Orthography in Practice: Corpus-based Verification of Writing Ktetics in MWUs in Croatian

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

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

Abstract. Writing upper and lower letter at the beginning of the word is one of the key orthographic problem in Croatian. The two main goals of this research are: 1 to determine whether there are frequent mistakes in writing upper and lower letter in ktetics (adjectives derived from geographic names) in corpus texts, 2 to use corpus tools (regular expressions) for orthographic analysis. According to [1: 112] upper and lower letter errors are defined as spelling errors and can be counted in one of the most common types of orthographic errors in Croatian. The emphasis will be on ktetics as components of various types of MWUs in which the orthographic rules prescribe a lower initial letter. Based on the corpus hrWaC 2.2 (using the Sketch Engine platform), the degree of deviation in the initial letter writing will be determined with regard to the meaning of the individual MWUs. The analysis includes four MWUs that denote different contents, in which the first component is ktetic; *zagrebačka katedrala* 'lit. Zagrebian cathedral, Eng. Zagreb cathedral' (sacral object), *osječko sveučilište* 'lit. Osijekian University, Eng. University of Osijek' (official institution), *sibirski haski* 'Siberian Husky' (dog breed), and *bečki odrezak* 'Wiener schnitzel' (types of steak). Given the results obtained, the conclusion will cover both: the types of corpus sources within which orthographic errors are most commonly as well as the possible causes of such errors. The research was made for the project *Rječnik velikoga i maloga početnog slova* 'Dictionary of upper and lower initial letter' which is being developed at the Institute of Croatian Language and Linguistics in Zagreb.

Keywords: Croatian, hrWaC, Lower Letter, Orthography, Upper Letter

1 Orthographic Rules and Errors in Corpus Texts

The rules of writing upper initial letter in Croatian, as well as in most other Slavic and European languages, primarily relate to two problems: writing upper initial letter at the beginning of the sentence and writing proper names - personal names and geographic names. These rules have been adopted by the speakers of the Croatian language already at the beginning of the schooling and are mainly used in written practice in the proper manner. In addition to these rules, Croatian orthography also includes other different rules that prescribe the writing upper or lower initial letter in

EUROPHRAS 2019, pages 46–52,
Malaga, Spain, September 25-27, 2019.
https://doi.org/10.26615/978-2-9701095-6-3_006

single words or MWUs depending on their sematic content - upper initial letter is used to describe the names of individual living beings, objects, historical events or geographic objects and areas, and words or MWUs with descriptive meaning in general use or denoting terms that belong to professional or scientific taxonomy are written in a lower initial letter. A large group of such MWUs are those in which the first component is ktetic - an adjective derived from a geographic place/object name (e.g., *zagrebački* 'Zagrebian, related to Zagreb', *pariški* 'Parisian, related to Paris', *švicarski* 'Swiss, related to Switzerland'). As the orthographic rules stipulates, ktetics at the beginning of multiword geographic names should be written in upper initial letter (*Zagrebačka gora* 'Zagrebian mountain' (a mountain north of Zagreb), *Pariška zavala* 'Parisian Basin', *Švicarska Konfederacija* 'Swiss Confederation'). At the beginning of multiword terms and in a general use, ktetics should be written in lower initial letter: *zagrebačka slavistička škola* 'lit. Zagrebian philological school, Eng. Zagreb philological school', *pariški odrezak* 'Steak Parisian', *švicarski franak* 'Swiss franc'). In written practice, however, these rules are often ignored. Wrong written records are common in writing of MWUs for which the orthographic rules prescribe a lower initial letter, but, besides the regular record, there are frequent records with a upper initial letter.

The corpus as a computer collection of written texts of different styles provides a good insight into the state of practice by showing how much the written record follows and how far it varies from the orthographic rules. According to classification made by [1: 112], upper and lower letter errors are type of spelling errors. In addition to spelling errors, punctuation, lexico-sematic errors, stylistic errors, typographical errors are also part of classification of errors mentioned in [1: 112]. All those types of errors are categorized as the errors made by humans and, as part of corpus texts, they can be detected in corpus search. Here, then, we start from the fact that in the case of orthographic errors made by a human, the corpus is not the source of these errors, but is only the platform on which those errors can be collected. In this context, the four limitations described by [2: 22–23] can be taken into account and which users should be aware of if they are using corpus results: "1 A corpus will not give information about whether something is possible or not, only whether it is frequent or not. (…) 2 A corpus can show nothing more than its own contents. (…) all attempts to draw generalizations from a corpus are in fact extrapolations. (…) 3 A corpus can offer evidence but cannot give information. (…) The corpus simply offers the researcher plenty of examples; only intuition can interpret them. 4 Perhaps most seriously a corpus presents language out of its context." Assertions described under 1 and 2 are considered as basic in this research.

