Korpusna analiza sintaktičko-semantičkih struktura s pomoću grafova: semantičke domene pojma *osjećaj*

Perak, Benedikt; Ban Kirigin, Tajana

Source / Izvornik: **Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, 2020, 46, 957 - 996**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.31724/rihjj.46.2.27

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:186:008677

UDK 811.163.42:811.111'322
Prethodno priopćenje
Rukopis primljen 5. XI. 2019.
Prihvaćen za tisak 10. VI. 2020.
doi.org/10.31724/rihjj.46.2.27

Benedikt Perak
Cultural Studies Department, Faculty of Humanities and Social Sciences, University of Rijeka
Sveučilišna avenija 4, HR-51000 Rijeka
bperak@uniri.hr

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

CORPUS-BASED SYNTACTIC-SEMANTIC GRAPH ANALYSIS: SEMANTIC DOMAINS OF THE CONCEPT *FEELING*

This research[1] exemplifies the corpus-based graph approach to the syntactic-semantic analysis of a concept feeling using the Construction Grammar Conceptual Network methodology. By constructing a lexical network from grammatically tagged collocations of the English and the Croatian web corpora, the structure of the semantic domains is revealed as a set of sub-graphs derived from the source lexeme's friend-of-a-friend graph. The sub-graph structures, calculated with the community detection algorithm, are interpreted as the semantic domains associated with the source lexeme's conceptual matrix. Lexical structures are analyzed using a centrality algorithm that determines the overall rank of the salience and semantic relatedness to the source concept *feeling*. This empirical approach can be used for developing NLP methods and tasks, such as computing semantic similarity, sense disambiguation, sense structuring, as well as for comparative corpus and cross-cultural studies. ConGraCNet has a web application on the page http://emocnet.uniri.hr/congracnet.

[1] 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 1408: www.emocnet.uniri.hr.

1. Introduction

This article demonstrates a graph approach to the syntactic-semantic analysis of the conceptualization of the psychological domain lexicalized as *feeling* in English and *osjećaj* (Eng. feeling) in Croatian using the Constructional Grammar Conceptual Network (ConGraCNet) methodology. We describe a lexicographic application of the corpus-based graph analysis of the coordinated syntactic-semantic construction. The application of the method (ConGraCNet) is available as a research web-app at http://emocnet.uniri.hr/congracnet. The application uses the syntactic relations of syntactically tagged corpora to represent various semantic relations in terms of a network structure. In this case, we present the graph analysis of a coordinated construction. The coordinated construction (Van Oirsouw 2019) conceptualizes the associability and conjunction of two or more entities, properties, or processes as in the example: *this can have an impact on your emotions and behavior*. The associability is marked with the use of logical *and*/*or* operators. By extracting the lexemes in the coordinated construction from the corpus we can rank the most strongly collocated lexemes. Typically, the collocated lexemes are associated with the members of the same categorical class but appear to relate to different conceptual abstractions, i.e. members of the hypernym or hyponym classes. For instance, the source lexeme *feeling* can be collocated with other cognition concepts of the same ontological order such as *belief*, *perspective*, but also with types of feelings, i.e. its subordinated lexemes: *hate*, *love*, *hope*. Our methodology starts with the presumption that lexical collocates in the coordination construction represent structures of the associative concepts (Dorow and Widdows 2003). These relations form networks of ontologically associated entities that could reveal different levels of classification granularity, i.e. categorical knowledge. The methodological problem we want to tackle is how to identify and classify different associated semantic domains by simply using the graph analysis of the syntactic relation. The proposed method constructs the lexical network from second-order collocates (friend-of-a-friend) of the source node feeling in the coordinated syntactic relation and then applies graph algorithms to reveal the structure of the (sub)graph(s). These (sub) graphs can be used to calculate the conceptual similarity, the conceptual salience in a conceptual network, and to classify the members in a graph sub-graph. The results of such an analysis can be beneficial for various lexicographic and

natural language processing applications and tasks such as semantic similarity measurement, identification of associated semantic domain clusters, or the word sense disambiguation. In terms of research on the emotion linguistic expression, the method can enhance the empirical insight into the lexical structures of the emotional categories used to express different emotional phenomena. Moreover, the empirical interpretations of the results with respect to a specific corpus, i.e. language community, open the possibility for a cross-cultural, synchronic, or diachronic usage-based analysis.

In this case study, we used two large web corpora: the English enTenTen13 and the Croatian hrWac13 corpus. After constructing friend-of-a-friend network by using coordinated collocates with over 1000 lexical nodes related to the source node *feeling* in English and *osjećaj* in Croatian, we performed centrality and community detection graph algorithms that yielded the hierarchical structure of the associated semantic domains.

The structure of the article is the following. In the second and third section, we explain the construction of the graph and subsequent application of the centrality and cluster identification algorithms for the lexeme *feeling* in English enTenTen13 corpus and *osjećaj* in hrWac13, respectively. The fourth section compares the results. The fifth section outlines the conclusions.

2. Syntactic-semantic networks of the source lexeme *feeling*

2.1. The ontological problem of feeling

The semantic analysis must consider the ontological status of the lexicalized phenomena. In other words, we must start with the question of what the word *feeling* refers to? The feeling is a phenomenon from the subjective psychological domain. It is something that emerges from the ability of some living organisms to perceive the environment. From the perspective of cognitive psychology, feeling is a general abstraction of the ability to differentiate the affect states, or the ability to sense a bio-psycho-social response to certain stimuli (Scherer, Schorr and Johnstone 2001; Scherer 2009). The feelings produce a sort of qualitative feature of hedonic valence and arousal from the sensorial inputs that is relevant

for the cognitive processes such as remembering, reasoning, etc., yielding in a structure of individual's tendencies to construct social identities and interactions. The ability to discern between different kinds of feelings is known as emotional granularity. An individual with high emotional granularity can discriminate affect states of an apparently similar level of valence and arousal, labeling them with discrete lexical concepts for emotions (Barrett 2006). A person with low emotional granularity would report their emotions in rather broad and coarse categories. In this cognitive sense, the value of the lexicographic research of the feeling domain is to display the sense structure of different lexemes, and possibly to enrich the emotional communication quality by choosing the most appropriate lexeme.

2.2. Lexicographic data on feeling

The traditional lexicographic resources give us ample result on the senses and associative concepts. For instance, the definition of the concept *feeling* according to the Merriam-Webster dictionary (Merriam Webster Thesaurus) is a subjective response to a person, thing, or situation. It is synonymous with lexical items such as *chord*, *emotion*, *passion*, *sentiment*. It has related words such as *impression*, *perception*, *sensation*, *sense*, *angle*, *attitude*, *outlook*, *perspective*, *standpoint*, *viewpoint*, *belief*, *conviction*, *judgment* (or *judgement*), *mind*, *notion*, *opinion*, *persuasion*, *verdict*, *view*, *receptiveness*, *receptivity*, *responsiveness*, *sensibility*, *sensitiveness*, *sensitivity*. However, lexicographic thesauri rarely give information on the structural differences between lexical features. The results are mostly represented as a list of words that can be taken as synonymous or otherwise semantically related. There is no detailed information on the interrelatedness of these words or the structure of different word senses. The idea behind the corpus-based graph research is to extract lexical concepts that are prototypically related to the lexical concept feeling in the original culturally distributed language usage. It is an empirical, bottom-up approach to the semantic analysis, and a cognitively grounded method of detecting associated concepts, the structure of the related semantic domains and word senses.

2.3. Data harvesting

The extraction of the conceptual networks of the concept feeling in English is performed using the large tokenized, lemmatized, and syntactically tagged enTenTen13 corpus available at the Sketch Engine service. We used a predefined WordSketch pattern (Sketch Engine, a) of noun lemmas connected with *and*/*or* to extract collocations. For each lexeme, we harvested up to the first 100 collocates ranked by the logDice statistic measure using the Sketch Engine API service. The lexical data has been stored locally in the Neo4j graph database. The code for the data harvesting process is available at the Github (Neo4j code).

2.4. English versus Croatian Corpus Queries

The query on English enTenTen13 corpus looks for collocations of the same part of speech class (POS), e.g. nouns, connected with *and*/*or* lexical connectives and words of the same sort separated by a comma, e.g. cat, mouse, and dog, see predefined grammatical constructions for enTenTen13 corpus in Sketch Engine (Sketch Engine, b). The query for the Croatian hrWac22 corpus involves lexemes of the same POS that are connected by *and*/*or* constructions involving either explicit appearances of *i* (Eng. and), *ili* (Eng. or), appearances of some other lexical connectives with the similar meaning in Croatian, as well as implicit appearances of and/or obtained by other forms of syntactic constructions. For example, the query includes the *te* connective which is very similar to *i* (Eng. and). In addition, the procedure also captures syntactical constructions that have a very similar meaning to using just and/or connectives, such as : *as well as* . . . , . . . *either* . . . *or* . . . , *neither* . . . *nor* . . . , that is, *kako* . . . *tako* . . . , *niti* . . . *niti* . . . in Croatian (see predefined grammatical relations for the Croatian hrWac22 corpus in Sketch Engine (Sketch Engine, c):

(1) My feelings and behavior are closely linked. . . (enTenTen13)

(2) This is a personal weblog and does not represent the feelings or opinions. . . (enTenTen13)

(3) ...kako da svoje misli i osjećaje slobodno izražava... (hrWac22) "...how to freely express his thoughts and feelings"

(4) ...promijeniti mu smjer misli ili osjećaje dok gleda sliku. (hrWac22) "...to change the direction of thoughts or feelings while looking at the picture"

(5) ...she has neither feelings nor consciousness (enTenTen13).

