Corpus Analysis of Complex Names with Common Nouns in Croatian

Ivana Matas Ivanković[0000-0002-9796-8346] and Goranka Blagus Bartolec[0000-0002-3577-7026]

Institute of Croatian Language and Linguistics, Republike Austrije 16, 10000 Zagreb, Croatia
{imatas,gblagus}@ihjj.hr

Abstract. The goal of this corpus-based research is to see can the complex names with common nouns in their composition be extracted from Croatian hrWaC v2.2 corpus by using regular expressions, i.e. to what extent the capital letter (not the one after the full stop, the exclamation mark or the question mark) can be taken as an indication of a name. Common noun can be used as a regular noun or as a constituent of a complex name, which, on one hand, makes it difficult to tag them automatically, and on the other hand, affects the lexicographic description. With the help of regular expressions, we searched for capitalized common nouns and for sequences in which a capitalized attribute is on the first place and the common noun follows it. After analyzing 1000 examples in each search, we divided results into two groups: names and sequences with an uppercase letter that are not names. Some of the causes of extracting "false" names are technical (e.g. interpunction: separating sentences with paragraph mark (¶), lack of interpunction at the end of sentence; whole parts of text written in upper case...), and some of them lie in the texts crawled for hrWaC, which are not written in accordance with Croatian orthography.

Keywords: Complex Names, Croatian Orthography, Corpus Search.

1 Introduction

Proper names denote a particular person, place, organization, ship, animal, event, or other individual entity. They can be divided into proper nouns (single words like *Europe*, *Mars*...) or complex names (phrases, multiword expressions like *United States of America*, *Faculty of Humanities and Social Sciences*...). Uppercase letter is usually a sign of proper name. In Croatian, uppercase letter stands at the beginning of the sentence and quotation, at the beginning of names (personal names, surnames and nicknames, names of animals, geographical names, names of inhabitants and members of the nation, other names, and possessive adjectives derived from the name) and at the beginning of words of respect and honor. In this work, we have focused on common nouns as a part of complex names. Our goal was to see can the complex names with common nouns in their composition be extracted from Croatian hrWaC

EUROPHRAS 2019, pages 106–113,
Malaga, Spain, September 25-27, 2019.
https://doi.org/10.26615/978-2-9701095-6-3_014

v2.2 corpus[1] by using regular expressions, i.e. to what extent the capital letter (after excluding the capital letter after the full stop, the exclamation mark or the question mark) can be taken as an indication of a name. Common nouns can be used as regular nouns or as constituents of complex names, which, on one hand, makes it difficult to tag them automatically, and on the other hand, if a noun is a part of complex name, "a new word sense must be created in the lexicon" [3: 443]. In lexicography, especially in determining the collocations of a word, it is important to establish the difference between a regular multiword expression and a complex name. Extracting of names from corpus could also be useful help in complementing a *Dictionary of upper and lower initial case in Croatian*, which is being developed in the Institute of Croatian Language and Linguistics.

The rest of the work is structured as follows. In the second chapter the detailed parameters of the corpus search are presented, the results of the search are analyzed in the third chapter, and conclusion is in the fourth chapter.

