https://doi.org/10.17234/INFUTURE.2019.6

Preliminary communication

A Corpus-Based Approach to Reevaluation of Croatian Verb Classification

Danijel Blazsetin
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
dblazset@ffzg.hr

Petra Bago
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
pbago@ffzg.hr

Summary

Croatian grammar textbooks have a long tradition of classifying verbs based on their morphosyntactic characteristics. Conclusions, such as the frequency or productiveness of a class, were drawn without having the insight into a big corpus. Corpora used in such descriptions were not described and were presumably made of literary works which is, in our opinion, describing a form of the Croatian language distant from its everyday use. The corpus used for analyzing verbs in this paper is hrWaC which contains 1.9 billion tokens and about 90,000 verbs. This corpus was selected with the intention of describing and analyzing a less formal and less standardized language This paper offers a corpus-based approach to the problem of verb classification and emphasizes the importance of NLP methods in the process of classification as they fasten and simplify it. The paper gives a brief introduction to verbs, their morphological characteristics and their classification. By extracting verbs from the Croatian web corpus hrWaC and processing them computationally, the paper gives an insight into the verb distribution in the Croatian language and points out some difficulties that were encountered during this study. Even though this paper aimed to reevaluate the existing data data, the present findings mostly confirm the claims of previous researches. A number of recommendations for future research are given, foremost, the need of the extension of the language material.

Key words: corpus linguistics, natural language processing, verb classification, grammar textbooks, Croatian language

Introduction

Reviewing Croatian grammar textbooks, one can find information about the frequency of Croatian verb classes. This paper uses modern technology tools and methods to reevaluate the existing statistics regarding Croatian verb classes. Applying natural language processing methods to this field, we can speed up the process of verb classification and get exact information i.e. the frequency of Croatian verb classes. It is important to mention that existing descriptions are mostly based on corpora of standardised Croatian language, whilst the frequency information given in this paper analyzes verbs from a web corpus which is comprised of diverse discourses, including texts written in the standardised version of the Croatian language as well as texts written in informal, colloquial version of the Croatian language. The presented model offers an automatic verb classification system that was applied and tested on the hrWaC web corpus. For the purposes of this paper, it was necessary to build three different groups of verbs that would simulate the corpus as a whole: *common verbs*, *occasional verbs*, *rare verbs*, as will be later explained in more detail. This paper also offers a comparative analysis of the existing works on frequency distributions of Croatian verb classes and the presented research.

Verb classification

Verb classifications are based on the verbs' morphological attributes. In Slavic languages there are two approaches to verb classifications: classifications founded on the verb's present tense base and classifications founded on the verb's infinitive base (Marković, 2012: 217-219). To understand the classifications, we must first define the morphological characteristics that determine a verb's class. If we compare a verb's infinitive form with its present tense form, we can see three elements: the verbs stem, the suffix that denotes its conjugational class and the derivational morph that indicates the verb's tense (Table 1).

Table 1. Comparison of the morphology of the infinitive and the present form.

infinitive form	*gled*-***a***-*ti* (Eng. *to watch)*)	*rad*-***i***-*ti* (Eng. to work)
present form	*gled*-***a***-*m* (Eng. I watch)	*rad*-***i***-*m* (Eng. I work)

Because of linguistic economy, words (including verbs) tend to organise themselves according to their similarities. Verb classifications aim to discover the patterns in the verbs' groupings and describe them (ibid: 197).

As mentioned, there are several ways of approaching the problem of verb classification. It seems like nowadays Croatian linguistics prefers the classifications based on the verbs' present tense form. Hence the presented model uses aforesaid classification as well (Bošnjak Botica, 2013: 65). We can find examples of this classification in the highly influential grammar textbook written by Josip Silić and Ivo Pranjković. Because of its prevalence, it is chosen to be the foundation of the verb class frequency analysis in this paper. The classification defines six verb types that can be divided into several classes (Silić, Pranjković, 2005).[1]

Related work

Information about the frequency and prototypness of verb classes can be found in Croatian grammar descriptions. However, it seems like the majority does not provide information about the corpus on which the classifications were applied to and as such prevents future researchers from comparing the diverging results of grammar textbooks and other related studies (cf. Babić et al., 1991; Raguž, 1997).