Notice that for Croatian, we do not include the comma sign in our search, although a comma is used when listing entities in Croatian in the same manner as in English. Here, for Croatian, we do not capture listings by comma, but we are aware that, when listing more than two entities, some intended and/or-connections are not captured since *and/or* are omitted in places. Such omissions are compensated by the huge size of the corpora. For example, the instances 'feelings, thoughts, and opinions' and 'opinions, thoughts, and feelings' having almost identical meaning, result in links between feelings and opinions even with no explicit *and* collocations between these two lexemes. Moreover, some omissions of and/or are captured by the iteratively computed friend-of-a-friend relation. Namely, as both *and* and *or* are commutative and associative logical connectives, conceptually related entities from such listings are eventually connected through the friend-of-a-friend relation that we use to construct the network of entities.

2.5. Graph construction and analysis

The methodology of graph construction and analysis relies on:

(1) Constructing a first-degree friend network from the collocates of the coordination syntactic-semantic construction. In this study, we constructed the friend network from 100 strongest collocates using the logDice measure.

(2) Constructing a second-degree friend-of-a-friend network with the coordination syntactic-semantic construction. In this study, we created the friend-of-a-friend network with 100 strongest collocates for each friend using the logDice measure.

(3) Identifying the sub-graph communities using a community detection algorithm. In this study, we applied Louvain community detection graph algorithm with granularity parameter 0.1.

(4) Identifying the prominent nodes using a centrality detection algorithm. In this study, we applied the PageRank algorithm to the whole graph.

The emergent network of lexical concepts associated with the source lexeme feeling comprises 1138 noun lexemes. The size of the network was modified

by selecting the first order (up to 100 friends) collocations of the coordination grammatical relation and subsequent second-order (up to 100 friends-of-a-friend) collocations linked with the same grammatical relation. The graph is processed using the Igraph Python package (Csárdi and Nepusz 2006) for statistical processing, centrality, and community detection algorithms. The Gephi application (Gephi) was used for the visualization of the graph Force layout with the collocation score as a measure of the relation weight. The layout of the graph is further enhanced by visualizing the size of the labels according to the value of the centrality algorithm PageRank, while the detected communities have different colors. This graph is represented in Figure 1.

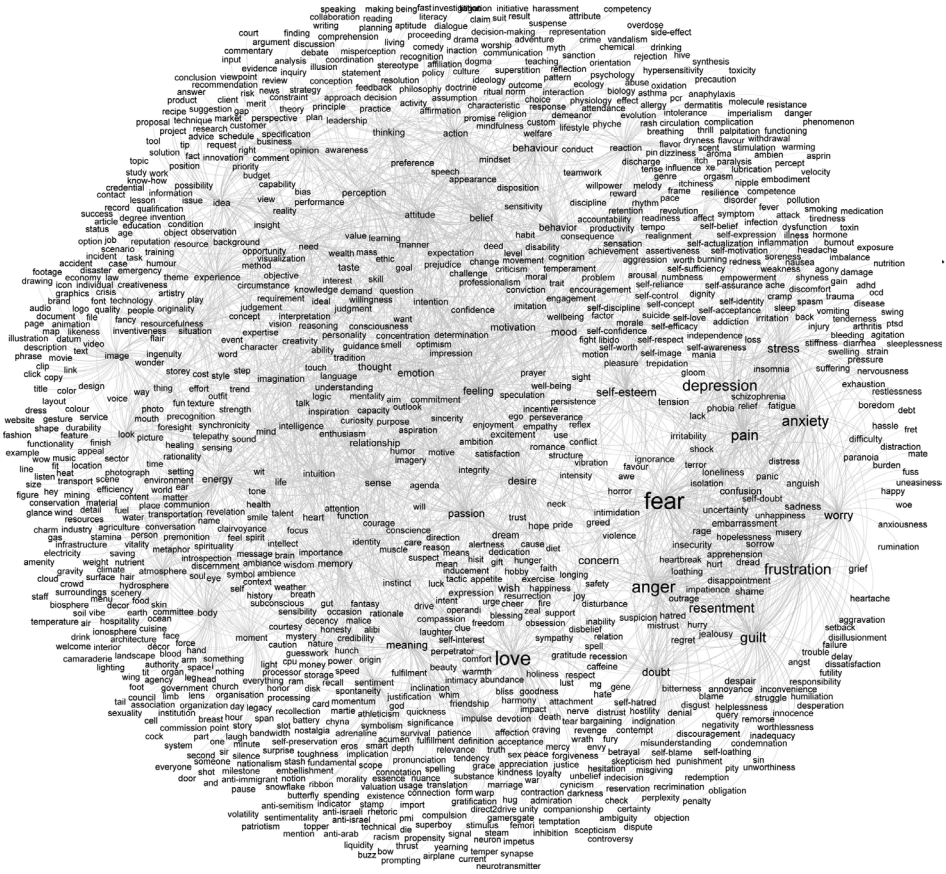

Figure 1: Graph representation of the associated noun lexemes for a noun lemma *feeling* in the enTenTen13 corpus

The network is constructed from the second degree (friend-of-a-friend) collocates in the coordinated "w" and/or syntactic relation with up to 100 strongest collocates according to the logDice statistic measure for identifying collocations. Label size represents the PageRank value, while the color represents the lexical community identified by the Louvain algorithm with the resolution 0.1.

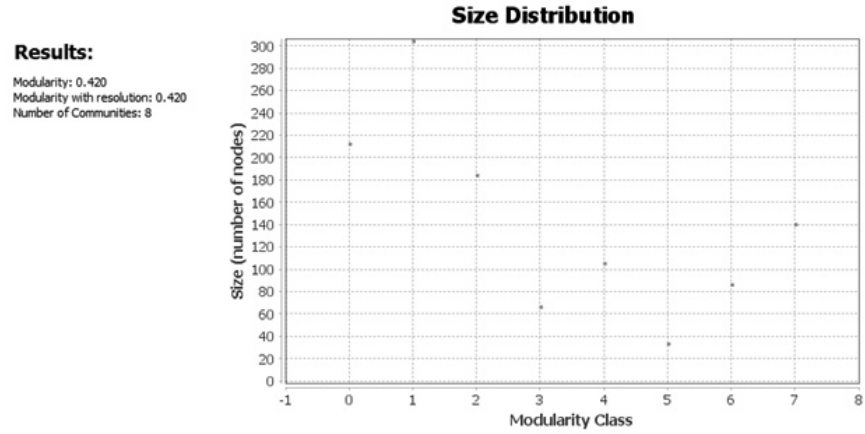

Figure 2: Size distribution of the communities (shown as dots) obtained with the Louvain algorithm

The highly interconnected lexemes form a cluster or a (sub)graph. The syntactic interconnectivity is semantically interpreted as a conceptual association that can be abstracted as a semantic domain. To identify associated semantic domains of the lexeme feeling, we applied the Louvain community detection algorithm (Blondel et al. 2008) on the whole second-order lexeme graph. With the resolution parameter set to 0.1, the algorithm identified 8 communities with the measure of modularity 0.4. The size of the communities varies from 30 to 300 nodes, as represented in Figure 2. By combining centrality and community detection algorithms we can discern the associative rank of the semantically related concepts as well as their respective interconnected communities. This method thus allows an empirical ranking of the semantically salient concepts and semantic domains abstracted from their interrelation. In this work, we did not perform an automatic labeling or external knowledge aligning of the semantic domains, but we presented the domains by their main representatives, identified with the centrality algorithm. In the following section, we show the abstracted semantic

domains with the Louvain algorithm and ranked according to the PageRank value of the distinctive nodes in the feeling conceptual network.

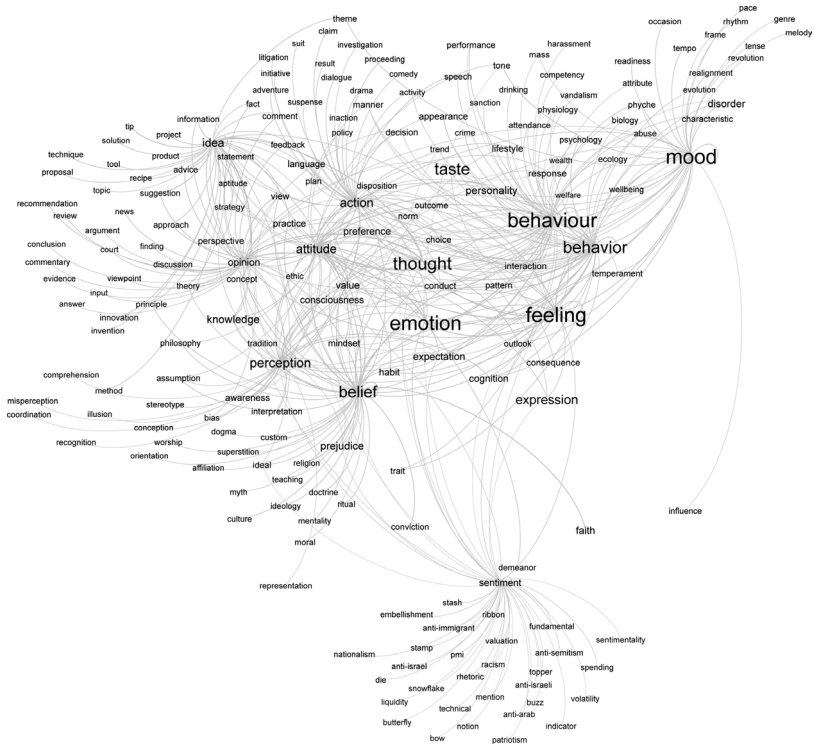

Figure 3: Feeling: Community *feeling*

Table 1: Nodes in the *feeling* community ranked according to the PageRank distribution

Lemma	PageRank
feeling	701
emotion	648
thought	499
belief	356
perception	342
attitude	304
mood	301
action	281
behavior	247
behaviour	236
idea	214
knowledge	183
personality	181
opinion	180
expectation	159
preference	154

Figure 3 shows the lexical community with the prominent source concept *feeling*. The centrality algorithm identified concepts *perception*, *emotion*, *mood*, *thought*, *behavior* as the lexical members with salient association, as well as the ontological similarity with the source concept *feeling*. These salient members of the sub-graph indicate the classification of the feeling as a member of the high-level category of affective, cognitive, and behavior concepts, in the sense of the Eleanor Rosch theory of prototype categorization hierarchy (Rosch, Mervis, Gray, Johnson and Boyes-Braem 1976). The high-level concepts lack informativeness because they represent the abstractions from basic-level categories. However, they provide the cognitive access to the highly elaborate structure of more basic-level and low-level categories.