2 Analysis

Proper nouns are tagged in hrWaC and the search of proper nouns[2] provided us with results like (first ten sorted by frequency and lemma): *Hrvatska*, *Zagreb*, *Ivan*, *EU*, *Europa*, *Zadar*, *Split*, *Rijeka*, *Hrvat*, *HDZ*. Wider list of 50 examples shows similar results: among them are the names of people (*Ivan*, *Marija*, *Ante*), cities (*Zagreb*, *Zadar*, *Split*), continents (*Europa* 'Europe', *Amerika* 'America'), states (*Hrvatska* 'Croatia', *Srbija* 'Serbia'), nations (*Hrvat* 'Croat', *Srbin* 'Serb'), football team (*Hajduk*). Two things should be noted: 1) Abbreviations tagged as proper nouns gave us abbreviated names of states and unions, like *EU*, *BiH* (*Bosna i Hercegovina* 'Bosnia and Herzegovina'), *SAD* (*Sjedinjene Američke Države* 'United States of America'), abbreviated names of parties, like *HDZ* (*Hrvatska demokratska zajednica* 'Croatian Democratic Union'), *SDP* (*Socijaldemokratska partija* 'Social Democratic Party'), but it also gave us *TV*, which is the abbreviation for the common noun 'television'. 2) Lemma "Bi", listed on 35th place by frequency, as a lemma does not exist in Croatian, and in HrWaC it stands for Bog (in examples like: *Taj put započinje krštenjem po kojem možemo* ***Boga*** *nazivati Ocem...* 'This path begins with baptism by which we can call God as Father') and for *Bin (Laden)* (in examples like *U njoj se opisuju aktivnosti Osame* ***Bin*** *Ladena...* 'In it, the activities of Osama Bin Laden are described...').

Since proper names are mainly correctly tagged and easy to find in HrWaC, we have focused on complex names consisting of a common noun, which form a large number of proper names. According to Croatian orthography [4], uppercase letter comes in one-word names and in multiword names, where the first word is capitalized as well as the word that itself is a name and a possessive adjective derived from the

[1] hrWaC 2.2 is Croatian web corpus by Tomaž Erjavec and Nikola Ljubešić, crawled in 2011 and 2013, cleaned, deduplicated, tagged with the Croatian specification from MULTEXT-East v5, word sketches created by Nikola Ljubešić. More about HrWaC see in [1] and [2].

[2] [tag="Np.*"]

name. This is why we searched for capitalized common noun or capitalized word before the common noun. The beginning of the sentence is marked with capital letter, so we have excluded capitalized words after the full stop, the exclamation mark or the question mark. In Search 1 (S1), we looked for capitalized common noun with the help of regular expression [word!="\.|\?|\!"][word="[A-Z].*"&tag="Nc..."], and in Search 2 (S2), we looked for the sequence in which a capitalized attribute is on the first place and the common noun follows it ([word!="\.|\?|\!"][word="[A-Z].*"][tag="Nc..."])[3]. The possibility of one more word between capitalized word ([word!="\.|\?|\!"][word="[A-Z].*"][][tag="Nc..."]) gave us very general results, so we did not examine sequences with one or more words between the attribute and the noun any further. In each search, we have examined 1000 examples.[4] To avoid large number of examples from one source, examples were shuffled.

3 Results

Although we searched primarily for complex names that consist of capitalized common noun at the beginning of the name (S1) or of the attribute written in capital letter followed by the common noun (S2), our search (especially S1) gave us also some one-word names (e.g. *Primorje* 'part of Croatia by the coast). In numbers, we got 539 names and complex names in S1 and 369 complex names in S2. On the other hand, we obtained sequences with the uppercase letter that are not names, 461 of them in S1 and 631 in S2. It seems like a lot, so in 3.2 we presented the reasons for these results.

3.1 Names and complex names

Names and complex names obtained by the search can be divided into two major groups in accordance with rules for upper case writing in Croatian orthography.

Complex names with all capitalized words. One group consists of names whose all parts, according to Croatian orthography, should be written with capital letters except for prepositions and conjunctions. These are:

- personal names and names of cities and villages: although proper nouns are tagged in HrWaC and in search we looked for common nouns, we obtained results with proper names, probably because they were not tagged regularly (e.g. *Osijek* 'city in Croatia') or they coincide with some form of common noun, like in: *Posljednjih mjeseci, naime, **Maksim Mrvica** prakticki nema vremena za odmor.* 'In recent

[3] In quoting examples, the whole sentence with the exact result that corresponds to the regular expression is given. The exact result is marked with bold letters in Croatian, but it is not marked in English translation (provided in single quotation marks) since the examples are not translated word-by-word. Mark (S1) means that example is the result of the Search 1, and (S2) that the example is the result of the Search 2.