The corpus that Jelaska (2003) based her research on was the so called *Moguš's corpus*. Compared to today's corpora, it is small, and compiled from texts written in the standardised variant of the Croatian language. As such, it represents an artificial form of the language (cf. Tadić, 1997: 389-390). Hence we believe that information gathered from that corpus cannot describe the language as it is in its everyday use. Jelaska and Bošnjak Botica (2019) made an extensive research on verb class frequency which contains 24,538 verbs, but they do not provide any information about the corpus the research was based on. The main drawback of these studies is that they do not give any information about the corpus they are based on.

We think that dealing with language should always include modern technologies ie. natural language processing methods. Our research combines corpus linguistics as well as NLP methods, and presents an automated verb classifier that could be used on any given corpus in order to define the frequency of verb classes. Some of the biggest advantages of our approach to verb classification are its automatization, upgradeability and feasibility. The corpus used for the purpose of this paper is the hrWaC corpus.

Methodology

Corpus hrWaC contains 1.9 billion tokens, is made of HTML documents found on the *.hr* top-level domain and is the first of its kind of Croatian language (Ljubešić, Klubička, 2014). It is an annotated and a searchable web corpus that can be accessed via Sketch Engine.[2] We find it crucial that the corpus is based on documents found on the web as it means that it does not only represent the standardised version of Croatian, but also includes the language used in fashion magazines, newspapers, blogs, advertisements, user responses, forum discussions etc. Thus it reflects written language in its everyday use. A corpus of this kind can attest the jargon of different groups; speakers' doubt in using the correct word forms; the usage and frequency of loanwords; problems with orthography and trends in a language. Unfortunately, such corpora have their disadvantages as well.

[1] The verbs of the first type can be sorted into eighteen classes, but in this paper, we ignored those classes as they would not be useful for the comparative analysis.

[2] https://www.sketchengine.eu/

One of the biggest issue is that they are not representative as they do not include all types of texts (for example literary works are barely found in web corpora because they are copyright protected) and the documents that make up the corpus might not be reliable (cf. Fletcher, 2011).[3] In the process of making this model we ran into some _web corpora_ specific issues that include incorrectly lemmatised or tagged words and improperly written words. This, on the one hand, means that the documents in the corpus did not overgo a process of selection, meaning that nothing is censored, but on the other hand it raises the problems of reliability of the statistics calculated from the corpus. Besides the benefits of such corpora, we would like to emphasize that one must always bear in mind the nature of alike corpora when he/she relies on its statistics.

Firstly, to classify verbs we had to gather information from the corpus hrWaC. As mentioned, to classify verbs one must _know_ the verb's infinitive form and its present tense form. Using Sketch Engine we created a verb frequency list ie. a lemma frequency list of the verbs. Since Sketch Engine only allows the export of 1,000 verbs long lists, we had to define categories that could represent the corpus as a whole. We outlined three categories: _common verbs_, _occasional verbs_, _rare verbs_. Naturally, the _common verbs_ category is constituted by the top of the frequency list and contains the 1,000 most frequent verbs. Then we defined the _rare verbs_ category, which includes 1,000 verbs that occur from 18 to 25 times in the corpus. We wanted to exclude _hapax legomenon_ as well as verbs that are occurring more than once, but are still very rare.[4] Then we had to define the _occasional verbs_ category. As the lower bound of the frequency of the _common verbs_ is 825 instances and the most frequent verb in _rare verbs_ occur 25 times, the _occasional verbs_ category had to be somewhere in between. In order to have a consistent methodology, it was mandatory to have a thousand verb category for the _occasional verbs_ as well. If we try do define this category on the arithmetic middle of the two numbers we will have a category with only a few hundred members so we decided that this category would include verbs that occur from 140 to 375 times. Therefore the category _occasional verbs_ contains 1000 verbs as well.