2 Frequency Analysis: Ktetics in Corpus Texts

For the analysis, four examples of MWUs with ktetic as the first component were selected: *zagrebačka katedrala* 'lit. Zagrebian cathedral, Eng. Zagreb cathedral' (sa-

cral object), *osječko sveučilište* 'lit. Osijekian University, Eng. Osijek[1] University' (official institution), *sibirski haski* 'Siberian husky' (dog breed), and *bečki odrezak* 'Wiener schnitzel' (type of steak). This research was made for the project *Rječnik velikoga i maloga početnog slova* '*Dictionary of upper and lower initial letter*' which is being developed at the Institute of Croatian Language and Linguistics in Zagreb. Selected examples are prototypical in four thematic contexts of written practice where similar MWUs are frequent: administration, news, culinary (recipes and food), and pets. An objective evaluation of the use of orthographic rules in written practice would be obtained by analyzing a wider range of queries than these four. However, the primary intention here is to show how much corpus tools contribute to a specific search that includes the distinction of upper and lower initial letters. Generally, the use of corpus helps the work at the *Dictionary* for determining the degree of deviation from the rules in the writing of particular content that require upper or lower initial letter. According to given data, it is possible to explain the writing of such content in the dictionary more systematically and include as many such examples in the list of entries.

In Croatian, the way of writing of these multiword units can be subsumed under one of the four specific orthographic rules for writing initial lowercase depending on their meaning: 1 rule for writing sacral objects, 2 rule for writing colloquial names of official institutions, 3 rule for writing zoological or botanical species, and 4 rule for writing kinds of food. For this research, we used the *Croatian web corpus* – hrWaC 2.2[2] (on Sketch Engine platform) which, as interpreted in [3: 2], has the features of the general and reference corpus and represents language or variety as a whole (vs. specialized corpora). hrWaC was selected for this research as the biggest available corpus of Croatian [4: 30] containing more than 2 billion words.[3] This corpus contains different text genres (newspaper texts, administrative and legislative texts, blogs, forums) that are indicators of different styles and levels of use of the Croatian standard language. Because of a wider insight into the research problem, the whole corpus was searched, so it is not used the option Text types functionality. Using the hrWaC, only multiword units in non-initial position in a sentence were selected in order to avoid examples with upper initial letter at the beginning of the sentence. Such a search has included two basic regular expressions [5], depending on whether we searched for attestations with upper or lower initial letter (Table 1). Regular expressions have varied with respect to what we wanted to get for a particular multiword unit.

[1] The city in eastern Croatia.

[2] https://old.sketchengine.co.uk/corpus/first_form?corpname=preloaded/hrwac22_rft1; last accessed 2019/04/19.

[3] There are two other corpora for the Croatian language: *Hrvatski nacionalni korpus* / *Croatian National Corpus* (http://filip.ffzg.hr/cgi-bin/run.cgi/first_form) as a general corpus, and *Hrvatska jezična riznica / Croatian Language Repository* (http://riznica.ihjj.hr/) which is limited to newspaper and literary texts. These two corpora are also available on the Sketch Engine platform.

Table 1. Regular expressions for MWUs *ktetics + noun* with lower or upper letter form

Form of initial letter	Regular expression
Upper initial letter	[word!="\."][lemma="ktetic" & word="[A-Z]*"][lemma="noun" & word="[a-z]*"]
Lower initial letter	[word!="\."][lemma="ktetic" & word="[a-z]*"][lemma="noun" & word="[a-z]*"]

2.1 MWU *zagrebačka katedrala*

The orthographic rule [6: 38] stipulates that types of sacral objects (church, cathedral, mosque, synagogue, temple) should be written in a lower initial letter. According to this rule, it is correct to write *zagrebačka katedrala*. Attestations for forms *Zagrebačka katedrala* i *zagrebačka katedrala* were selected in hrWaC using regular expressions:

[word!="\."][lemma="zagrebački" & word="[A-Z].*"][lemma="katedrala" & word="[a-z].*"]

[word!="\."][lemma="zagrebački" & word="[a-z].*"][lemma="katedrala" & word="[a-z].*"].