[4] In each search, a few sentences which made no sense to us were excluded manually, e.g. (S1) *...shvatio sam da se sve svodi na samo farmanje NPCa ili josh sladje ljudi...*'...I understood that everything comes to ?farmanje NPCa? or even better of people...'

months, Maksim Mrvica has practically no time to rest.' (*mrvica* means 'crumb', and *Mrvica* is a surname); *Dalnji plan je bio stići do Ploča*. 'The further plan was to reach Ploče.' (*ploče* means 'panels', and *Ploče* is a city)

- names of gods, their periphrastic names, and the names of other supreme religious persons: (S1) *Svaki oblik nijekanja života i samougušivanja strasti putem mučenja* ***tijela Poslanik*** *islama je osudio izrekama kao...* 'Any form of denial of life and self-denial of passion through torture of the body The Prophet of Islam condemned with the sayings like...'; (S1) *Što ti imaš sa mnom, Isuse****, Sine*** *Boga Svevišnjega?* 'What do you have to do with me, Jesus, the Son of the God Most High?'
- names of states: (S1) *Velika Britanija evakuirat će u srijedu u Ujedinjene* ***Arapske Emirate*** *svo osoblje veleposlanstva u Teheranu.* 'The United Kingdom will evacuate all embassy staff in Tehran on Wednesday into the United Arab Emirates.'

Names and complex names with capitalized word at the beginning. The other group obtained in our search are names and complex names in which only the first word is capitalized and also the word that itself is a name and a possessive adjective derived from the name. These are:

- geographic entities: (S2) *Što se događa* ***na Afričkome rogu****?* 'What is going on on the Horn of Africa?'
- institutions, organizations, associations, factories, state and public services, banks, libraries, faculties and colleges, schools, companies, and other objects and their parts: (S2) *Općinu Gornji Kneginec predstavila* ***je Turistička zajednica*** *Općine Gornji Kneginec...* 'Gornji Kneginec Municipality was presented by the Tourist Board of the Gornji Kneginec Municipality...'
- religious holidays, state holidays and memorials: (S2) *...organizirali su u nedjelju veliku gradsku biciklijadu u povodu* ***obilježavanja Dana grada*** *Zadra...* '...on Sunday, they organized a large city bike tour on the occasion of celebrating the Day of the city Zadar...'
- official texts, documents, laws, regulations, agreements: (S1) *...mora biti izravno propisano* ***u Zakonu*** *o medijima.* '...it must be directly specified in the Media law.'
- cultural, artistic, political, scientific and other social events, conferences, congresses, festivals, sports competitions...: (S2) *Napokon, Jug dobio Partizana u Beogradu te ušao u* ***finale Lige prvaka***. 'Finally, Jug beat Partizan in Belgrade and entered the Champions League final.'
- counties, administrative units: (S1) *... predstavljene su aktivnosti Krapinsko-zagorske županije* ***i Grada Zaboka*** *u mjesecu lipnju.* '...the activities of Krapina-Zagorje County and the City of Zabok were presented in June.'
- artistic, cultural and social groups: *...poslušnik kraljice pokušava Luciju uvjeriti u ljepote La La Landa, u kojem vječno* ***ratuju Alfe****, Bete i Game...* '...the queen's servant tries to convince Lucia in the beauty of La La Land, in which Alfas, Betas and Gamas are eternally at war...'

Some complex names are taken from other languages (mostly English): (S2) *Posljednja sesija skupa, nazvana "****Buffer zone*** *polities" II: Croatian Principality,*

sadržavala je četiri izlaganja. 'The last session of the conference, called "Buffer Zone polities" II: Croatian Principality, consisted of four expositions.'

Mistakes. The search also gave us some complex names where the target word is correctly written, but other words are not: (S2) *Kao **tajnik Hrvatskog Sabora** Kulture obišao sam sva mjesta u Hrvatskoj gdje se njeguje kulturni amaterizam...* 'As the secretary of the Croatian Parliament of Culture, I have visited all the places in Croatia where cultural amateurism is cultivated...' Correct writing would be *Hrvatski sabor kulture*.