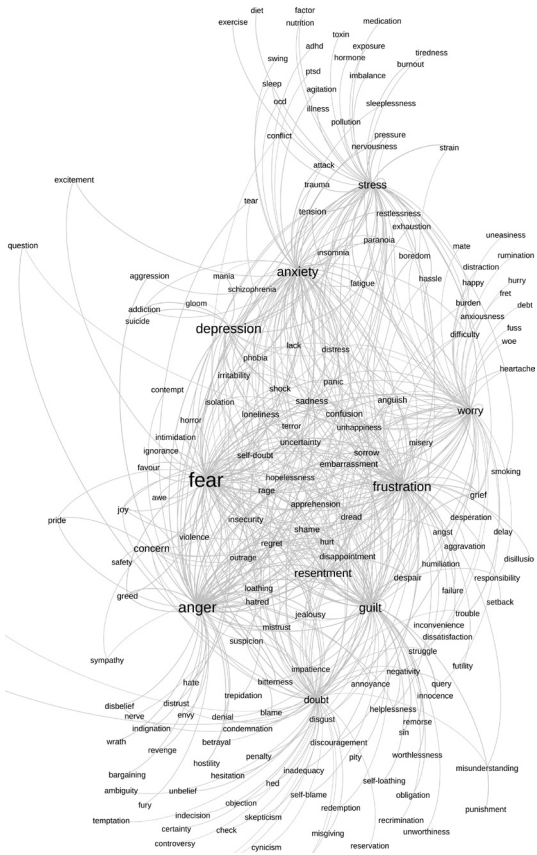

Figure 4: Feeling: Community *fear*

Table 2: Nodes in the *fear* community ranked according to the PageRank distibution

Lemma	PageRank
fear	526
anger	470
anxiety	457
depression	454
frustration	431
guilt	405
stress	388
worry	310
concern	309
confusion	295
sadness	280
grief	280
resentment	256
anguish	256
tension	250
loneliness	248
joy	244
shame	241
despair	241

The next sub-graph in Figure 4 is profiled with the salient concepts *fear*, *anger*, *frustration*, *worry*, *depression*, *anxiety*, *stress*. This lexical community represents emotional lexemes with negative hedonic valence and heightened arousal, according to the emotional dimensions of the core affect theory (Russell and Barrett-Feldman 1999). Their PageRank values indicate the importance of these more basic-level categories in the construction of the semantic matrix of *feeling*.

Another prominent community with negative hedonic valence and high arousal is related to the lexeme *pain* shown in Figure 5. The lexemes in this community refer to the deteriorating bio-psychological states, symptoms, or mechanisms of reaction such as reaction, inflammation. It is interesting to note that pleasure is classified in this community, although it obviously ontologically does not fit in the classification. This result probably emerges from the mixture of low discriminative clustering parameters and frequent formulaic use of antonyms in the *and-or* construction. In this representation, the pleasurable hedonic state of

pleasure is profiled as the opposition to the pain. The formulaic use of antonyms in the *and/or* constructions often refer to the high-level category by profiling extreme items of phenomena in the category, rather than just profiling the relation between antonyms. This antonym construction, in conjunction with other syntactic constructions, can provide valuable syntactic-semantic procedures for extracting the antonyms from a corpus.

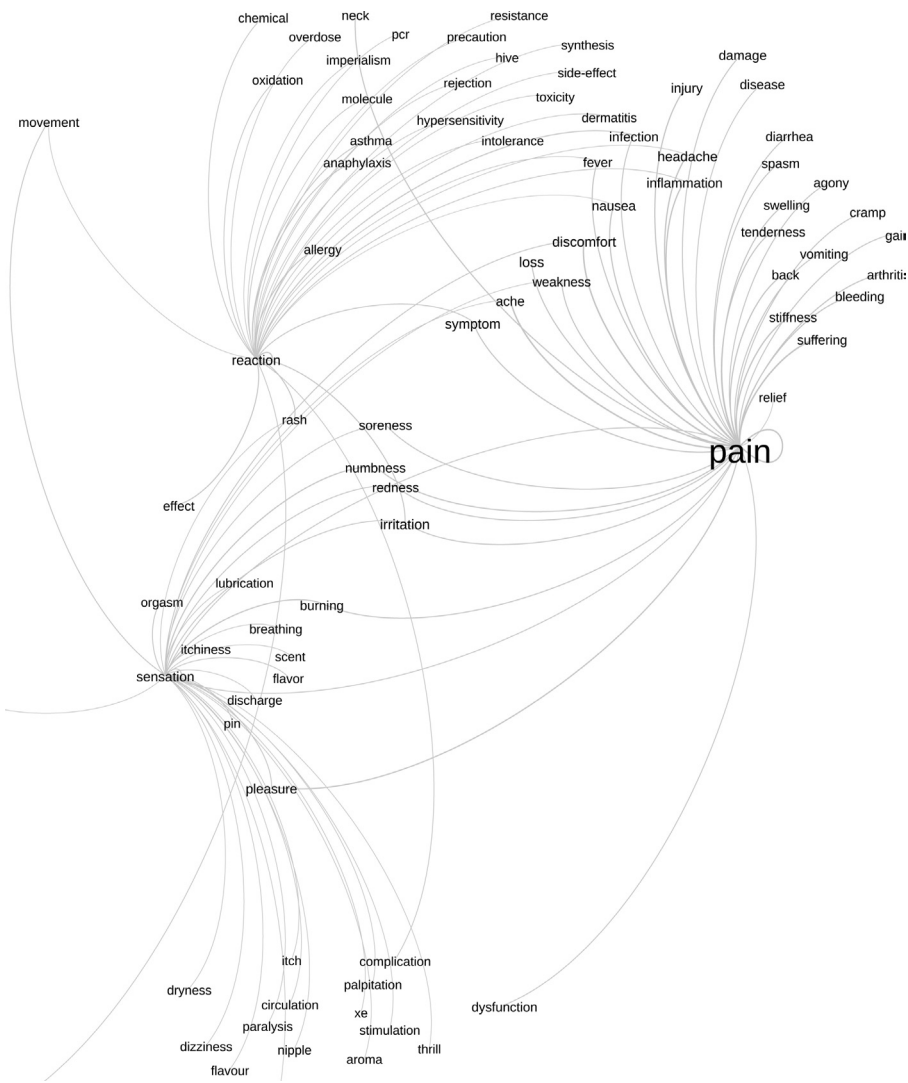

Figure 5: Feeling: Community *pain*

Table 3: Nodes in the *pain* community ranked according to the PageRank distribution

Lemma	PageRank
pain	398
reaction	186
irritation	129
loss	119
headache	104
symptom	100
discomfort	100
nausea	80
pleasure	78
inflammation	66
infection	66
injury	54
damage	54
disease	54
movement	46
fever	42
burning	38
numbness	38

The next community, shown in Figure 6, indicates the relation of the feeling to affective features of desire and motivation. It is represented with lexemes *desire*, *passion*, *motivation*, *sense* that have a common feature of a psychological state of wanting to achieve something with relatively high arousal and positive hedonic valence.

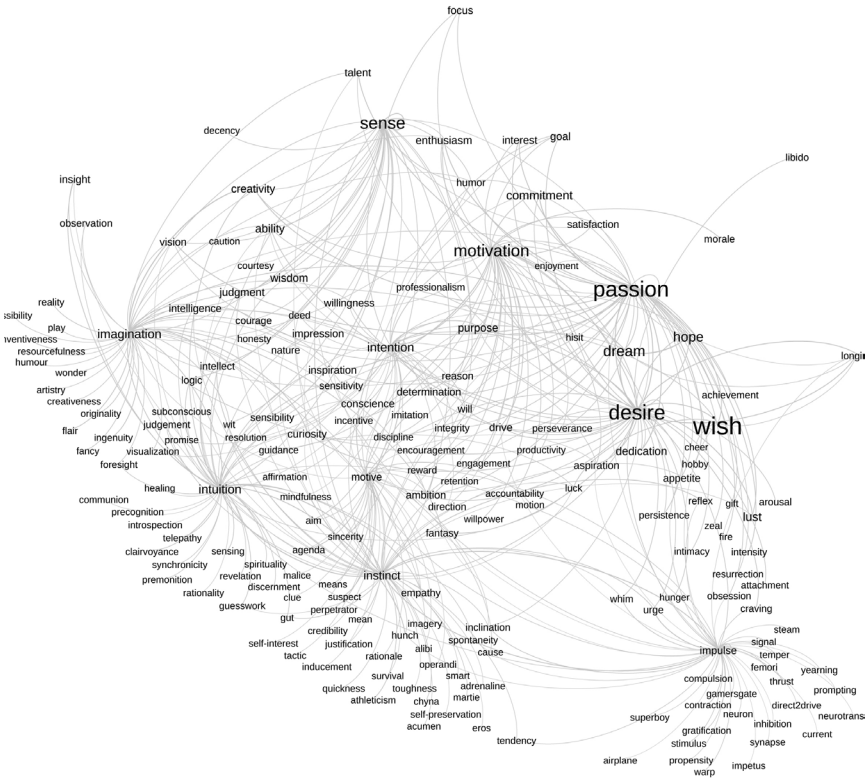

Figure 6: Feeling: Community *desire*

Table 4: Nodes in the *desire* sub-graph ranked according to the PageRank distribution

Lemma	PageRank
desire	326
passion	284
motivation	251
sense	231
intention	225
dream	215
imagination	199
ability	185
commitment	175
hope	163
intuition	157
impression	142
creativity	125
enthusiasm	125

