results were acquired from a complex set of queries and graph algorithms, they intuitively represent the conceptual associative dimension of the lexeme *mir*, and as such can help us to begin to understand how conceptual relations are influenced and motivated by the political views of the speakers in question and their respective institutional functions. The possibility of such an insight demonstrates the true value of graph-based socio-linguistically and morpho-syntactically tagged corpus analysis: it allows us not only to simply count lexical items or allocate frequencies of syntactical relations, but to explore the socio-cognitive aspects of conceptualization by revealing a dynamic pragmatic context that is determined by the graph structure of the nodes and the relationships that link them together.

5 Conclusion

In this essay, I have introduced a graph-based data management and analysis framework that seeks to integrate parliamentary data with a NLP dependency tagged corpus. The proposed framework can ingest multiple data formats with disparate internal structures, while producing flexible and intuitive information structures that are highly conducive to comparative data analysis. Making use of the readily available Neo4j database, NLP tools, and a Python implementation of graph algorithms, it can easily be adapted to a wide variety of discipline-specific research approaches and customized to meet the requirements of individual researchers.

In our continued work on parliamentary data analysis, we plan to: a) harvest recent parliamentary data, b) implement new NLP tools for the Croatian language provided by the CLASSLA initiative, and include named-entity recognition (NER) data in the data tagging, c) carry out further research on stylometric analysis, particularly with regard to its diachronic dimension, d) integrate data from external sources of parliamentary data (<https://edoc.sabor.hr>, wikipedia, etc.), and e) extend existing semantic graph research to other dependency relations.

Special attention will be paid to further enhancing the web infrastructure for Croatian parliamentary data,⁹ which has recently been established with the help of funding from the University of Rijeka and the University of

⁹<http://emocnet.uniri.hr/croparl>

Zagreb University Computing Centre (SRCE), as such resources are essential for establishing a platform for collaborative scholarship on parliamentary data that encourages the exchange of information at an international level.

Acknowledgements

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

Figures and Tables

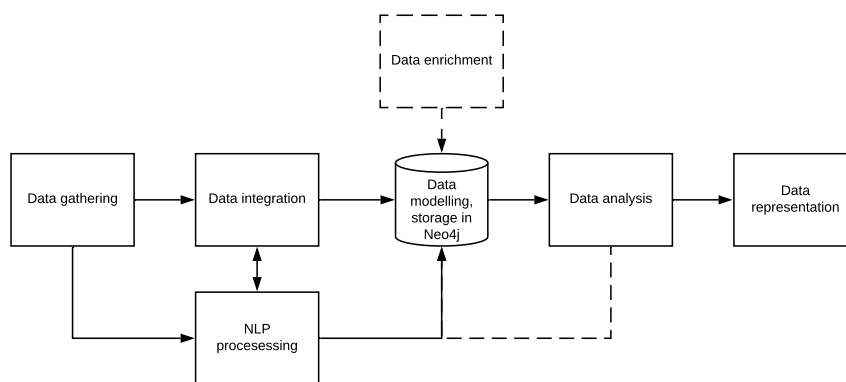


Figure 1: Croatian parliamentary data processing pipeline

```

# newpar
# sent_id = 1
# text = Hvala lijepo, predsjedniku Odbora za Ustav, Poslovnik i politički sustav.
1 Hvala hvala NOUN _ Case=Nom|Gender=Fem|Number=Sing 0 root _ _
2 lijepo lijepo ADV _ Degree=Pos4 advmod _ SpaceAfter=No
3 , , PUNCT _ 2 punct _ _
4 predsjedniku predsjednik NOUN _ Case=Dat|Gender=Masc|Number=Sing 1 obj _ _
5 Odbora odbor NOUN _ Case=Gen|Gender=Masc|Number=Sing 4 nmod _ _
6 za za ADP _ Case=Acc 7 case _ _
7 Ustav ustav NOUN _ Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing 5 nmod _ SpaceAfter=No
8 , , PUNCT _ 9 punct _ _
9 Poslovnik Poslovnik PROPN _ Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing 5 conj _ _
10 i i CCONJ _ 12 cc _ _
11 politički politički ADJ _ Animacy=Inan|Case=Acc|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing 12 amod
12 sustav sustav NOUN _ Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing 5 conj _ SpaceAfter=No
13 . . PUNCT _ 1 punct _ _
  