The three derived lists were merged and, as we only had their infinitive form, their present tense form was defined. This step was done manually.[5] Throughout this process verbs that were not suitable for the analysis were removed, resulting in a corpus that counts 2,588 verbs.[6]

The program compares a string to several regular expressions ie. words to patterns. The present tense form and the infinitive form of a verb are paired and are part of a list of all the verbs. The program compares the infinitive form and the present tense form of a verb to class-specific patterns. For example, the pattern for the infinitive form of the verbs in the fourth class of the third type (e. g. _držati_ (Eng. _to hold_)) is defined as follows: _r'.*(š|č|ž|j|št|žd)ati\b'_. The value of this expression is _True_ if the string ends with _šati, čati, žati, jati_, _štati_ or _ždati_. The regular expression for the present tense form of the same class (e. g. _držim_ (Eng. _I hold_)) is _r'.*im\b'_. This expression returns _True_ if the string ends with _im_. If both of the expressions' values are _True_, the program classifies the verb (its infinitive and present form) into the matching class and removes it from the list which the program iterates through. Because some verbs would pass several regular expressions, there is a defined order of comparing the verbs with the patterns. For example, the verb _razgledavati_ (Eng. _to sightsee_) both in its infinitive and present form _razgledavam_ (Eng. _I sightsee_) matches the pattern of the first class of the fifth type ie. _r'.*ati_\b'-r'._*am\b'_, but it belongs to the second class of the fifth type ie. _r'.*avati_\b'-r'.*am\b'. As it is seen, defining the order is mandatory and is a crucial step in the classification process. When the program is finished, the user gets a document with a _.txt_ extension that contains the classified verbs.

[3] Even though these corpora are not representative in its narrower sense, they can, for example in our situation, be a representative corpus of the language on the web.

[4] We must not forget that we are working with a huge web corpus and that same incorrectly written words can occur many times in the corpus. Even the category _common verbs_ contains incorrectly written verbs that had to be excluded from the statistics.

[5] We think that in the future, when dealing with huge ammount of data (verbs), we will automate the process of defining the present tense form by utilizing hrLeX (http://nlp.ffzg.hr/resources/lexicons/hrlex/).

[6] We excluded verbs which are not correctly written, which are wrongly lemmatised or which contain typographical errors.

Results and discussion

Before the comparison, we would like to give a brief overview of the existing frequency data of the verb classes. There is only a handful of grammar textbooks that contain information about verb frequency and a few research papers that give insight into verb class frequency. It seems like exact data regarding the number of verbs in a class does not preoccupy Croatian linguists (Marković 2012: 220). Instead of listing all the conclusions about verb class frequency found in grammar textbooks, we will only be illustrating how grammar textbooks inform the reader about verb class frequencies. In Babić et al. (1991) one can find such statements: ”There are approximately sixty verbs like *vidjeti-vidim* (Eng. *to see-I see*)[7] and two hundred more derivatives.” or ”There is a lot of verbs like *misliti-mislim* (Eng. *to think-I think*)”. Raguž (1997) states the following: ”There are a few hundred verbs of the type *vidjeti-vidim* (Eng. *to see-I see*)” and ”There are approximately 6,000 verbs like *misliti-mislim* (Eng. *to think-I think*)”. As we can see, the given information are not precise. There is more accurate and exact information in the works of Jelaska (2003) and Jelaska and Bošnjak Botica (2019). Jelaska (2003) categorized 16,000 of the most frequent verbs that were extracted from the *Moguš's corpus*, while Jelaska and Bošnjak Botica (2019) categorized 24,538 verbs, however the used corpus is unknown. In the following tables (Tables 2, 3 and 4) we can see the results of their research and their comparison with our research (Table 5).