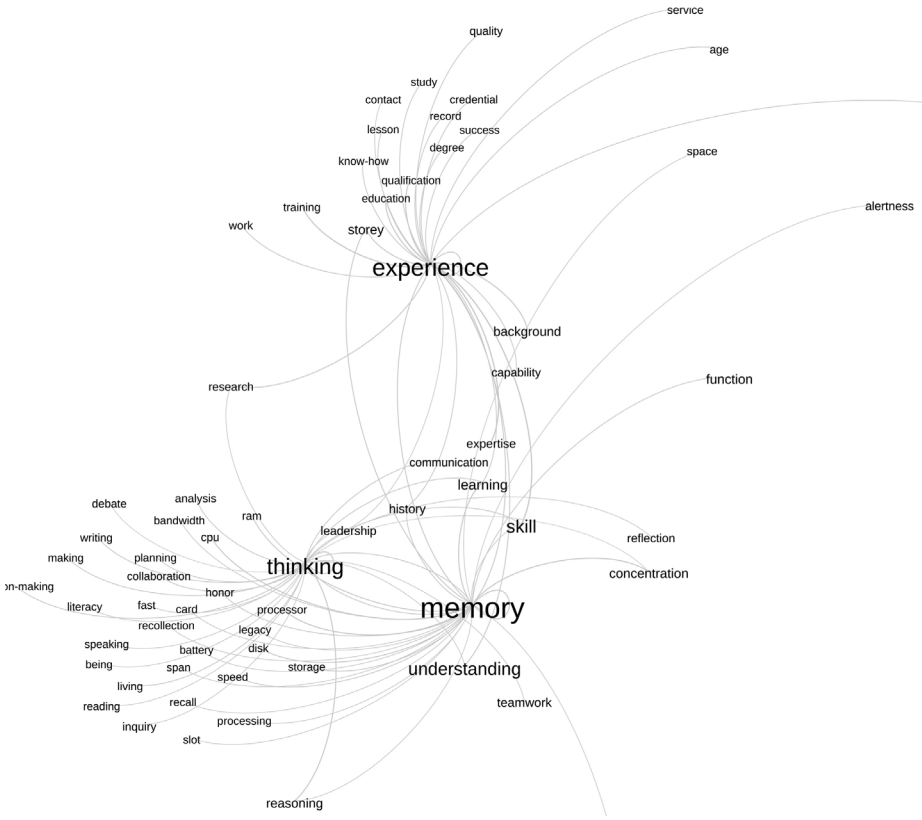

Figure 7: Feeling: Community *thinking*

Table 5: Nodes in the *thinking* sub-graph ranked according to the PageRank distribution

Lemma	PageRank
thinking	237
experience	222
skill	215
memory	211
understanding	146
learning	99
reasoning	72
concentration	71
storey	69
level	69

The community in Figure 7 presents the facet of the feeling domain as meronomically related to the prominent cognitive abilities like *thinking*, *memory*, *understanding*, *learning.* This corroborates the appraisal theories of emotions (Scherer, Schorr and Johnstone 2001) that claim that affective phenomena necessitate the cognitive component of the appraisal. The community with the prominent concept *self-esteem* indicates the social role of the feeling in self-awareness and identity, such as *confidence*, *shyness*, *self-hatred*, *well-being*, *self-discipline*, *self-confidence*, etc. see Figure 8.

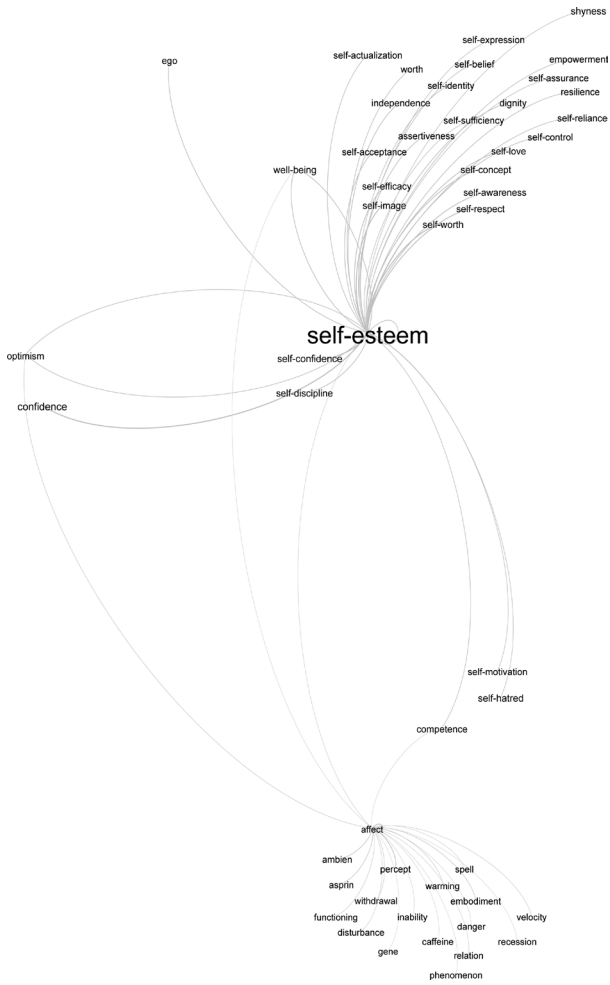

Figure 8: Feeling: Community *self-esteem*

Table 6: Nodes in the *self-esteem* sub-graph ranked according to the PageRank distribution

Lemma	PageRank
self-esteem	234
confidence	119
shyness	60
self-hatred	56
well-being	48
self-discipline	38
self-confidence	38
ego	33
optimism	24
competence	24
empowerment	22
resilience	22
self-efficacy	22
worth	22
self-worth	22
self-awareness	22
self-image	22
self-concept	22
self-respect	22

The community *love* in Figure 9 exemplifies the conceptualization of the domain *feeling* related to the romantic disposition and behavior with high arousal and preference towards positive hedonic valence. It is interesting to note that *energy*, *strength*, and *power*, as well as *attention*, *compassion*, and *romance* are the prominent members of the community.

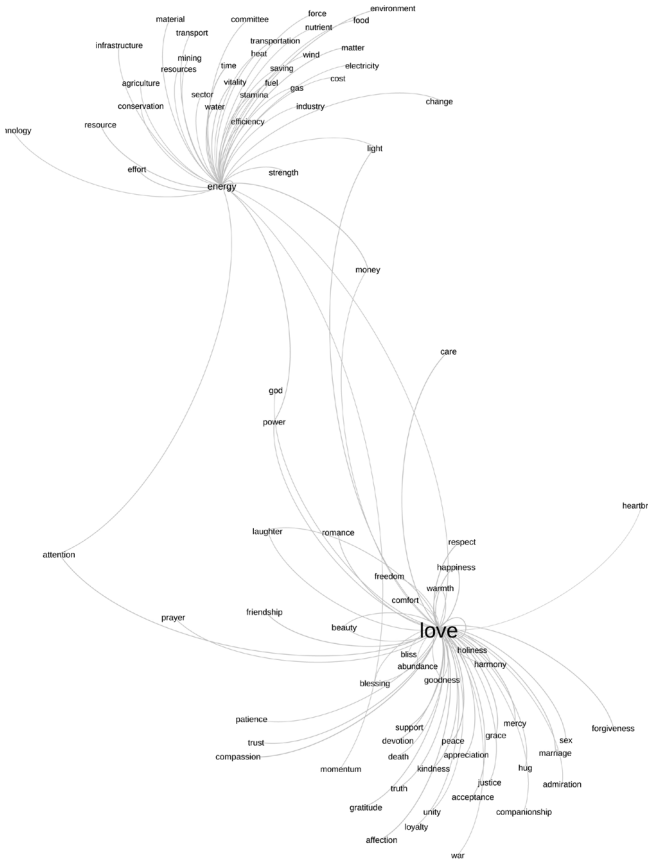

Figure 9: Feeling: Community *love*

Table 7: Nodes in the *love* sub-graph ranked according to the PageRank distribution

Lemma	PageRank
energy	230
love	206
attention	112
power	80
compassion	77
romance	66
change	64
happiness	61
prayer	61
forgiveness	61
resource	52
strength	52
affection	51
respect	49
beauty	48
light	48
trust	48
devotion	47

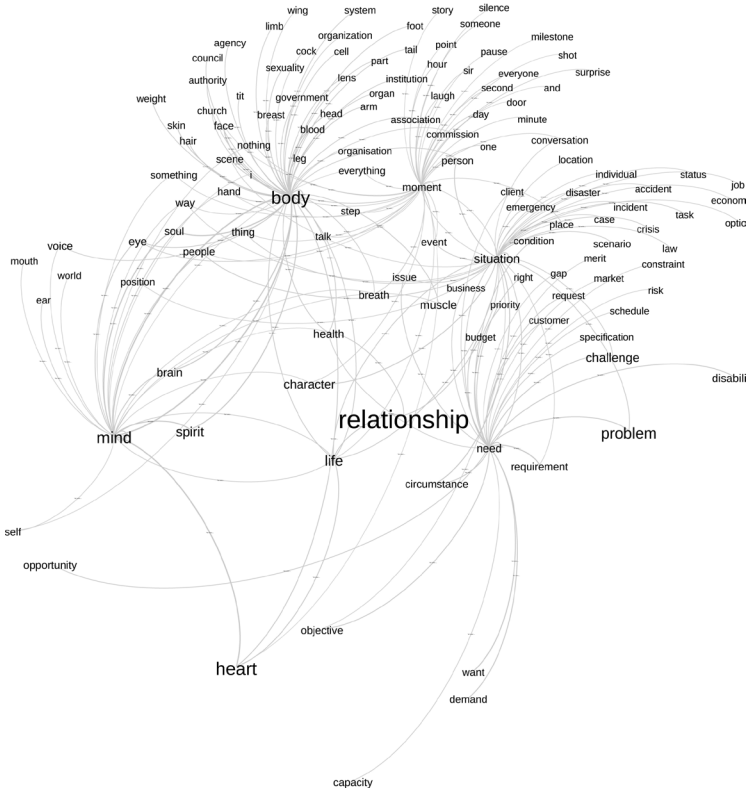

Figure 10: Feeling: Community *relationship*

Table 8: Nodes in the *relationship* sub-graph ranked according to the PageRank distribution

Lemma	PageRank
heart	182
mind	174
relationship	165
problem	150
spirit	126
character	104
body	91
life	88
challenge	87
brain	71
need	69
moment	65

The network structure of the *love* community in Figure 9 could have been further categorized into two sub-graphs with *love* and *energy* as the central nodes. In terms of the community detection algorithm this can be achieved with lowering the clustering resolution parameter. This maps to the general idea of emotion granularity: if we magnify resolution, we can obtain smaller but more distinct abstractions of semantic domains, that are latently inherent in the more emergent semantic domains. The community in Figure 10 profiles the feeling domain in terms of the relationship and causally prominent embodied components of such interpersonal relation, like *heart*, *body*, *brain*, their psychological components, *mind*, *spirit*, as well as the co-occurring identity and interactional states, such as *character*, *situation*, *challenge*, *problem*. The community in Figure 11 presents the meronomical relation of the feeling concept to the network of concepts that form a domain related to the process of cognitive representation of objects that contain meaning.