Frequency results are presented in Table 2.

Table 2. Results for MWU *zagrebačka katedrala* in hrWaC

Form of writing	Frequency	%
Zagrebačka katedrala	441	20
zagrebačka katedrala	1791	80

2.2 MWU osječko sveučilište

Descriptive colloquial forms of public institutions which are not in official use according to orthographic rule should be written in lower initial letter [6: 30]. Since the official name of the university in Osijek is *Sveučilište Josipa Jurja Strossmayera u Osijeku* 'Josip Juraj Strossmayer University of Osijek', MWU *osječko sveučilište* is colloquial and unofficial form and should be written in a lower initial letter. Attestations for forms *Osječko sveučilište* i *osječko sveučilište* were selected in hrWaC using regular expressions:

[word!="\."][lemma="osječki" & word="[A-Z].*"][lemma="sveučilište" & word="[a-z].*"]

[word!="\."][lemma="osječki" & word="[a-z].*"][lemma="sveučilište" & word="[a-z].*"].

Frequency results are presented in Table 3.

Table 3. Results for MWU *osječko sveučilište* in hrWaC

Form of writing	Frequency	%
Osječko sveučilište	71	36
osječko sveučilište	127	64

2.3 MWU *sibirski haski* and *bečki odrezak*

The orthographic rule [6: 36] stipulates that zoological species and kinds of food should be written in a lower initial letter. According to this rule, proper written records are *sibirski haski* (for dog breed) and *bečki odrezak* (type of steak). Attestations for forms *sibirski haski*, *Sibirski haski*, *bečki odrezak*, and *Bečki odrezak* were selected in hrWaC using regular expressions:

[word!="\."][lemma="sibirski" & word="[A-Z].*"][lemma="haski" & word="[a-z].*"]

[word!="\."][lemma="sibirski" & word="[a-z].*"][lemma="haski" & word="[a-z].*"]

[word!="\."][lemma="bečki" & word="[A-Z].*"][lemma="odrezak" & word="[a-z].*"]

[word!="\."][lemma="bečki" & word="[a-z].*"][lemma="odrezak" & word="[a-z].*"]

Frequency results are presented in Table 3.

Table 3. Results for MWU *sibirski haski* and *bečki odrezak* in hrWaC

Form of writing	Frequency	%
Sibirski haski	18	27
sibirski haski	48	73
Bečki odrezak	12	10
bečki odrezak	111	90

3 Ktetics in Corpus Texts: Examples of Good or Bad Orthographic Practice?

Based on results obtained for all four MWUs it is evident that written attestations which confirm orthographic rules prevail in the corpus texts. Statistically, the correct writing of ktetics at the beginning of MWUs prevails in the range of 64% (*osječko sveučilište*) up to 90% (*bečki odrezak*). These frequency results confirm the relatively high level of correct writing of MWUs with the ktetics as first component. It is therefore possible to conclude the following:

1 corpus text writers know the orthographic rules that prescribe the writing of a upper or lower initial letter in the ktetics, but also other rules that include the contents of the MWUs included in this research, whether or not they contain ktetics. Speakers of the Croatian language in written practice often have doubts regarding the writing of sacral objects, especially buildings or places for different religious ceremonies, so it was expected that corpus texts would overwrite erroneous attestations, that is, with upper initial letter. The searching, however, confirmed that have attestations with the correct written records have higher frequency.[4]

2 It was also expected that the erroneous written records will prevail in writing colloquial forms of official institutions, because, in written practice, such forms are often taken as official although they are not. [5] Here, we can distinguish two types of corpus texts as the key reason why correct orthographic records are prevalent in the obtained results - newspaper texts (Croatian daily newspapers) and official sites of various public institutions and religious communities. In general, in such texts both the higher level of spelling knowledge and the higher level of language culture is represented because it takes into account that the texts are aligned with the Croatian standard language. As the second reason, it is possible to point out the fact that interest in orthography has increased in Croatia in recent years because, apart from the printed orthographic manuals, there are numerous public available free internet pages with orthographic and language advice topics (e.g. http: // pravopis.hr/, http://bolje.hr/, http://jezicni-savjetnik.hr/, http://hjp.znanje.hr/). In this way, one can quickly and easily reach the appropriate information on how to properly write.