3.2 Uppercase letter, no name

In S1 we got 461 results which do not match our intention to find names consisting of common noun, and in S2 we got 631 of such results. We established several reasons why our searches extracted sequences with uppercase letters which are not complex names.

Problems with interpunction. There are three typical situations that resulted in uppercase letter which is not a part of complex name.

Paragraph mark (¶). In many examples, paragraph mark is followed by a word with uppercase letter. It probably means that in the original source, before processing it for the corpus, the text was structured in two lines, and the paragraph mark is probably the sign of a new line and, in accordance with that, the sign of a new sentence (e.g. (S1) *Čuvati med od visoke temperature ¶ Mišljenje da kristalizirani med nije prirodan posve je pogrešno.* 'To keep honey from high temperature ¶ The opinion that crystallized honey is not natural, is completely wrong.').

Lack of interpunction. The search gave us results where there is no interpunction at the end of the sentence and before the uppercase letter, which is probably, like with paragraph mark, the result of two lines in original binding in one line in corpus: (S2) *Borac za **pravdu Uz** članove obitelji vijence na njegov grob položili su i saborska zastupnica...* 'A fighter for justice Along with family members, wreaths on his grave laid also a member of parliament...'

Another punctuation mark. In some examples, uppercase letter comes after punctuation mark other than the full stop, the exclamation mark or the question mark: (S2) *Razmislio je na trenutak i rekao: **- Zelim milijun** dolara svake godine do kraja moga zivota.* 'He thought for a moment and said: - I want a million dollars every year to the end of my life.' Unlike the full stop, the question mark or the exclamation mark, some other marks (like dashes, brackets, and colons) can be followed by uppercase or lowercase letter depending on the context, so this situation could not have been predicted and avoided with more precise search quest. Here we have also counted ordinal numbers, which in Croatian are marked with the full stop mark after numerals. If the

full stop is a part of ordinal number, as examples from HrWaC show, it is not treated as the full stop at the end of sentence (even when it is at the end). Cases like this should have been excluded because the regular expression said no full stop, but they still showed up as results of our search: (S1) *Članak **110. Izvjestitelji** sredstava javnog priopćavanja imaju pravo pratiti rad Vijeća...* 'Article 110th Public relations media reporters shall have the right to monitor the work of the Council...'

Text written in upper case. Sometimes, a whole part of the text can be written in upper case for stylistic reasons and this affects the results of the search, giving us examples which are not complex names like: (S2) *... sada zarađuju tu sumu u dolarima, eurima, funtama A I DUPLO SU MLAĐI TAKVI SU VANI **I NEMA TEORIJE** DA SE VRATE...* '... now they are earning that sum in dollars, euros, pounds AND THEY ARE TWO TIMES YOUNGER THEY ARE ABROAD AND THERE IS NO THEORY FOR THEM TO COME BACK...' Abbreviations are also written in upper case, and search extracts them as capitalized words: (S2) *Snimanje filmova **u HDTV** kvaliteti* 'Making films in HDTV quality'. Although some abbreviations are tagged as proper names (as pointed out in 2nd chapter), some of them showed up in our search for common nouns although, they should have been tagged as proper names: (S1) *... a da bi toplinska energija onda poskupjela za 25 posto, **prenosi HRT*** '... and then the heat energy would increase by 25 percent, reports HRT (Croatian Television)'.

Attribute + noun. Search 2 showed results in which obtained sequences are not complex names but a sequence of an attribute and a noun. Attribute can be; a) possessive adjective (which in Croatian is written with uppercase letter): (S2) *Božić je **blagdan Kristova rođenja***. 'Christmas is the feast of the Christ's birth'; b) a pronoun capitalized as a sign of respect: (S2) ***Čestitam Vašim čitateljima** i Vama na otvaranju ove teme.* 'I congratulate your readers and you on opening this theme'; c) a noun in function of an attribute (in Croatian, this is considered as a syntactic influence of English): (S2) *... najljepše od svega je da su sva **tri Ferrari motora** crkla u utrci.* 'The most beautiful of all is that all three Ferrari engines crashed in the race.'