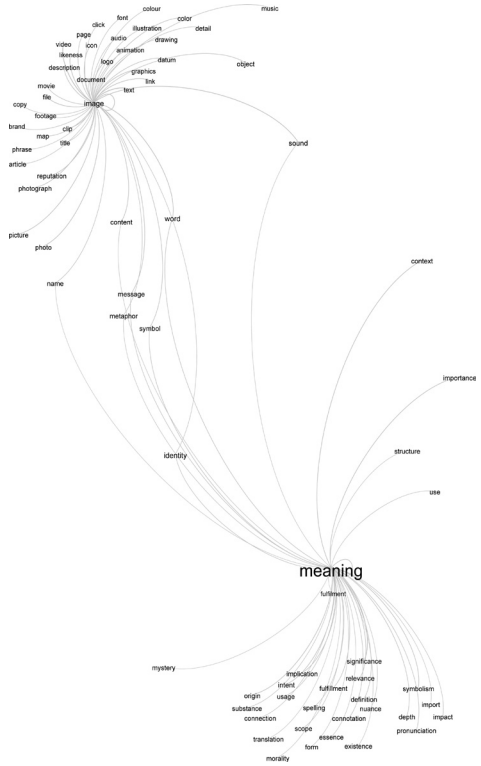

Figure 11: Feeling: Community *meaning*

Table 9: Nodes in the *meaning* sub-graph ranked according to the PageRank distribution

Lemma	PageRank
image	132
word	113
meaning	110
identity	83
sound	81
reputation	49
intent	38
object	32
use	30
structure	30
content	25

The last community with the smallest centrality measure, shown in Figure 12, seems to refer to the ambient conditions that are conducive for the emergence of the feeling. Some of the prominent categories, such as *style*, *design*, *outfit*, *texture* indicate the interaction practices and human manipulation of the ambient design and would be related to inducing positive hedonic valence.

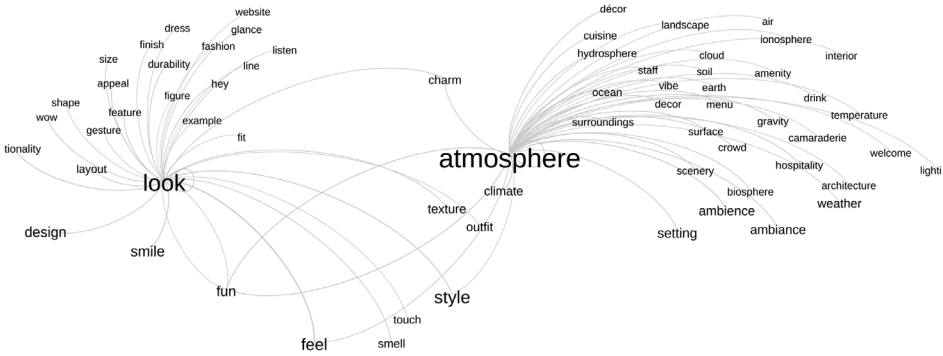

Figure 12: Feeling: Community *atmosphere*

Table 10: Nodes in the *atmosphere* sub-graph ranked according to the PageRank distribution

Lemma	PageRank
style	76
look	75
atmosphere	72
feel	52
smile	48
design	44
setting	33
fun	30
texture	29
outfit	29
ambience	28
ambiance	28
weather	28

The eight communities show a rather coherent structure that represents the domain matrix of the lexical concept feeling in the large enTenTen13 corpus. The PageRank centrality measure gives a sense of the inter-domain relatedness and prominence. In the next section, we will present the results of the corresponding concept in the Croatian corpus – *osjećaj.*

3. Constructing the graph of the lexeme *osjećaj* (Eng. feeling) in hrWac22

According to the *Croatian language portal* (HJP) (Hrvatski jezični portal) *osjećaj* (Eng. feeling) is a psychological form of experiencing that expresses a certain emotional relation (feeling of loneliness, reaction, guilt, feeling of low social value, etc.); a capability to feel referring to the senses; a special emotion reaction; a special inclination for something (a feeling for literature); thinking, comprehension, especially irrational. The graph analysis of the Croatian lexical concept *osjećaj* (Eng. feeling) is performed on the data extracted from the hrWac22 corpus. The same ConGraCNet method for constructing a coordinated collocation based conceptual network for the lexical concept *osjećaj* (Eng. feel-

ing) is performed on the Croatian hrWac22 corpus yielding a somewhat different configuration of the sub-graphs and represented semantic domains.

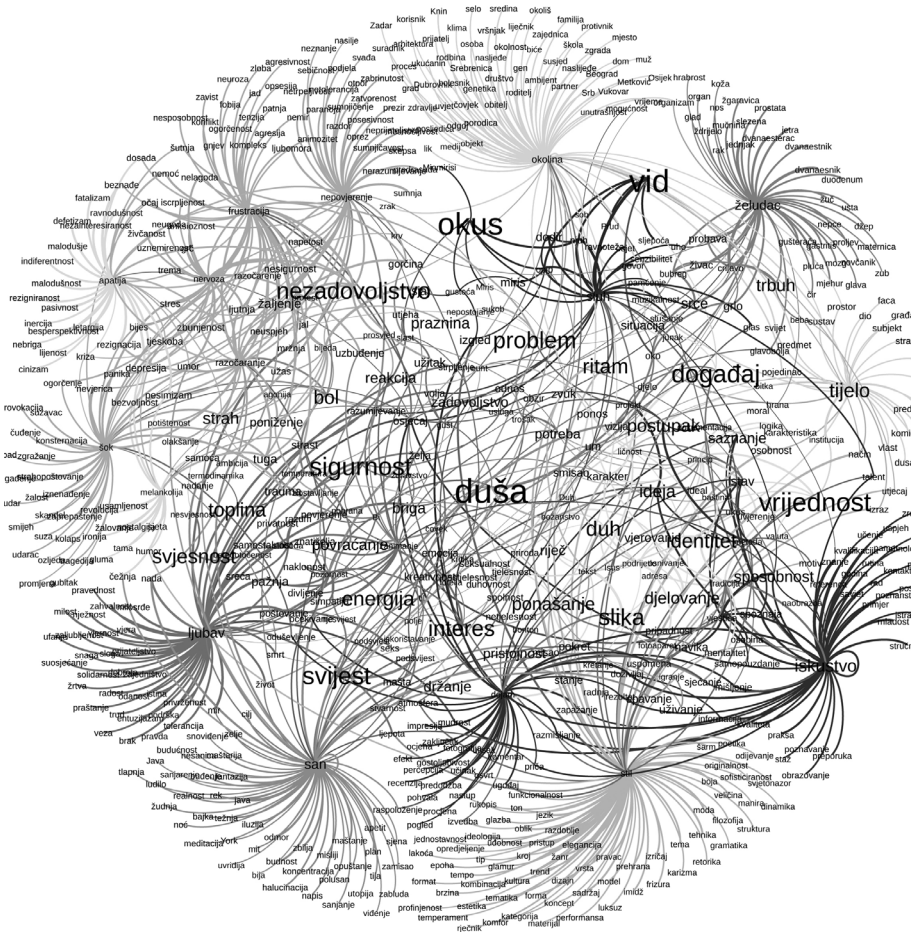

Figure 13: Graph representations of coordination collocations: *osjećaj* (Eng. feeling). Network of the coordinated construction reflects the associative structure of the concepts and semantic domains related to the concept *osjećaj* (Eng. feeling).

Figure 14: Size distribution of the communities obtained with the Louvain algorithm for the hrWac22

The resulting conceptual network of the source lexeme *osjećaj* is represented in Figure 13 with 646 lexical concepts. Using the Louvain community detection algorithm with the resolution parameter set to 1, collocation logDice score as the weight parameter, the algorithm identified 10 communities with the measure of modularity 0.66, as shown in Figure 14. The most prominent community, according to the PageRank values, is the community that refers to the conceptualized parts of the cognizer attributed with sensing of the feeling, see Figure 15. These are *soul* as the carrier of the *consciousness* and *mind* in the dualistic cultural model. There are some body parts: *stomach*, *nerve*, *heart* that point to the visceral model of the sensing.