```

Figure 2: A sample CoNLL-U output of the UD Pipe parser

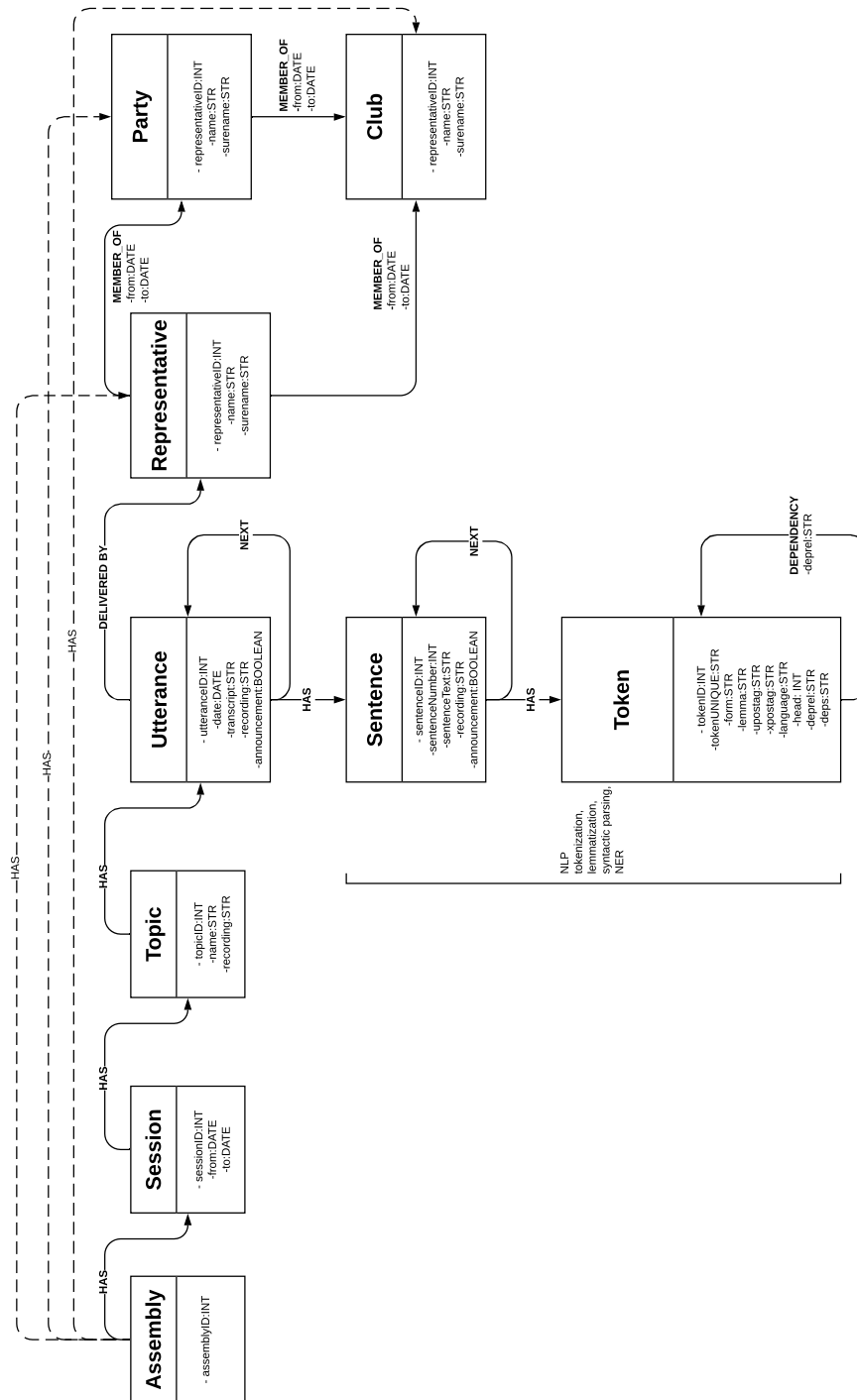


Figure 3: Informational ontology with node labels, relation labels, and their key:value properties: (Assembly) -[HAS]-> (Session) -[HAS]-> (Topic) -[HAS]-> ((Utterance) -[HAS]-> (Sentence) -[HAS]-> (Token) -[SEQUENCE]-> (Token) -[DEPENDENCY]-> (Token)) -[DELIVERED BY]-> ((Representative) -[MEMBER OF]-> (Club) ,-[MEMBER OF]-> (Political party)).

| Rank | Lemma | Eng. | Count |
|------|--------------|----------|---------|
| 1 | problem | problem | 100,684 |
| 2 | ministar | minister | 94,195 |
| 3 | građanin | citizen | 91,425 |
| 4 | sustav | system | 89,971 |
| 5 | rad | labor | 87,245 |
| 6 | ministarstvo | ministry | 85,868 |
| 7 | klub | club | 85,413 |
| 8 | proračun | budget | 85,325 |
| 9 | vrijeme | time | 80,573 |
| 10 | replika | reply | 80,302 |
| ... | ... | ... | |

Table 1: Most frequent noun lemmas

| Person | Example |
|--------------------------|--|
| Jurjević, Marin | Htio bih odmah na početku kazati pošto govorimo o djeci, a ponekad i u nekim raspravama ima dosta politiziranosti, htio bih kazati da se djeca ne rađaju ni radi politike, ni radi države, ni radi klase, ni radi nacije, da se rađaju radi <i>ljubavi</i> kao što je netko već jutros rekao i da je to posljedica možda jednog od posljednjih prirodnih odnosa između muškarca i žene, a koji su opet posljedica <i>ljubavi</i> . |
