A Construction Grammar Approach in the NooJ Framework: Semantic Analysis of Lexemes Describing Emotions in Croatian

Dario Karl(✉), Božo Bekavac, and Ida Raffaelli

Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
dariokarl.sl@gmail.com, {bbekavac,iraffaelli}@ffzg.hr

Abstract. The paper deals with semantic analysis of several lexemes encoding emotions in Croatian. The paper embraces the Construction grammar approach and shows how some of its basic theoretical tenets perfectly comply with the computational capabilities of NooJ. Using examples of the noun *strah* 'fear', the aim of the research is to point out the possibilities in annotating specific constructional meanings in NooJ, like different connotations of chosen lexemes, generalized uses of that constructions (their distributions in more abstract constructions like noun phrases), their relations with other constructions (other intensifiers of emotions, causative sentences etc.) and various distinctive features of their specific meanings (pragmatic features, as well as semantic and morphosyntactic), which all reflect different linguistic and cognitive phenomena in the language use.

Keywords: Construction grammar · Constructions · NooJ · Meaning
Croatian language

1 Introduction

The aim of this paper is to analyze the applicability of NooJ linguistic tool in the theoretical framework of Construction Grammar (CxG). The reason for choosing the CxG as a theoretical and methodological framework was to showcase various possibilities in formalizing linguistic data in a rather simple way, while holding meaning of a construction – basic linguistic unit – as the center of the research. A corpus-based analysis enables formalization of the constructions that have a specific meaning, distinguishable from the basic meaning of a certain lexeme that is part of a construction. Taking the example of the lexeme *strah* 'fear' we show the results of implementing the NooJ tool in the C&G framework. Moreover, we point out how the constructions and various usages of the lexeme are formalized in NooJ and later on tested with randomly chosen concordances from corpora, mirroring real language usage. The main goals of the research are twofold: (a) to investigate whether NooJ has potential for recognition of specific constructions and annotate them with corresponding meaning and to what degree, (b) to see to what extent CxG can be applied as a theory in building and

I. M. Mirto et al. (Eds.): NooJ 2018, CCIS 987, pp. 114–123, 2019.
https://doi.org/10.1007/978-3-030-10868-7_11

improving existing linguistic resources providing, moreover, with information about pragmatic and semantic features. Since CxG has never been used in NooJ, the main aim of the paper is to point to the compliance of NooJ linguistic tool with a theory that revolves around meaning and has its roots in Cognitive science and cognitive linguistics.

2 Construction Grammar – A Cognitive Science-Based Theory

In the article *Regularity and idiomaticity in grammatical constructions: The case of 'let alone',* Fillmore, Kay and O'Connor (1988) introduced the notion of a construction as a grammatical unit based on its syntactic and semantic features that cannot be described via regular 'rules' of grammar: "Constructions on our view are much like the nuclear family (mother plus daughters) subtrees admitted by phrase structure rules, except that (1) constructions need not be limited to a mother and her daughters, but may span wider ranges of the sentential tree; (2) constructions may specify, not only syntactic, but also lexical, semantic, and pragmatic information; (3) lexical items, being mentionable in syntactic constructions, may be viewed, in many cases at least, as constructions themselves; and (4) constructions may be idiomatic in the sense that a large construction may specify a semantics (and/or pragmatics) that is distinct from what might be calculated from the associated semantics of the set of smaller constructions that could be used to build the same morphosyntactic object" (1988: 501).

Accordingly, constructions, with all their syntactic, semantic, pragmatic and other features, make up the structure of language as it is. One construction can be made out of several different constructions, but its basic definition is that it can only be considered as a construction if its meaning could not be presupposed by just knowing the meaning of units within that construction: "an idiomatic expression or construction is something a language user could fail to know while knowing everything else in the language" (1988: 504).

The importance of construction as basic units in the language has been recognized also by Adelle Goldberg, who has pointed to an intertwined relationship between a verb and its arguments, (1995: 8–9; 11). It was actually Adelle Goldberg (1995; 2006) and William Croft (2001) who defined construction as any linguistic unit, regardless of its formal complexity or the level of abstractness, that has a meaning distinctive from any other in language. The increasing number of various researches shows a tendency to use constructions as keys for encapsulating and describing the entire grammatical knowledge of a speaker[1].