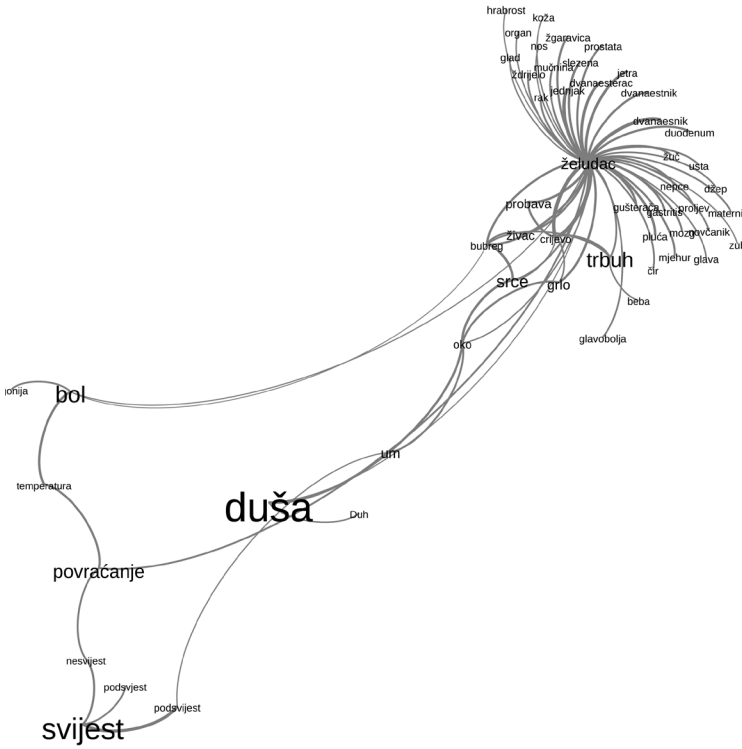

Figure 15: *Osjećaj* (Eng. feeling): Community *duša*

Table 11: Nodes in the *duša* community ranked according to the PageRank distribution

Lema	Eng	PageRank
duša	soul	49
svijest	concio-usness	36
bol	pain	29
trbuh	stomach	26
povraćanje	vomit	24
srce	heart	22
želudac	stomach	21
um	mind	18
probava	digestion	17
živac	nerve	17
oko	eye	15

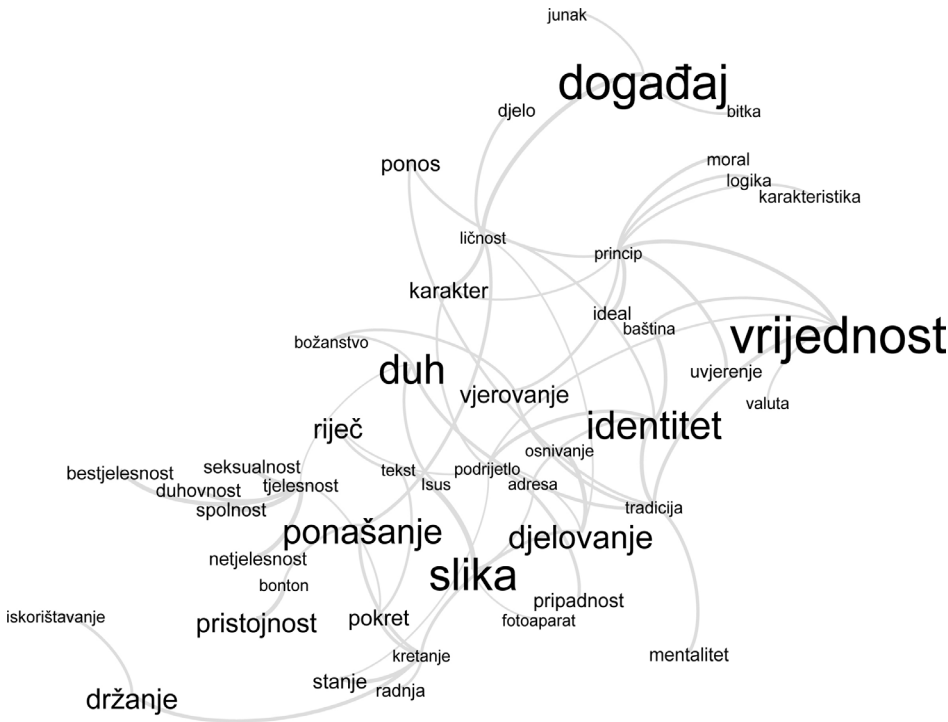

Figure 16: *Osjećaj* (Eng. feeling): Community *vrijednost*

Table 12: Nodes in the *vrijednost* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
vrijednost	value	39
događaj	event	37
slika	image	34
duh	spirit	31
identitet	identity	30
ponašanje	behaviour	37
djelovanje	act	25
držanje	keeping	22
riječ	word	21
pristojnost	decency	21

The next community *value*, shown in Figure 16, seems to conceptualize the value that is related with the feeling, mostly in the sense of the social identity:

pride, *image*, *spirit*, *identity* and the value of feeling emerging from social interaction: *event*, *act*, *word*, *decency*, *movement*.

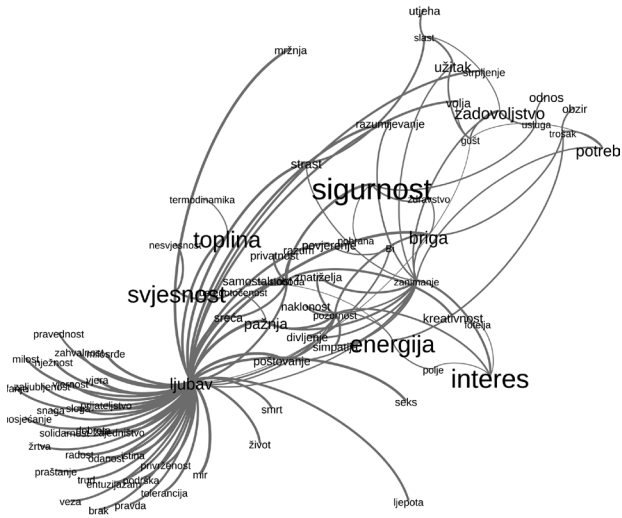

Figure 17: *Osjećaj* (Eng. feeling): Community *sigurnost*

Table 13: Nodes in the *sigurnost* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
sigurnost	safety	37
interes	interest	33
energija	energy	30
svjesnost	consciousness	30
toplina	warmth	29
briga	care	23
zadovoljstvo	pleasure	22
ljubav	love	21
potreba	need	21
pažnja	attention	20
užitak	pleasure	19
strast	passion	17
odnos	relationship	17
kreativnost	creativity	17
volja	will	17
povjerenje	trust	16

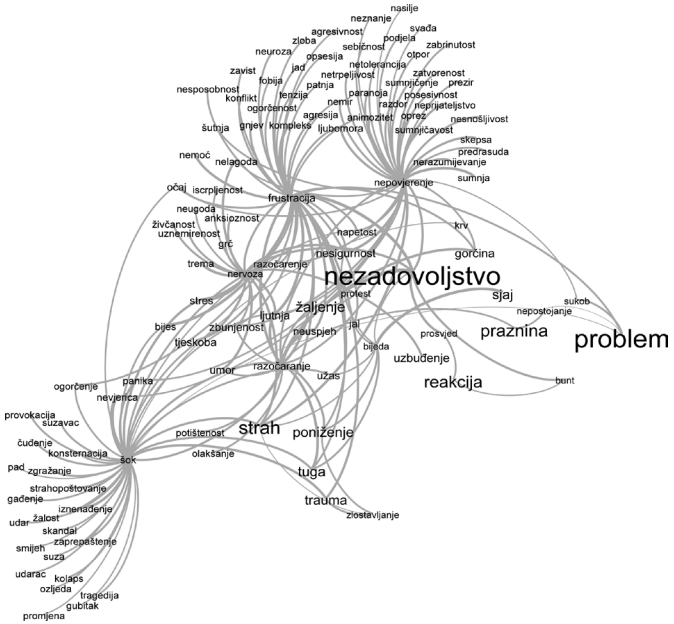

Figure 18: *Osjećaj* (Eng. feeling): Community *nezadovoljstvo*

Table 14: Nodes in the *nezadovoljstvo* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
nezadovoljstvo	dissatis-faction	35
problem	issue	34
strah	fear	25
praznina	emptiness	24
reakcija	reaction	23
žaljenje	grief	20
poniženje	humiliation	20
tuga	sadness	19
sjaj	glow	19
trauma	trauma	18
gorčina	bitterness	17
uzbuđenje	excitement	17
nesigurnost	insecurity	16
ljutnja	anger	16

Figure 17 illustrates the community that refers to the domain of positive affective states and processes and a heightened level of arousal: *safety*, *warmth*, *care*, *pleasure*, *need*, *creativity*, *will*, *trust*, *respect*, *independence*, *happiness*. Somewhat like the enTenTen13 corpus, concepts like *love* and *passion* are conjoined with *interest*, *energy*, *attention*, shown in Figure 9.

The affective states with negative hedonic valence are classified in the sub-graph shown in Figure 18. These are concepts with a mixed level of arousal. High arousal is related to states such as *fear*, *bitterness*, *excitement*, *anger*, *confusion*, *anxiety*, *stress*, and *grief* while low arousal features share *humiliation*, *sadness*, *trauma*, *fatigue*. It seems that this sub-graph could be further granulated to show the distinction between these two groups of states. This sub-graph is commensurable with the *fear*, *anger* community in enTenTen13, Figure 4. The next community profiles the concepts that form a domain related to the objects that promote certain feelings with social implications: *rhythm*, *expression*, *performance*, *stage act*, *influence*, *elegance*, *genre*, *personality*, *style*, see Figure 19.