| Sučec-Trakoštanec, Ivana | Dakle, ja mislim da je žena kao majka, žena kao dio obitelji i ne znam zašto su se ljudi smijali kada se govorilo o <i>ljubavi</i> pa ja onda ne moram ni <i>ljubav</i> spomenuti ali dakle, žena koja izabere da će sa svojim suprugom barem u prijateljstvu imati određeni broj djece a usput završi fakultet, usput doktorira, usput i radi nazadna ja mislim da je ona vrlo slobodna osoba, da je ona suvremena i da je ona ravnopravnija od kolega muškaraca jer emancipacija se živi a ne priča... |
| ... | ... |

Table 2: Lexeme *ljubav* ('love') in context as used by two parliamentary representatives

| Rank | Representative | Utterance Count |
|------|---------------------|-----------------|
| 1 | Šeks, Vladimir | 45,635 |
| 2 | Bebić, Luka | 22,770 |
| 3 | Leko, Josip | 18,825 |
| 4 | Zgrebec, Dragica | 18,454 |
| 5 | Stazić, Nenad | 16,293 |
| 6 | Reiner, Željko | 16,224 |
| 7 | Batinić, Milorad | 10,486 |
| 8 | Jandroković, Gordan | 9,673 |
| 9 | Jarnjak, Ivan | 9,044 |
| 10 | Milinović, Darko | 6,496 |
| 11 | Brkić, Milijan | 5,342 |
| 12 | Petrov, Božo | 5,039 |
| 13 | Šuker, Ivan | 4,659 |
| 14 | Friščić, Josip | 3,780 |
| 15 | Grubišić, Boro | 3,736 |
| ... | ... | |