[1] For a detailed insight into differences in Construction grammar approaches and its implementation in the analysis of Croatian language structures see Katunar 2015.

2.1 Constructions and Cognitive Grammar

In a wider theoretical context of Cognitive linguistics, CxG is a theoretical and methodological framework considered as a syntactic alternative to Cognitive Grammar (Katunar 2015: 3). According to Belaj and Tanacković Faletar (2014), Cognitive Grammar considers a linguistic unit to be exclusively made of phonological and semantic pole or a connection between those two structures, whereas CxG considers linguistic form to be a syntactic structure. Within the CxG theoretical framework, the grammatical form would be a separate level in the formal structure (Belaj and Tanacković Faletar 2014: 33–34). In Cognitive Grammar the notion of grammatical form is a direct result of relation between semantic and phonological form. The other difference between these approaches is that Cognitive Grammar research is focused on schematic description of language and cognitive phenomena primarily describing the semantic pole (Katunar 2015: 36), whereas CxG is focused on formalized approaches to linguistic structures. These kinds of formalisms are considered as a theoretical and methodological backbone of our research enabling formalization of constructions in the NooJ linguistic tool. In contrast to Generative Grammar formalism, CxG formalisms include meaning, as well as pragmatic information, as core language features in describing constructions. Also, what distinguishes CxG from Generative Grammar is a negation of syntactic-centric and derivative approaches to grammar, interpreting grammatical relations as subordinate to semantic and pragmatic ones (Belaj and Tanacković Faletar 2014: 20–21). Moreover, since CxG is a usage-based model, the analysis of the language data is entirely usage based.

2.2 Usage-Based Model

Both Construction Grammar and Cognitive Grammar belong to cognitive-based usage models of language description that take language usage as a foundation for explaining mental structures of language and cognitive mechanisms connected to language (Katunar 2015: 27–28). One of the most important features of usage-based models is the importance of corpus-based research. It enables an in-depth analysis of frequencies, pointing to: (a) the linguistic structures that are more entrenched or innovative, (b) the interconnection between language and other cognitive systems and (c) the influence of language production on linguistic structures, among others (see Katunar 2015: 30–31; Barlow and Kemmer 2000).

Accordingly, we consider frequencies found in corpora as significant data that provide an insight into: (a) real language use, (b) distribution of particular lexemes in different language structures, (c) lexicalization patterns that can be defined as formalized constructions and (d) meanings conveyed by such constructions.

3 Methodology

The presented research is based on the analysis of the lexeme *strah* 'fear' that encodes the emotion of fear in Croatian language. Its distribution data were analyzed in the Croatian National Corpus (CNC)[2] and Croatian Web Corpus (HrWac)[3]. Both resources contain written texts automatically lemmatized and MSD tagged using standard heuristic methods. As a lexicographical source we used the online Croatian Language Portal (HJP) for digital overview of the already defined meanings. Defining lexeme distribution plays an important role in the context of usage-based models, especially when based on language corpus analysis. As Katunar points out (2015: 132–133), distribution is used to indicate semantic relations between two or more linguistic units (tokens in this case). Frequency data, mutual information and the logDice statistical method, for example, show if there is a semantic relation and a specific meaning that results from that distribution pattern. When a certain distribution pattern is lexicalized and its usage frequent enough, it is possible to consider it as a construction. Note that there are certain linguistic patterns that have an exact predictable meaning in all the contexts they are used in, but are still considered as a construction based on the merit of usage frequency (Goldberg 2006: 64).

3.1 Corpora Results for the Lexeme *strah*, Meaning 'fear'

The lexeme *strah* has 12 230 tokens in CNC and 177 740 in HrWaC. Its distribution patterns consist of a high frequency of prepositions:

(a) [*strah od čega*] ('fear of something') when the lexeme includes a prepositional phrase (PP)
(b) [*razlog za strah*] ('reason for fear'), when it is a part of another NP + PP
(c) [V + *bez straha*] ('without fear'), when it is a part of another V + PP
(d) [*V + u strahu*] ('in fear'), when it is a part of another V + PP construction.