3 Although corpus research has established that attestations of good orthographic practice are statistically predominant, corpus texts also contain written records that do not comply with spelling rules. Such written records are, as the study has shown, in a statistically lower ratio, but are not negligible - in the MWUs involved in this research, they range from 10% to 36%. Here are some possible explanations as to why, apart from the correct written records, there are also records that deviate from the orthographic rules:1 when writing some contents, e.g. religious buildings, it is assumed that, if something is written with a upper initial letter, has a greater importance

---

[4] Lower initial letter also prevails for *varaždinska katedrala* (488 results) 'Varaždian cathedral, Eng. Varaždin cathedral' in relation to *Varaždinska katedrala* (93 results).

[5] Lower initial letter also prevails for *zadarsko sveučilište* (530 results) 'lit. Zadarian university, Eng. University of Zadar' and *riječko sveučilište* (191 results) 'lit. Rijekian university, Eng. University *of* Rijeka', over upper initial letter in Zadarsko sveučilište (443 results) and Riječko sveučilište (145). The results were reversed for the *zagrebačko sveučilište* (727 results) 'lit. Zagrebian university, Eng. University of Zagreb' i *splitsko sveučilište* (176 results) 'lit. Splitian university, Eng. University of Split', in relation to *Zagrebačko sveučilište* (1124 results) and *Splitsko sveučilište* (556 results).

in the individual minds of speakers, but actually is a deviation from orthographic rule, 2 when writing some contents, such as religious buildings or colloquial and unofficial names of official institutions, the speakers do not know or have not checked the official names of the institutions and write such names in the upper initial letter instead of lowercase letter, and 3 as a possible reason, the influence of orthographic rules from other languages (most commonly English and German) can be mentioned in writing some single words and MWUs. For this reason, some of the MWUs involved in this research (*bečki odrezak*, *sibirski haski*) sometimes are written in upper initial letter, which is contrary to the orthographic rules in Croatian for writing zoological species and types of dishes. For a stronger argumentation of this assertion, the parallel English and Croatian, and German and Croatian Corpora should certainly be included in the research, which was not carried out in our work.

4 Finally, writing upper and lower initial letter is an important orthographic problem in Croatian. Only four orthographically interesting MWUs were included as prototypes in this research. For more insight into the orthographic practice of the Croatian language based on the corpus, it is necessary to include much more MWUs. The corpus as a source of large collection of texts gives a wide insight into how orthographic rules for writing upper and lower initial letter are implemented in practice, i.e. in written use. Good corpus tools, such as regular expressions for upper and lower letter, facilitate the availability of the requested results, suggesting that the corpus-based approach becomes an integral part in orthographic researches.

References

1. Jakubíček, M., Bušta, J., Hlaváčková, D., Pala, K.: Classification of Errors in Text. In: RASLAN 2009: Recent Advances in Slavonic Natural Language Processing, pp 109–119. Masaryk University, Brno (2009).
2. Hunston, S.: Corpora in applied linguistics. Cambridge University Press, Cambridge (2002).
3. Nesselhauf, N.: Corpus Linguistics: A Practical Introduction, http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf, last accessed 2019/04.
4. Ljubešić, N., Klubička, F.: {bs,hr,sr}WaC –Web corpora of Bosnian, Croatian and Serbian. In: Proceedings of the 9th Web as Corpus Workshop (WaC-9), pp 29–35. Association for Computational Linguistics, Gothenburg (2014).
5. Regular expressions, https://www.sketchengine.eu/user-guide/user-manual/concordance-introduction/regular-expressions/, last accessed 2019/04.
6. Jozić, Ž.: Hrvatski pravopis. Institut za hrvatski jezik i jezikoslovlje, Zagreb (2013).