Table 2. Percentage of the classes' representation in regards to all the verbs

Type	Class	100	100 (by type)	16,000 (by type)
a	*gledati-gledam* (Eng. *to watch)*)	22%	22%	36%
i	*moliti-molim* (Eng. *to pray*)	26%	37%	30%
	voljeti-volim (Eng. *to love*)	6%		
	bježati-bježim (Eng. *to run away*)	5%		
e	*dignuti-dignem* (Eng *to lift*)	0%	12%	29%
	vjerovati-vjerujem (Eng. *to believe*)	4%		
	davati-dajem (Eng. *to give*)	1%		
	smijati se-smijem se (Eng. *to laugh*)	2%		
	plesati-plešem (Eng. *to dance*)	5%		
ø	*naći-nađem* (Eng. *to find*)		29%	5%

Source: Jelaska (2003: 56)

[7] The translations of the Croatian verbs were added by the authors.

Table 3. Number of verbs in the classes

Representative verbs	Class frequency	Verb type	Verb type frequency
gledati-gledam (Eng. *to watch*)	9,590		
		a	9,590
moliti-molim (Eng. *to pray*)	7,011		
vidjeti-vidim (Eng. *to see*)	509		
trčati-trčim (Eng. *to run*)	225		
		i	7,745
pisati-pišem (Eng. *to write*)	1,325		
smijati se-smijem se (Eng. *to laugh*)	337		
putovati-putujem (Eng. *to travel*)	2,621		
davati-dajem (Eng. *to give*)	67		
viknuti-viknem (Eng. *to yell*)	1,463		
		e1	5,813
naći-nađem (Eng. *to find*)	1,390		1,390
		e1+e2	7,203
Total number	24,538		24,538

Source: Jelaska, Bošnjak Botica (2019: 64)

Table 4. Number of verbs in the classes based on the research presented in this paper

		Representative verb	Number	Number by type	% by class	% by type
I. type[8]		*ići-idem* (Eng. *to go*)	278	278	10.7	10.7
II. type		*viknuti-viknem* (Eng. *to yell*)	96	96	3.7	3.7
III. type	1. class	*pisati-pišem* (Eng. *to write*)	143	172	5	6
	2. class	*pljuvati-pljujem* (Eng. *to spit*)	1			
	3.class	*grijati-grijem* (Eng. *to heat*)	28		1	
IV. type	1. class	*raditi-radim* (Eng. *to work*)	821	897	31.7	34.5
	2. class	*vidjeti-vidim* (Eng. *to see*)	49		1.8	
	3. class	*trčati-trčim* (Eng. *to run*)	27		1	
V. type	1. class	*kopati-kopam* (Eng. *to dig*)	810	952	31.3	36.2
	2. class	*proučavati-proučavam* (Eng. *to study*)	142		4.9	
VI. type	1. class	*kupovati-kupujem* (Eng. *to buy*)	49	187	1.8	6.6
	2. class	*smanjivati-smanjujem* (Eng. *to reduce*)	138		4.8	
∑			2,582+5	2,582+5	100	100

[8] The first verb type was not separated into classes. The Silić and Pranjković grammar textbook (2005) defines 18 classes in the first type. The criteria for the classes are really specific, hence we believe that it would be redundant to separate the verbs in the first type to classes as we will not use those numbers in the comparative analysis.

Table 5. The comparison of the works of Jelaska (2003), Jelaska and Bošnjak Botica (2019) and our research

	Jelaska (2003)	Jelaska and Bošnjak Botica (2019)	Our research
I. type	5%	5.66%	10.7%
IV. type	30%	32.02%	34.5%
V. type	36%	39.08%	36,2%
II. type	29%	23.23%	16.3%
III. type			
VI. type			