It can also be found frequently next to particular verbs like:

(e) [*izazivati strah*] ('to cause fear')
(f) [*utjerati strah*] ('to instill fear')

And nouns/pronouns:

(g) [*strah koga/čega*] ('fear of somebody or something'),
(h) [*mene* (N+D) *je (biti) strah čega* (N+G)] ('I am/was scared of something'), being a part of a verb's argument structure

Online dictionary (HJP) has listed the following meanings:

- unpleasant emotion, state of anxiety and concern as a physiological response to a sense of danger (death, disease, punishment etc.) [*od straha da; od straha pred; u strahu od; sa strahom*] – for fear of, out of fear, with fear…

[2] http://filip.ffzg.hr/cgi-bin/run.cgi/first_form.
[3] http://nl.ijs.si/noske/all.cgi/first_form?corpname=hrwac;align.

- fright, reluctance or respect [*strah od starijih*] – fear of older (teenagers)
- concern for someone's safety [*u strahu zanjezin život*] – fearing for her life

Next to a couple of frozen expressions:

- *nema straha* – no need to fear
- *umrijeti od straha* – to be scared to death
- *strah i trepet* – fear and terror
- *u strahu su velike oči* – a frightened man sees danger everywhere

Taking into consideration that one of the main goals of this research was to define linguistic patterns that form constructions (Katunar 2015: 38), the distribution of the lexeme *strah* 'fear' points to several constructions which convey a specific meaning and/or are more frequent, i.e. which have a higher level of entrenchment:

- Construction [*strah za koga/što*], 'to care for something dear or someone's well-being, fearing that something bad could happen'.
- Constructions with the prepositions *od* ('of') and *prema* ('towards') and with lexemes that denote people in the meaning of 'awe': [*strah od starijih*] – 'being afraid of older people (teenagers)'; [*strah prema nastavnicima*] – 'fear of teachers'; it should be noted that the preposition *od* is mostly lexicalized in this meaning with the lexeme *strah*.
- Frequent usages of the lexeme *strah* with adjectives: [A + N] construction – the most frequent example being [*paničan strah*] ('a strong fear, panic').
- Constructions with prepositions and verbs conveying the meaning of unpleasant emotion, state of anxiety and concern as a physiological response to a sense of danger: [*strah od smrti*] – 'fear of death'; [*živjeti u strahu*] – 'to live in fear'.
- Constructions [*iz straha*] – 'out of fear' and [*zbog straha*] – 'because of fear' - a speaker can express fear as a cause that prevents a certain action.
- Frozen expressions with the aforementioned specific meanings as examples of substantive idioms, as defined by Fillmore and all.

3.2 Lexicalization Patterns

Distribution patterns of the lexeme *strah* in combination with HJP data gave us insight into the meanings that are lexicalized when the lexeme is found in different contexts. The next challenge was to analyze why and how do certain constructions acquire specific meanings. Raffaelli (2017: 175) notes that "the term lexicalization pattern comprises word-formation patterns as well as other grammatical (e.g. syntactic) patterns used in naming different concepts". Lexicalization patterns thus represent constructions as formal structures (morphosyntactic) which gain specific meanings and are shared by a larger number of speakers. Lexicalization patterns exhibit different degree of conventionalization and entrenchment (2017: 174). There is a connection between the frequency of a used construction (its distribution patterns), its level of conventionalization as a degree of speakers' shared knowledge about a certain construction and, thus, the cognitive entrenchment of that construction. In general, lexicalization is viewed as a naming process of a concept and lexicalization patterns as constructions that have been lexicalized and conceptually recognized by a large number of speakers.

Consequently, in this paper we make a clear distinction between distribution patterns (all the contexts in which a certain lexeme appears) and lexicalization patterns (distribution patterns that gain a status of constructions since they encode a specific meaning, different from lexical meaning of a certain lexeme). In NooJ, only those distribution patterns which have a high frequency of occurrences in the corpora and are considered as constructions, have been formalized and annotated.