Table 3: Utterances per representative

| | Name | F | p in All | Auth | p in Auth |
|----|-----------------------------|----|----------|-----------|-----------|
| 1 | Antičević Marinović, Ingrid | 56 | 7.13e-07 | 1,513,103 | 3.7e-05 |
| 2 | Jurjević, Marin | 32 | 4.07e-07 | 226,649 | 0.000141 |
| 3 | Kosor, Jadranka | 29 | 3.69e-07 | 756,311 | 3.8e-05 |
| 4 | Lalić, Ljubica | 25 | 3.18e-07 | 305,994 | 8.2e-05 |
| 5 | Sumrak, Đurđica | 22 | 2.8e-07 | 308,075 | 7.1e-05 |
| 6 | Marić, Goran | 20 | 2.55e-07 | 569,439 | 3.5e-05 |
| 7 | Letica, Slaven | 18 | 2.29e-07 | 216,754 | 8.3e-05 |
| 8 | Beus Richebmergh, Goran | 17 | 2.16e-07 | 575,021 | 3e-05 |
| 9 | Pernar, Ivan | 16 | 2.04e-07 | 461,554 | 3.5e-05 |
| 10 | Grubišić, Boro | 15 | 1.91e-07 | 1,484,462 | 1e-05 |

Table 4: Lexeme *ljubav* ('love') (corpus count = 908, frequency = 11.56 per million words) per representative

| | Name | F | p in All | Auth | p in Auth |
|----|---------------------|-----|-----------|-----------|-----------|
| 1 | Šeks, Vladimir | 560 | 7.13e-06 | 1,247,153 | 0.000449 |
| 2 | Tafra, Višnja | 224 | 2.852e-06 | 95,950 | 0.0023345 |
| 3 | Kajin, Damir | 127 | 1.617e-06 | 2,171,645 | 5.85e-05 |
| 4 | Đakić, Josip | 125 | 1.592e-06 | 501,417 | 0.0002493 |
| 5 | Brkić, Milijan | 110 | 1.401e-06 | 141,999 | 0.0007747 |
| 6 | Jandroković, Gordan | 86 | 1.095e-06 | 404,767 | 0.0002125 |
| 7 | Pernar, Ivan | 85 | 1.082e-06 | 461,554 | 0.0001842 |
| 8 | Špoljar, Dunja | 84 | 1.07e-06 | 60,456 | 0.0013894 |
| 9 | Čosić, Krešimir | 83 | 1.057e-06 | 242,607 | 0.0003421 |
| 10 | Zgrebec, Dragica | 79 | 1.006e-06 | 699,250 | 0.000113 |

Table 5: Lexeme *mir* ('peace') (corpus count = 4665, frequency = 59.4 per million words) per representative

| | Nominals | Clauses | Modifier | Function |
|---------------------|---------------------------------------|----------------------------------|----------------------------------|----------------------|
| Core Arguments | nsubj nsubj obj iobj | csubj csubj ccomp xcomp | | |
| Non-Core Dependents | obl vocative expl dislocated | advmod | aux discourse | cop mark |
| Nominal Dependents | nmod appos nummod | acl | amod | det clf case |
| Coordination | MWE | Loose | Special | Other |
| conj cc | fixed flat compound | list parataxis | orphan goeswith reparandum | punct root dep |

Table 6: Structure of syntactic relations

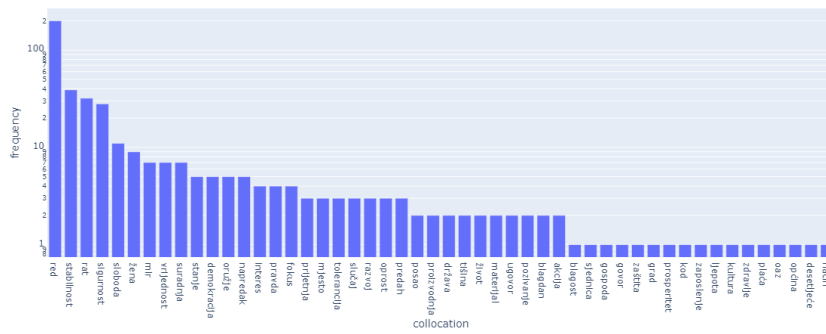


Figure 4: Fifty most frequent collocates for the lexeme *mir* ('peace')