Firstly, we would like to emphasize that the research seen in Jelaska (2003) and Jelaska and Bošnjak Botica (2019) analyzes significantly more verbs than our research. However, it is our belief that, even though our research analyzes fewer verbs, due to it being based on three different frequency categories, it can serve as a valid element in the comparison. The statistics conducted in different studies do not differ much. This means that the verbs in hrWaC are similar to those in the *Moguš's corpus*. Thus, putting emphasis on rare verbs might give us a more exciting insight into verb classification. We shall highlight the differences and similarities between the existing statistics here. Verbs like *gledati-gledam* (Eng. *to see-I see*)[9] are the most frequent in every statistic and are followed by the verb type *misliti-mislim* (Eng. *to think-I think*). Classes inside the verb type *misliti-mislim* (Eng. *to think-I think*) differ though. According to Jelaska (2003), 6% of the verbs belong to the class *voljeti-volim* (Eng. *to love-I love*), while 5% to the class *trčati-trčim* (Eng. *to run-I run*). In Jelaska and Bošnjak Botica (2019) and our research, these percentages are 2% and 1%, respectively. It is interesting how the verbs *bosti-bodem* (Eng. *to stab-I stab*)[10] in Jelaska (2003) and Jelaska and Bošnjak Botica (2019) make only 5% of all the verbs, while in our research it is as high as 10.7%. It is our opinion that this percentage would gradually decrease if we added more verbs to our analysis.[11] It is usually said that the verb type *dignuti-dignem* (Eng. *to lift-I lift*) is frequent and productive (cf. Babić et al. (1991); Raguž (1997). However, Jelaska and Bošnjak Botica (2019) and this research found that only 5.96% and 3.7% of all verbs, respectively, belong to this type.

As shown by our analysis, the differences are modest, therefore we believe that comparing different frequency categories of our research could be useful and could give insight to the system of verb classification.

In this research, verbs like *misliti-mislim* (Eng. *to think-I think*) and *gledati-gledam* (Eng. *to see-I see*) are the most frequent in every frequency category. However, it has to be emphasized that while in the common verbs category *misliti-mislim* (Eng. *to think-I think*) makes 40.3% of all the verbs and *gledati-gledam* (Eng. *to see-I see*) only 28.8%, in the rare verbs category *misliti-mislim* (Eng. *to think-I think*) decreases to 30.8% and *gledati-gledam* (Eng. *to see-I see*) raises to 46.8%. The fact that there are verbs in the rare verbs category such as **odblokirati* (Eng. *to unblock someone on social media platforms?), *štrumpfetati* (Eng. *to act like Smurfette; to be an easy girl?*) indicates the prototypness of the class *gledati-gledam* (Eng. *to see-I see*). It is more likely that Croatian speakers will make up the word *štrumpfetati* and not *štrumpfetjeti*.[12]

As expected, the percentage of the verb type *bosti-bodem* (Eng. *to stab-I stab*) decreases with the growth of the number of the analysed verbs. However, atypicality does not always correlate with high frequency. This is supported by the fact that 5.7% of the verbs in the rare verbs category belong to the class *bosti-bodem* (Eng. *to stab-I stab*). Such verbs in the rare verbs category are: *rastresti* (Eng. *to shake up*) *prigristi* (Eng. *to have a bite*), **štići* (Eng. *to arrive*), *crpsti* (Eng. *to draw out*).

The *viknuti-viknem* (Eng. *to yell-I yell*) class is fairly low in all the categories: 2.3%, 3.7% and 5.4%.

[9] To avoid confusion, in this section we will not name a verbs' class to determine it, but a prototype of its class (eg. *gledati-gledam* (Eng. *to see-I see*) instead of type V. class 1.). This is necessary as Jelaska and Bošnjak Botica use a slightly different classification in their works.

[10] These verbs are traditionally classified into the first verb type. They are unique and unusual because their suffix that denotes its conjugational class is a zero morph (ø) and the stem cannot be seen from the infinitive verb form (eg. *jes-ø-ti, jed-e-m* (Eng. *to eat-I eat*)).

[11] We will discuss this statement below.

[12] However, this paper does not aim to define the prototypness of the Croatian verb classes.

Babić et al. (1991) mention that the class *kupovati-kupujem* (Eng. *to buy-I buy*) is big, however in our research there are only 49 verbs out of 2,587 which are classified in this category. Because of the verbal aspect pairs in Croatian language, we can indeed produce a lot of verbs that will belong to this category. However, it seems they are barely present in the written language of the Internet. It would be interesting to compare this frequency distribution to verbs extracted from a spoken corpus.