4 Creating Grammars in NooJ

For creating NooJ syntactic grammars we have used the existing resources (dictionaries, morphological and lexical grammars) for Croatian language (Bekavac et al. 2007) that have successfully morphosyntactically annotated a large portion of tokens in Croatian texts. These annotations are described in this paper as metalinguistic generalizations that describe certain morphosyntactic, semantic and/or pragmatic categories in language. The advantage of NooJ in this theoretical and methodological framework is that it is flexible when it comes to inserting new annotations and offers different approaches to it (manually through grammars or dictionaries, but also automatically for describing wide linguistic phenomena or a specific and narrow linguistic category). Linguists actually have freedom to describe and name the category as they find appropriate, as long as it is correctly assigned in grammars later on. This functionality was used for inserting a couple of semantic annotations in the existing dictionaries and, ultimately, increased the precision of the grammars (Fig. 1).

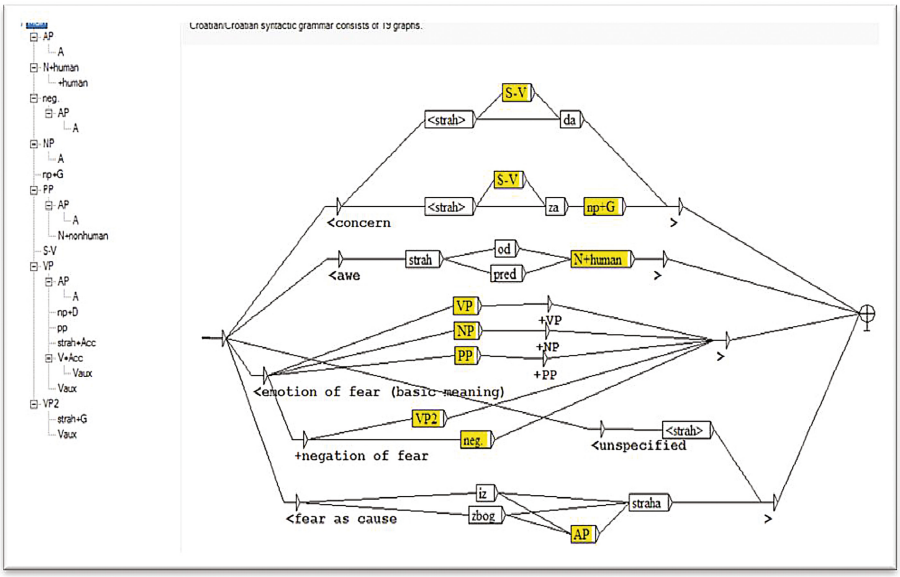

Fig. 1. NooJ syntactic grammar with annotated meanings for the lexeme *strah*

5 Results and Evaluation

Grammars were applied on 100 randomly extracted concordances from the Croatian National Corpus (CNC) which contain the targeted lexeme. We imported extracted concordances in NooJ and applied the existing Croatian language resources on them. A maximum recall of 100% represented one hundred lexeme usages from the CNC which were given an annotation, i.e. a construction was recognized by the grammar. The precision of NooJ grammars was measured by a number of correctly assigned annotations, both from the total recall percentage and from the total of 100 lexeme usages that were tested (Fig. 2).