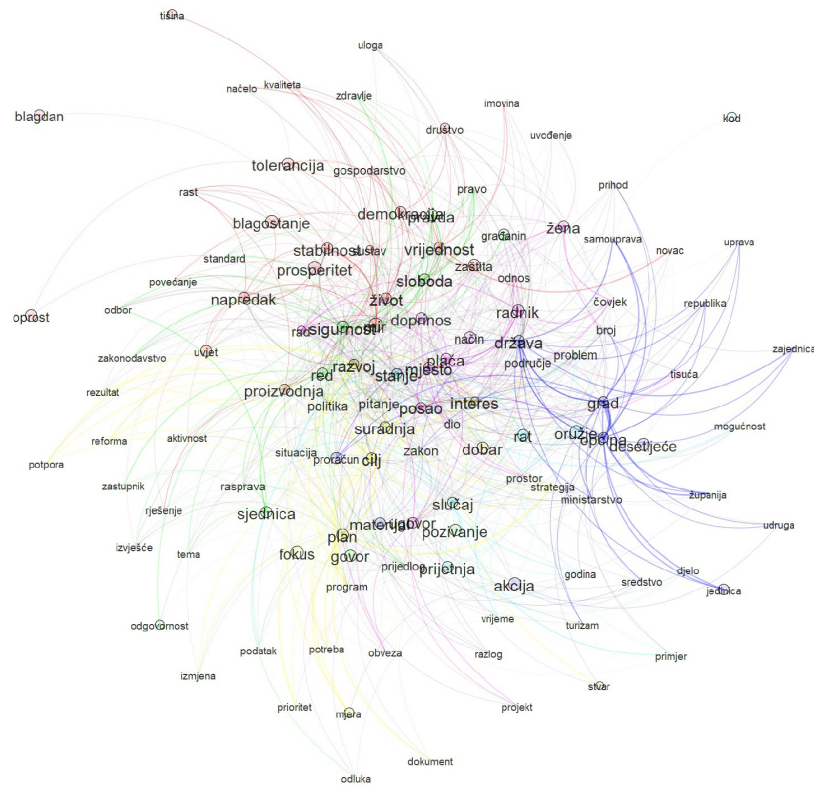


Figure 5: FoF network for *mir* ('peace')

| C | Nodes | Eng |
|---|---|--|
| 1 | mir, stabilnost, vrijednost, demokracija, napredak, tolerancija, oprost, proizvodnja, blagdan, život, tišina, prosperitet, blagostanje, rast, sustav, gospodarstvo, zaštita, kvaliteta, aktivnost, povećanje, uvjet, društvo, novac, imovina, načelo, uloga, rješenje | peace, stability, value, democracy, progress, tolerance, forgiveness, production, holiday, life, silence, prosperity, well-being, growth, system, economy, protection, quality, activity, increase, condition, society, money, property, principle, role, solution |
| 2 | red, sigurnost, sloboda, pravda, pozivanje, govor, sjednica, zastupnik, rasprava, tema, prijedlog, pitanje, izvješće, standard, zdravlje, pravo, zakonodavstvo, građanin, odbor, odgovornost, odluka | order, security, freedom, justice, calling, speech, session, representative, debate, topic, proposal, question, report, standard, health, law, legislation, citizen, committee, responsibility, decision |
| 3 | materijal, država, akcija, desetljeće, općina, grad, ministarstvo, zakon, proračun, prihod, samouprava, županija, razlog, zajednica, broj, republika, sredstvo, uprava, jedinica, turizam, udruga | material, state, action, decade, municipality, city, ministry, law, budget, revenue, self-government, county, reason, community, number, republic, means, administration, unit, tourism, association |
| 4 | suradnja, interes, fokus, razvoj, plan, dobar, cilj, područje, reforma, izmjena, politika, strategija, program, potpora, prioritet, rezultat, potreba, stvar, mjera, dokument | cooperation, interest, focus, development, plan, good, goal, area, reform, change, policy, strategy, program, support, priority, result, need, thing, measure, document |
| 5 | žena, mjesto, ugovor, posao, plaća, doprinos, radnik, rad, način, uvođenje, odnos, čovjek, dio, prostor, tisuća, obveza, projekt | woman, place, contract, job, salary, contribution, worker, work, manner, introduction, relationship, man, part, space, thousand, obligation, project |
| 6 | rat, stanje, oružje, prijetnja, slučaj, kod, problem, vrijeme, situacija, godina, primjer, djelo, mogućnost, podatak | war, condition, weapon, threat, case, code, problem, time, situation, year, example, work, possibility, data |