Influence of foreign languages. Some sequences can not be treated as complex names since capitalized word in them is taken from a foreign language (mostly English): (S2) *Pa program je predvidio Park-Ride sustav…* 'Well, the program anticipated Park-Ride system...'

Mistakes. Some examples given by the search are capitalized and correctly extracted from HrWaC, but they are not in accordance with Croatian othography – they are the result of author's ignorance or, possibly, stylistic intention of giving the capitalized word a greater meaning: (S2) *...započinje suradnja **s Aikido klubom** u Mantovi.* '...the co-operation with the aikido club in Mantova begins'; (S1) *...lako bi se moglo dogoditi da Rusija postane posljednje utočište **Bijele Rase***. '...it may well be that Russia

becomes the last retreat of the White Race.' In Croatian orthography, there is no reason to write *white race* or *aikido* with uppercase letters.

4 Conclusion

In our research, we tried to extract complex names that consist of a common noun in Croatian corpus HrWaC. Since proper names form a specific group of words and they are mainly properly tagged in HrWaC, we have focused on common nouns, which can be used as regular nouns or they can be a part of complex names. Establishing the difference between complex names and the regular use of common nouns is important for lexicographic description. Names received in such a way could also serve as an additional extraction tool of entries for a *Dictionary of upper and lower initial case in Croatian, which* is being developed in the Institute of Croatian Language and Linguistics. In our search we got 539 names of 1000 examples in S1 (when we looked for capitalized common noun at the beginning of the name) and 369 of 1000 examples complex names in S2 (when we looked for sequence in which a capitalized attribute is on the first place and the common noun follows it). Some of the results are proper names, which should have been tagged as proper nouns and should not have come up as a result of our search for common nouns (probably they were not tagged regularly or they coincide with some form of common noun). We also obtained sequences with the uppercase letter that are not complex names, 461 of them in S1 and 631 in S2. One of the reasons for such a large number of "false" names can be linked to processing texts for corpus, namely inconsistencies of interpunction (paragraph mark, no interpunction...).

This search also emphasized the aspect of liability of sources in corpus. The texts collected for corpus come from public domain, they are often unedited and with many orthographic and grammatical errors. On the other hand, propositions for writing uppercase and lowercase letters in Croatian are very detailed and sometimes ask for extralinguistic knowledge. Besides, some rules have changed over the years (when parts of the earth, like *istok* 'East', refer to people and culture, they had to be written with uppercase letter – *Istok*, but according to new Croatian Orthography [4], they should be written in lowercase – *istok*). In order for texts from public domain to be a reliable source for linguistic purposes, their authors should be in step with orthographic changes and familiar with some general facts, which often is not the case.

Although it seems like search gave as many negative results, obtained names can contribute to the work on *Dictionary of upper and lower initial case in Croatian*, in checking the existing list and its enrichment, since some of obtained names were not previously included in it.

References

1. Ljubešić, N., Erjavec, T.: hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. Text, Speech and Dialogue 2011, 395–402, http://nlp.ffzg.hr/data/publications/nljubesi/ljubesic11-hrwac.pdf (2011).
2. Ljubešić, N., Klubička, F.: {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In: Bildhauer, F., Schäfer, R. (eds.) Proceedings of the 9th Web as Corpus Workshop (WaC-9), pp. 29–35. Association for Computational Linguistics, Gothenburg (2014).
3. Coates-Stephens, S.: The Analysis and Acquisition of Proper Names for the Understanding of Free Text, Computers and the Humanities, 26 (5/6), 441–456 (1992).
4. Jozić, Ž., Blagus Bartolec, G., Hudeček, L., Lewis, K., Mihaljević, M., Ramadanović, E., Birtić, M., Budja, J., Kovačević, B., Matas Ivanković, I., Milković, A., Miloš, I., Stojanov, T., Štrkalj Despot, K.: Hrvatski pravopis. Institut za hrvatski jezik i jezikoslovlje, Zagreb (2013).