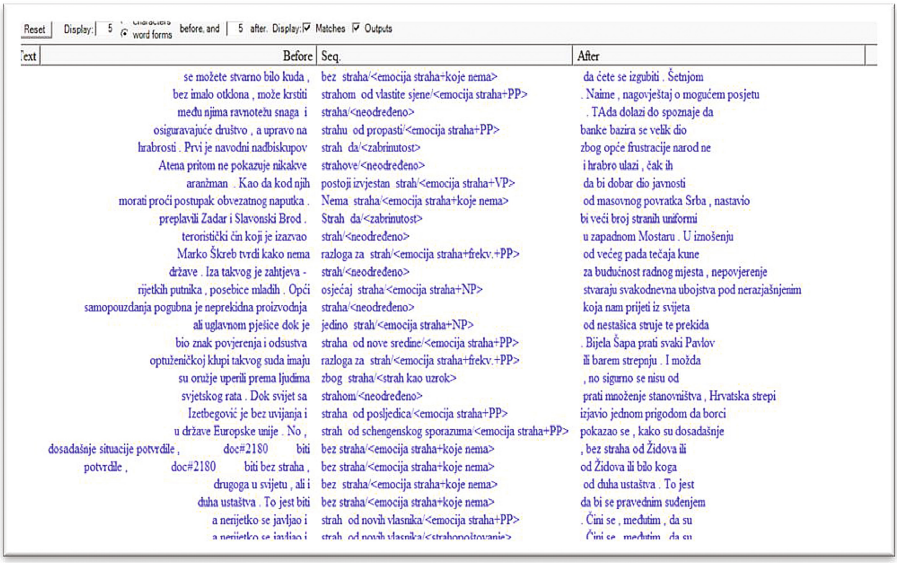

Fig. 2. Annotated concordances of the lexeme *strah*

Considering the fact that Croatian language has variable word order and that the speakers have a possibility to insert various constituents within certain frozen expressions and semi-frozen expressions[4], we had to be careful while creating grammars because NooJ syntactic grammars analyze only sequences of tokens. Everyday language use does in fact enable a speaker to break even the most formal and frequent constructions and insert various new and innovative expressions within those constructions. Those kinds of usages could have decreased the final precision of the grammars, but since we had a statistical insight for the distribution patterns in which

[4] Expressions where one or more elements fully or partially vary.

the analyzed lexemes might be used, we had that kind of usages in mind and annotated their possible (re)occurrence in the grammars (Table 1).

Table 1. Recall and precision percentages of NooJ syntactic grammars comparing results of the lexeme *strah* with other lexemes describing emotions of happiness, anger and sorrow

Emotion	Lexemes	Recall	Correctly annotated usages	Total precision
Happiness	*Sreća* ('happiness')	73%	97%	71%
	Sretan ('happy')	71%	88%	63%
Anger	*ljutiti se* ('to be angry')	52%	98%	51%
	Ljut ('angry')	69%	97%	67%
Fear	*Strah* ('fear')	78%	86%	67%
	Strašan ('horrible')	76%	83%	63%
Sorrow	*Tuga* ('sadness')	100%	94%	94%
	Tužan ('sad')	61%[a]	61%	45%

[a]39 usages were not annotated, but 23 were annotated with two possible meanings.

Table 2. Recall and precision percentages for constructions containing the lexeme *strah*

Construction	Meaning	Recall	Correctly annotated usages	Precision
[*strah od čega*]	'fear of something'	23%	18%	78%
[*razlog za strah*]	'reason for fear'	3%	3%	100%
[*V (živjeti) + u strahu*]	'(to live) in fear'	6 (1)%	5%	83%
[*izazivati strah*]	'to cause fear'	/		
[*utjerati strah*]	'to instill fear'	/		
[*strah + subj. + koga/čega*]	'someone is in fear of somebody or something'	/		
[*sa strahom]*	'with fear'	2%	2%	100%
[A + *strah*]	'adjective + fear'	19%	17%	89%
[*strah od N + human (starijih)*]	'being in fear of older people (teenagers)'	2%	1%	50%
[*strah + za + koga*]	'concern for someone's safety'	4%	4%	100%
[*zbog straha*]	'because of fear' – fear as a cause	5%	5%	100%
[*nema straha*]	'no need to fear'	1%	1%	/
[bez straha]	'without fear'	13%	11%	100%
[*umrijeti od straha*]	'to be scared to death'	/	/	/
[*strah i trepet*]	'fear and terror'	1%	1%	100%
[*u strahu su velike oči*]	'a frightened man sees danger everywhere'	/	/	/
[]	Unspecified usage	22%	/	/

The total recall of all the NooJ syntactic grammars for the analyzed lexemes was 72.50%. Out of those annotated usages, NooJ grammars have correctly annotated 88% constructions and their overall precision was 65.125%. This means that over a third of lexeme usages were correctly annotated with their meaning by the grammars.

The numbers for the lexeme *strah* ('fear') roughly correspond to the overall number, but different meanings and usages of the lexemes had different results. The most frequent was the 'emotion of fear' (66 usages), 22 of them being a construction with prepositions: [*u strahu*] ('with fear') and [*od straha*] ('out of fear'). On the other hand, 'awe' had 3 usages, 'expression of concern' had 4 and 'fear as a cause for preventing actions', such as [*zbog straha*] ('because of fear') had 5, with a 100% precision (Table 2).