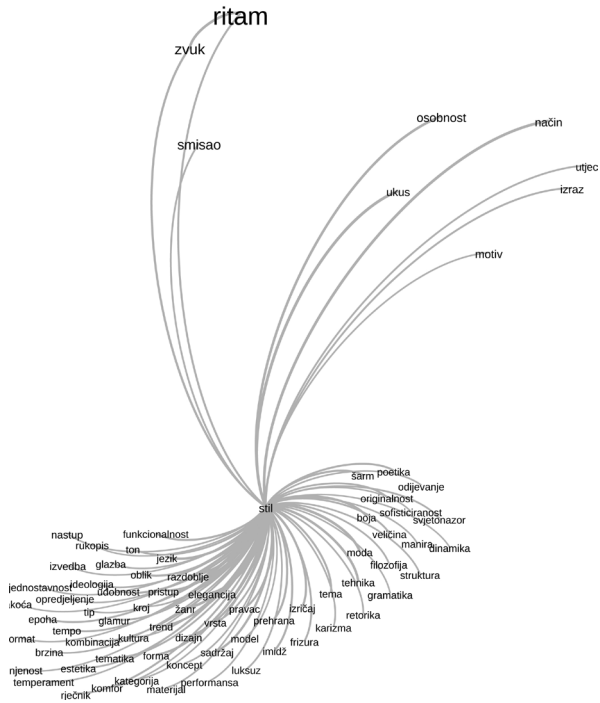

Figure 19: *Osjećaj* (Eng. feeling): Community *ritam, zvuk, smisao*

Table 15: Nodes in the *ritam, zvuk, smisao* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
ritam	rhythm	31
zvuk	sound	19
smisao	sense	18
osobnost	personality	16
stil	style	15
motiv	motive	15
način	means	15
izraz	expression	15
ukus	taste	15
izvedba	performance	14
nastup	stage act	14
utjecaj	influence	14
elegancija	elegance	14
žanr	genre	14
moda	mode	14

The community illustrated in Figure 20 is similar to the English enTenTen13 desire sub-graph Figure 6, referring to the domain of psychological motivation and expectation.

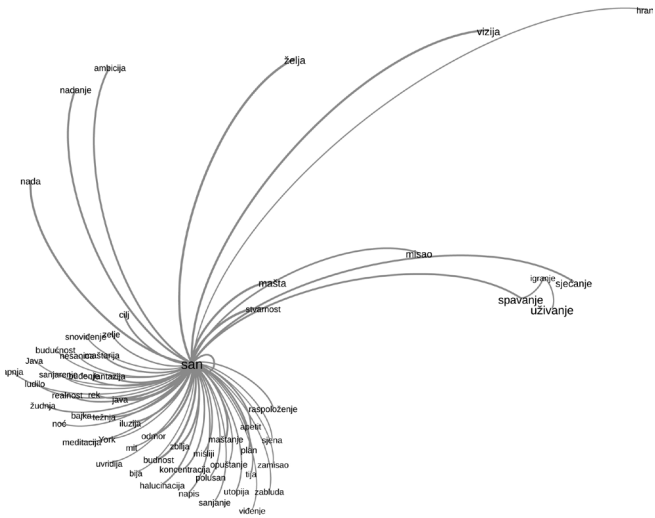

Figure 20: *Osjećaj* (Eng. feeling): Community *san, uživanje, spavanje, želja*

Table 16: Nodes in the *san, uživanje, spavanje, želja* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
san	dream	21
uživanje	enjoyment	19
spavanje	sleep	18
vizija	vision	17
želja	desire	17
mašta	fantasy	16
sjećanje	remem-berance	16
misao	thought	16
nada	hope	15
nadanje	aspiration	15
ambicija	ambition	14
stvarnost	reality	14
raspoloženje	mood	14

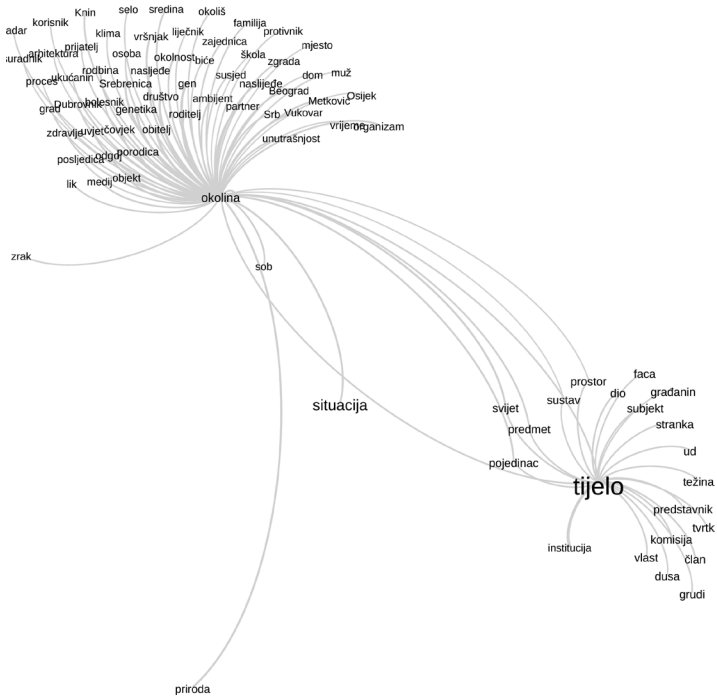

Figure 21: *Osjećaj* (Eng. feeling): Community *tijelo, situacija*

Table 17: Nodes in the *tijelo, okolina, situacija* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
tijelo	body	30
situacija	situation	20
okolina	environment	16
svijet	world	15
pojedinac	individual	15
predmet	item	15
prostor	space	15
sustav	system	15
komisija	commission	15
član	member	15
dio	part	15
vlast	government	15
stranka	party	15

The sub-graph composed of the lexemes *body*, *situation*, *environment*, *world* is shown in Figure 21. It is comparable to the ambient community in enTenTen13, Figure 12, but with additional profiling of the institutional identities of government, party that form the group dynamics of the feeling. The sub-graph with the lexical concept *osjećaj* (Eng. feeling) in Figure 22 is comparable to the enTenTen13 community related to the abstract psychological concepts, Figure 3. hrWac22 connects the affective and cognitive concepts of *knowledge*, *attitude*, *habit*, *feeling*, *cognition*, *emotion*, *memory*, *thinking*. Slightly differently, though, hrWac22 highlights the *procedure*, *experience*, and *feeling* as conceptually similar categories.

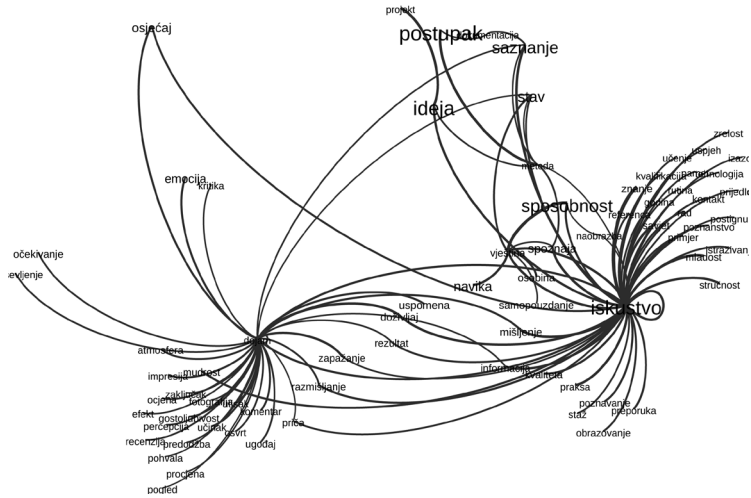

Figure 22: *Osjećaj* (Eng. feeling): Community *ideja, postupak, osjećaj*

Table 18: Nodes in the *ideja, postupak, osjećaj* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
postupak	procedure	27
ideja	idea	26
iskustvo	experience	26
sposobnost	capability	24
saznanje	knowledge	22
stav	attitude	20
navika	habit	18
osjećaj	feeling	18
spoznaja	cognition	17
emocija	emotion	17
uspomena	memory	15
mišljenje	thinking	15
osobina	feature	15
znanje	knowledge	15
očekivanje	expectation	15

The depression community, Figure 23, profiles a domain with hedonically negative affective states with a saliently low level of arousal such as *pessimism*, *solitude*, *apathy*, *resignation*, *melancholy.* This is comparable to the *depression,*

fear community in enTenTen13, Figure 4 Slightly lower on the PageRank scale is the community that refers to the senses: *sight*, *taste*, *hearing, smell*, *touch*, shown in Figure 24. The PageRank score indicates that senses do not have such a strong prominence in the coordination construction of the lexeme *osjećaj* (Eng. feeling) in hrWac22.

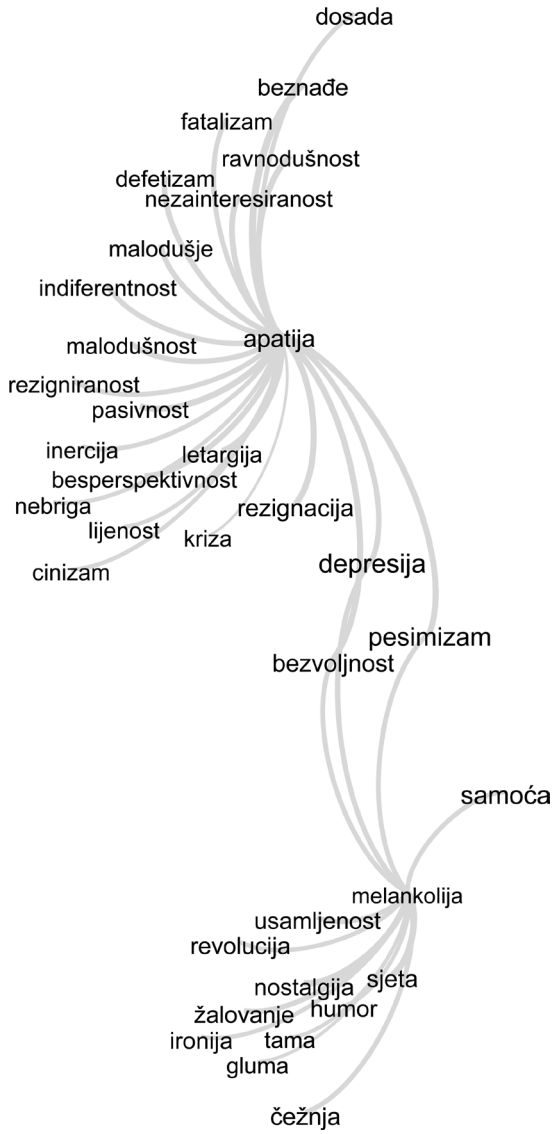

Figure 23: *Osjećaj* (Eng. feeling): Community *depresija*

Table 19: Nodes in the *depresija* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
depresija	depression	16
pesimizam	pessimism	16
samoća	solitude	16
bezvoljnost	apathy	15
rezignacija	resignation	15
apatija	apathy	15
čežnja	longing	15
sjeta	melancholy	15
nostalgija	nostalgia	15
beznađe	hopelessness	15
žalovanje	grief	15
usamljenost	loneliness	15
humor	humor	14
revolucija	revolution	14
ironija	irony	14
dosada	boredom	14
tama	darkness	14
kriza	crisis	14