Table 7: Clusters of the FoF network for *mir* ('peace') pruned by *degree* < 4, *clustering*=Louvain, *method*=mvp

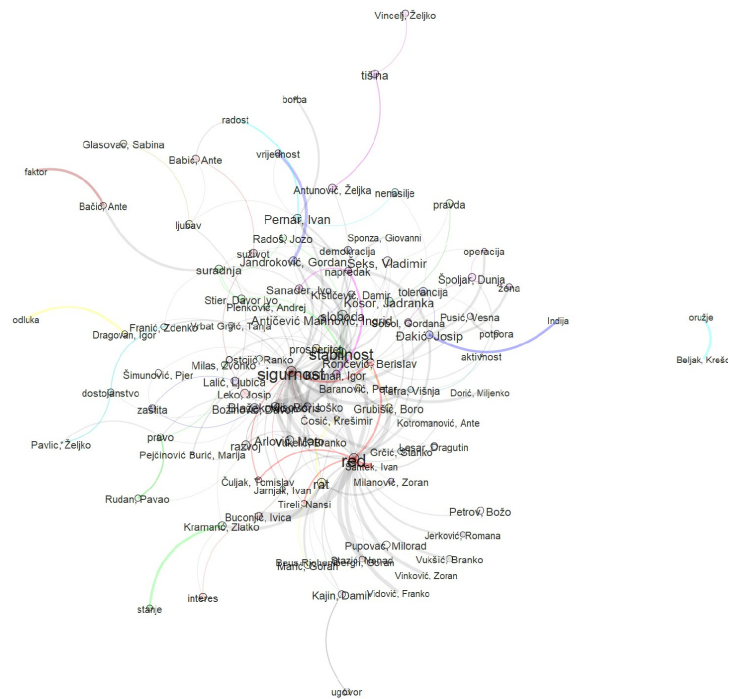


Figure 6: Associative conceptual profiling of the concept *mir* through the coordinated dependencies of representatives

| C | Nodes |
|----|--|
| 1 | <i>red</i> 'order', <i>sigurnost</i> 'safety', Šantek Ivan, Rončević Berislav, Čuljak Tomislav, Tireli Nansi |
| 2 | <i>stabilnost</i> 'stability', <i>suradnja</i> 'cooperation', Radoš Jozo, Stier Davor Ivo, Plenković Andrej |
| 3 | Indija 'India', <i>tolerancija</i> 'tolerance', Đakić Josip |
| 4 | Ćosić Krešimir, <i>rat</i> 'war', Marić Goran |
| 5 | Kolman Igor, <i>napredak</i> 'progress', Sanader Ivo |
| 6 | Pernar Ivan, <i>radost</i> 'joy', <i>nenasilje</i> 'nonviolence' |
| 7 | <i>sloboda</i> 'freedom', Krstičević Damir, Šeks Vladimir |
| 8 | <i>suživot</i> 'coexistence', Leko Josip, Babić Ante |
| 9 | Rudan Pavao, <i>pravo</i> 'law', Milas, Zvonko |
| 10 | Lalić Ljubica, <i>zaštita</i> 'protection', Božinović Davor |
| 11 | Grubišić Boro, Baranović Petar, <i>prosperitet</i> , 'prosperity' |
| 12 | Antunović Željka, <i>tišina</i> 'silence', Vincelj Željko |

Table 8: Prominent associations of the lexeme *mir*: 97 nodes and 50 clusters pruned by *weighted degree*<4, *clustering*=Louvain, *method*=cpm, *resolution*=0.34