We analyzed 2,587 verbs, presenting the data as 2,582+5. The five isolated verbs belong to the irregular class and they do not fit into any of the classes. These verbs are: *biti-jesam* (Eng. *to be-I am*), *moći-mogu* (Eng. *to can-I can*), *spati-spim* (Eng. *to sleep-I sleep*), *zaspati-zaspim* (Eng. *to fall asleep-I fall asleep*) and *htjeti-hoću* (Eng. *to want-I want*).

Conclusion

In this paper we offer a corpus-based approach to the problem of verb classification in Croatian language extracting verbs from the hrWaC corpus. As we have seen there are studies that are based on a bigger corpus of verbs. However we believe that the uniqueness of our research lays in the fact that it is based on a web corpus, hence mirrors a variation of Croatian language that is similar to its everyday use. It has to be stated that a research which analyzes 3,000 verbs cannot reflect a true and comprehensive picture of the language, even if frequency categories were assembled. Thus the results of this research have to be dealt with reservations. However, it seems that a similar approach to the problems of verb classifications in Croatian language could shed light on some tendencies in the language and present new numerical data. Although this paper aimed to reevaluate the existing statistics regarding the number of verbs in various verb classes, it, first of all, points out the significance of the usage of NLP methods in language research and linguistics. The strength of the presented tool lays in its reusability and easy application. Future studies could fruitfully explore the issue of verb classification by analysing the corpus as a whole. We believe that a throughout analysis could either decisively confirm the existing data regarding verb classes or verify our initial hypothesis ie. the authors of Croatian grammar textbooks did not have access to this big amount of data so the information about the frequency of verb classes should be reevaluated. This study tried to offer an approach to the process of reevaluation. A verb analysis as shown in this paper could also be useful in the making of a language learning program for those learning Croatian as a second language as it is known that verbs of the same class have the same inflection, and derivational phenomena can also be generalised. Such approaches (ie. that take into account the prototypness and frequency of the verbs and their classes) in language teaching are already a trend and this type of clearly *digitally born data* could expand the previously proposed programs database. If, in the future, a corpus of contemporary and standard Croatian language is made, with the application of this method anyone can come to conclusions regarding the verb class frequencies. On the other hand, the paper highlights that statistics and data made by programs always have to be supervised by humans. The future of linguistics is based on the interdisciplinary approach to language investigation thus researchers have to accept the challenges and incorporate computational methods and tools into their field of interest. However, we should percieve computational methods as tools which help us analyze language and not as an approach that substitutes linguists.

References

Babić, S., Brozović, D., Moguš, M., Pavešić, S., Škarić, I., Težak, S. (1991). Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika: Nacrti za gramatiku. Zagreb: HAZU: Globus

Bošnjak Botica, T. (2013). Opća načela podjela na glagolske vrste u hrvatskome u perspektivi drugih bliskih jezika. // Lahor 1, 15, 63-90

Fletcher, W. H. (2012). Corpus analysis of the world wide web. // The encyclopedia of applied linguistics / Chapelle, C. A. (ed.). Hoboken, NJ: John Wiley & Sons

Jelaska, Z. (2003). Proizvodnja glagolskih oblika hrvatskoga jezika kao stranoga jezika: od infinitiva prema prezentu. // Zbornik Zagrebačke slavističke škole 2002. / Botica, S. (ed.). Zagreb: FFpress Filozofski fakultet, 48-63

Jelaska, Z., Bošnjak Botica, T. (2019). Conjugational Types in Croatian. // Rasprave: časopis instituta za hrvatski jezik i jezikoslovlje 45, 1, 47-74

Ljubešić, N., Klubička, F. (2014). {bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian. // Proceedings of the 9th Web as Corpus Workshop (WaC-9). Bildhauer, F., Schäfer, R. (eds.). Gothenburg: Association for Computational Linguistics, 29-35

Marković, I. (2012). Uvod u jezičnu morfologiju. Zagreb: Disput

Raguž, D. (1997). Praktična hrvatska gramatika. Zagreb: Medicinska naklada

Silić, J., Pranjković, I. (2005). Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta. Zagreb: Školska knjiga

Tadić, M. (1997). Računalna obrada hrvatskih korpusa: povijest, stanje i perspektive. // Suvremena lingvistika 23, 43-44, 387-394