The results have shown that NooJ has the ability to produce high percentages of precision for automatic semantic annotation, but the question from the research itself arose about the causes that lay in the back of varying results, not just for lexemes, but for the meanings themselves, besides the possible incorrectly annotated tokens or variable word order.

6 Conclusions and Future Work

In this work we used NooJ to formalize lexicalization patterns of a chosen lexeme, describing one of the basic emotions in Croatian language. One of the biggest advantages of NooJ is the flexibility of inserting new and multiple annotations. Since we have considered annotations as being the metalinguistic generalizations of morphosyntactic, semantic and pragmatic information, NooJ makes it rather simple for a linguist to note a certain linguistic phenomenon and test it straight away. One example that was important for us was including annotation of [+human] in a large number of tokens in a simple and quick manner. It has notably increased the precision of constructions such as *strah od starijih* ('fear of older people') and helped differentiate them from *strah od smrti* ('fear of death') and other constructions with the overall meaning of 'of anxiety and concern as a physiological response to a sense of danger'.

Furthermore, NooJ grammars had a higher precision for constructions with a lower recall, while primary or basic meanings made up around 50% of lexeme usages. Also notable was the fact that the primary meanings had more lexicalized patterns, i.e. constructions formed in different prepositional, noun and verb phrases (*strah od smrti* 'fear of death', *u strahu/sa strahom* 'in fear/with fear', *osjećati strah* 'to feel fear', *iznenadan strah* 'sudden feeling of fear'…), which made them more challenging to formalize in NooJ. Other linguistic phenomena, such as causative constructions and intensifiers, were formalized in NooJ without problem (*zbog straha* 'because of fear', *silan strah* 'strong fear'…), although it would be interesting to see if it is possible to formalize argument structure constructions the way Goldberg showed in her research (1995; 2006).

In its present form our work contains only lexeme that encodes emotion of fear in Croatian language. The project described in this paper has served us as an experiment and a starting point for developing grammars for other lexemes using proposed methodology. For further research in the area of automated semantic recognition, more

metalinguistic generalizations, such as pragmatic and semantic annotations, would be necessary in resources for Croatian language.

Evaluation results have shown that NooJ provides sufficient means for involving further applications of this construction-based methodology. Focusing on fine tuning of developed grammars could increase both precision and recall of results, taking into account that we experimented with highly complex language structures. Current research was based on semasiological structure, but it is also possible to analyze onomasiological structures and see how and which patterns are more likely to lexicalize a particular meaning.

References

Barlow, M., Kemmer, S.: Usage-Based Models of Language. CSLI Publications, Stanford (2000)

Bekavac, B., Vučković, K., Tadić, M.: Croatian resources for NooJ. 2007 NooJ Conference Book of abstracts

Belaj, B., Tanacković Faletar, G.: Kognitivna gramatika hrvatskoga jezika. Disput, Zagreb (2014)

Croft, W.: Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford University Press, Oxford (2001)

Fillmore, Ch., Kay, P., O'Connor, M.C.: Regularity and idiomaticity in grammatical constructions: the case of 'let alone'. Language **64**(3), 501–538 (1988)

Goldberg, A.E.: A Construction Grammar Approach to Argument Structure. The University of Chicago Press, Chicago (1995)

Goldberg, A.E.: Constructions at Work. The Nature of Generalization in Language. Oxford University Press Inc, New York (2006)

Katunar, D.: Ustroj leksikona u konstrukcijskoj gramatici – primjer prijedloga u hrvatskom jeziku. Faculty of Humanities and Social Sciences, Doctoral thesis, Zagreb (2015)

Raffaelli, I.: Conventionalized patterns of colour naming in Croatian. In: Cergol Kovačević, K., Udier, S.L. (eds.) Applied Linguistics Research and Methodology, Proceedings from the 2015 CALS Conference, pp. 171–186. Peter Lang Verlag, Franfurt am Main (2017)