References

- Agić, Ž. and Ljubešić, N. (2015). Universal Dependencies for Croatian (that work for Serbian, too). In *The 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, pages 1–8, Hissar. INCOMA Ltd. Shoumen, <https://aclanthology.org/W15-5301.pdf>.
- Amancio, D. R. (2015). A Complex Network Approach to Stylometry. *PLOS ONE*, 10(8):1–21, DOI: 10.1371/journal.pone.0136076.
- Andrews, P. and da Silva, F. S. C. (2013). Using Parliamentary Open Data to Improve Participation. In *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance*, pages 242–249. DOI: 10.1145/2591888.2591933.
- Berntzen, L., Johannessen, M. R., Andersen, K. N., and Crusoe, J. (2019). Parliamentary Open Data in Scandinavia. *Computers*, 8(3):65, DOI: 10.3390/computers8030065.
- Cederberg, S. and Widdows, D. (2003). Using LSA and Noun Coordination Information to Improve the Recall and Precision of Automatic Hyponymy Extraction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, volume 4, pages 111–118. DOI: 10.3115/1119176.1119191.
- Erjavec, T. and Pancur, A. (2019). Parla-CLARIN: TEI Guidelines for Corpora of Parliamentary Proceedings. In *Book of Abstracts of the TEI2019: What is text, really? TEI and beyond*, number 157. <https://gams.uni-graz.at/o:tei2019.bookofabstracts>.
- Glavaš, G., Nanni, F., and Ponzetto, S. P. (2019). Computational Analysis of Political Texts: Bridging Research Efforts Across Communities. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Tutorial abstracts*, pages 18–23, Florence, Italy. Association for Computational Linguistics, DOI: 10.18653/v1/P19-4004.
- Granickas, K. (2014). Open Data as a Tool to Fight Corruption. Technical Report 4, European Public Sector Information Platform.
- Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., et al. (2014). DCEP-Digital Corpus of the European Parliament. In *LREC 2014 (Language Resources and Evaluation Conference)*, pages 3164–3171. http://www.lrec-conf.org/proceedings/lrec2014/pdf/943_Paper.pdf.

- Hofmann, K., Marakasova, A., Baumann, A., Neidhardt, J., et al. (2020). Comparing Lexical Usage in Political Discourse Across Diachronic Corpora. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 58–65, Marseille, France. European Language Resources Association, <https://aclanthology.org/2020.parlaclarin-1.11>.
- Ide, N. and Véronis, J. (1998). Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1):1–40, DOI: 10.5555/972719.972721.
- Janssen, K. (2011). The Influence of the PSI Directive on Open Government Data: An Overview of Recent Developments. *Government Information Quarterly*, 28(4):446–456, DOI: 10.1016/j.giq.2011.01.004.
- Lassinantti, J., Ståhlbröst, A., and Runardotter, M. (2019). Relevant Social Groups for Open Data Use and Engagement. *Government Information Quarterly*, 36(1):98–111, DOI: 10.1016/j.giq.2018.11.001.
- Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69, DOI: 10.1145/1459352.1459355.
- Perak, B. and Rodik, F. (2018). Building a Corpus of the Croatian Parliamentary Debates Using UDPipe Open Source NLP Tools and Neo4j Graph Database for Creation of Social Ontology Model, Text Classification and Extraction of Semantic Information. In Fišer, D. and Pančur, A., editors, *Conference on Language Technologies & Digital Humanities*. <https://www.bib.irb.hr/960280>.
- Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123, DOI: 10.5555/972719.972724.
- Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA), <https://aclanthology.org/L16-1680>.
- Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 With UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver. Association for Computational Linguistics, DOI: 10.18653/v1/K17-3009.

- Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Scientific Reports*, 9(1):1–12, DOI: 10.1038/s41598-019-41695-z.
- Webber, J. (2012). A Programmatic Introduction to Neo4j. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, pages 217–218, New York, NY. Association for Computing Machinery, DOI: 10.1145/2384716.2384777.
- Widdows, D. (2003). Unsupervised Methods for Developing Taxonomies by Combining Syntactic and Statistical Information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 276–283. <https://aclanthology.org/N03-1036>.
- Widdows, D. and Dorow, B. (2002). A Graph Model for Unsupervised Lexical Acquisition. In *COLING 2002: The 19th International Conference on Computational Linguistics*, volume 1, pages 1–7. DOI: 10.3115/1072228.1072342.