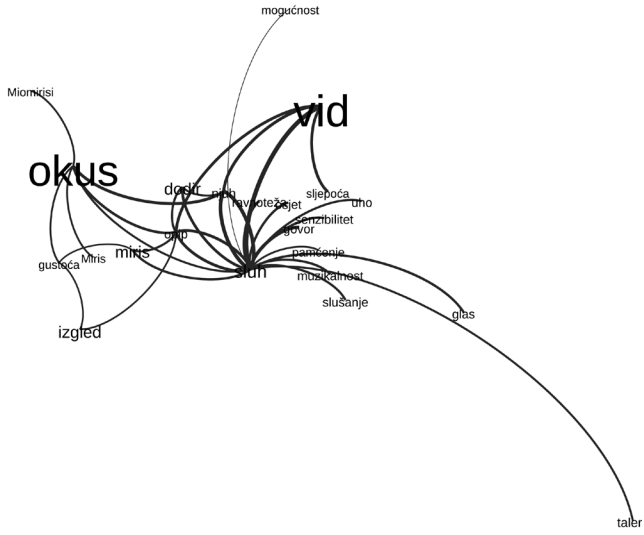

Figure 24: *Osjećaj* (Eng. feeling): Community *okus, vid, sluh*

Table 20: Nodes in the *okus, vid, sluh* sub-graph ranked according to the PageRank distribution

Lemma	Eng	PageRank
vid	sight	46
okus	taste	45
sluh	hearing	20
miris	smell	19
dodir	touch	19
izgled	look	18
talent	talent	15
njuh	smelling	15
uho	ear	15
opip	touch	15
glas	voice	15
ravnoteža	balance	15

4. Comparative analysis of lexical items *feeling* and *osjećaj*

Our study includes a comparison of the two corpora from English and Croatian because we wanted to explore the potential of the ConGraCnet methodology to analyze syntactic-semantic networks of different languages. We contrasted English with Croatian, as languages that have syntactically different structures. However, due to the universality of the concept *feeling*, the universality of the coordinated syntactic relation (Van Oirsouw 2019), and the comparable tagging scheme we have performed the analysis without major changes in the methodological procedure.

The two networks of *feeling* and *osjećaj* are constructed from different sets of texts and a rather diverse grammatical morpho-syntactic systems. Despite these major differences, the resulting lexical graphs display commensurable conceptual and semantic domain distribution represented by the sub-graphs. In this section, we outline the differences in the graph construction and compare conceptually the results of the two graphs.

By abstracting the features of the identified semantic domains in terms of the cognitive theory of emotion, we could conclude that both graphs show the promi-

nent conceptual relatedness of the *feeling* and *osjećaj* lexemes with the concepts pertaining to the abstract cognitive abilities (Figure 3, Figure 7, and Figure 22) involved in the processing of the feeling. We identified commensurable semantic domains representing the affect categories with negative hedonic valence and high arousal (Figure 4 and Figure 18). The community algorithm extracted from the hrWac22 graph additionally a low arousal component (Figure 23). Both graphs have semantic domains referring to the intense physical reaction to negative stimuli (Figure 5 and Figure 15). Semantic domains with positive valence show similar relation of love and energy (Figure 9 and Figure 20), as well as the motivational dimensions (Figure 6 and Figure 17). The perceptual features of the affect process are identified more distinctly in the hrWac22 sub-graph (Figure 24). The dimensions of social identification and interaction components are represented in the somewhat different and less subjectively recognizable similarity between identified sub-graphs (Figure 10 and Figure 21). Moreover, it seems that enTenTen13 highlights self-esteem as a prominent social feature of the emotion (Figure 6). The feeling as a component of the meaning construction aspect is represented in relation to the sub-graphs (Figure 11 and Figure 16). Lastly, the grounding of the feeling in some environment is referred in the enTenTen sub-graph (Figure 12).

5. Conclusion

This paper demonstrates the ConGraCNet procedure for identifying associative concepts and semantic domains using a corpus-based graph analysis of the syntactic constructions on the example of the concept *feeling*. The study deals with examples in English and Croatian corpora/languages. Due to the cognitive-linguistic universality of the coordination syntactic-semantic construction, the procedure is suitable for analyzing lexical networks and even network structure comparison in different languages. The major prerequisite, however, is a morpho-syntactically tagged corpus. In our web application of ConGraCNet http://emocnet.uniri.hr/congracnet/ we have successfully included corpora from other European languages.

The qualitative comparative analysis of the lexical networks from enTenTen13 and hrWac22 corpora revealed commensurable associative semantic domains associated with the source lexemes. There is expected variation due to the size

and structure of the corpus, different NLP tools used to parse the texts, as well as the structural linguistic differences and cultural patterns. In our future work, we will incorporate some graph comparison methods for a quantitative assessment of the cross-/corpus/language commensurability.

This bottom-up approach can be incorporated into a range of lexicographic applications to enrich the data with structural information about conceptual relation, distance, or salience within a certain semantic domain in a particular corpus, i.e. language community. For instance, the use of the centrality and community detection algorithms for calculation of the semantic domain association rank can be used to enhance the dynamic word sense ordering in lexicographic applications for semantically rich and polysemous words. In our future work, we will also describe the impact of other centrality and community detection algorithms as well as the procedures for fine-tuning of network construction and algorithm parameters.

References

Blondel, Vincent D.; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10. 2–12. doi.org/10.1088%2F1742-5468-%2F2008%2F10%2Fp10008.

ConGraCNet. 2020. http://emocnet.uniri.hr/congracnet (accessed 2 July 2020).

Csárdi, Gábor; Nepusz, Tamás. 2006. The igraph software package for complex network research. *InterJournal, complex systems* 1695/5. 1–9.

Dorow, Beate; Widdows, Dominic. 2003. Discovering corpus-specific word senses. *10th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. Budapest. 79–82.

Feeling. 2020. *Merriam Webster Thesaurus*. Merriam–Webster, Incorporated. Springfield. https://www.merriam-webster.com/thesaurus/feeling (accessed 2 July 2020).

Feldman Barrett, Lisa. 2006. Valence is a basic building block of emotional life. *Journal of Research in Personality* 40/1. 35–55. doi.org/10.1016/j.jrp.2005.08.006.

Gephi. https://gephi.org (accessed 2 July 2020).

Hrvatski jezični portal. http://hjp.znanje.hr (accessed 2 July 2020).

Perak, Benedikt. 2019. *Neo4j code*. https://github.com/bperak/ConGraCNet/blob/master/sketch2Neo4j_ enTenTen13.py (accessed 2 July 2020).

Rosch, Eleanor; Mervis, Carolyn B.; Gray, Wayne D.; Johnson, David M.; Boyes-Braem, Penny. 1976. Basic objects in natural categories. *Cognitive psychology* 8/3. 382–439.

Russell, James A.; Feldman Barrett, Lisa. 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology* 76/5. 805–815. doi.org/10.1037//0022-3514.76.5.805.

Rychlý, Pavel. 2006. A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, 2008*. Eds. Sojka, Petr; Horák, Aleš. Masaryk University. Brno.

Scherer, Klaus R. 2009. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364/1535. 3459–3474. doi.org/doi:10.1098/rstb.2009.0141.

Scherer, Klaus R.; Schorr, Angela; Johnstone, Tom. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press. Oxford.

Sketch Engine. a. https://www.sketchengine.eu/ (accessed 2 July 2020).

Sketch Engine. b. https://bonito.sketchengine.eu/corpus/wsdef?corpname= preloaded/ententen13_tt2_1 (accessed 2 July 2020).

Sketch Engine. c. https://bonito.sketchengine.eu/corpus/wsdef?corpname= preloaded/ hrwac22_ws (accessed 2 July 2020).

Van Oirsouw, Robert R. 2019. *The syntax of coordination*. Routledge. Abingdon.

Korpusna analiza sintaktičko-semantičkih struktura s pomoću grafova: semantičke domene pojma *osjećaj*

Sažetak

Ova studija prikazuje metodu ConGraCNet na primjeru korpusne sintaktičko-semantičke analize s pomoću grafova pojma *osjećaj*/*feeling*. Analizom mreža leksičkih kolokacija koordinirane konstrukcije iz korpusa *enTenTen* i *hrWac* struktura semantičkih domena ishodišnih pojmova razlučuje se algoritmom prepoznavanja graf-zajednica. Leksičke se zajednice sagledavaju kao apstrakcija semantičkih domena povezanih s pojmovnom matricom ishodišnoga leksema. Korištenjem algoritmom centralnosti koji prepoznaje istaknuto umrežene lekseme određuje se stupanj povezanosti semantičke domene s izvornim pojmom. Ovaj empirijski pristup može se upotrebljavati za razvijanje NLP metoda za prepoznavanje semantičke sličnosti, razlučivanja višeznačnosti, strukturiranje značenja te za komparativne korpusne i međukulturne studije. Metoda ConGraCNet objavljena je kao mrežna aplikacija na stranici http://emocnet.uniri.hr/congracnet.

***Keywords*:** coordination, lexical graph analysis, emotions, centrality algorithm, community identification algorithm, corpus

***Ključne riječi*:** koordinacija, leksička graf-analiza, emocije, algoritam centralnosti, algoritam za prepoznavanje zajednice, korpus