UDK 811.163.42'373.46:159.937.5
159.937.511.5:811.163.42
004:81'373.46
Izvorni znanstveni članak
Prihvaćeno za tisak: 17. svibnja 2022.
https://doi.org/10.22210/suvlin.2022.093.03

Kristina Kocijan
University of Zagreb, Faculty of Humanities and Social Sciences
krkocijan@ffzg.hr

How we color the world with words

This paper presents a computational approach to the automatic detection of language patterns, specifically those dealing with expressing colors in the Croatian language. It investigates different lexicalization patterns of color terms, mainly compounds and multiword units, in order to classify them and prepare them for usage in the design of an algorithm that will automatically recognize and annotate these expressions in Croatian text. The paper also presents a comparative analysis of different classes of color terms found in a corpus built from books intended for younger (CLC) and older (ALC) populations. Finally, the research data is presented through a dictionary of three types of color terms categorized as multiword expressions.

1. Introduction

Our world is made of colors: according to Brown and Lenneberg (1954) 7.5 million different colors but according to the RGB color model that is based on trichromatic color vision (red–green–blue) and widely used in web design, this number goes up to 16.777.216 possible colors, although the vast majority of them is known only by their RGB code. However, these numbers relate mostly to the digital world while in the world of nature, as Taylor pointed out "things are typically associated with quite narrow segments of the color continuum" (2003: 13). The importance of colors is seen in the fact that we find them in all sorts of texts, ranging from children's stories and novels to texts on art, psychology, instructional design (Žufić and Kalpić 2009), chemistry, physics, or medicine, throughout history and in different languages (cf. Steinvall 2002). It is important to learn all lexicalization patterns we use when expressing a color so that we can build adequate electronic dictionaries and electronic resources that can be used for any automatic text analysis concerned with colors. However, this problem remains in general lexicography as well, and

understanding a color shade has been proven difficult for both native and foreign speakers of Croatian (Jelaska 2021).

The paper is structured in the following manner: in section two, we will provide background on color–related research in the Croatian academic community, after which (section three), our research objectives and methods will be presented. Detailed analysis of different patterns important from the computational perspective will be given in section four and the results of our research in section five. We will conclude the paper with a dictionary of three types of color terms belonging to the category of multiword expressions.

2. Theoretical Background

This section provides an overview of color–related research in Croatia from the perspectives of lexicography, linguistics, teaching, translation studies, and analysis of literary works. We will try to show how related research from all these disciplines was utilized in designing an algorithm for automatic detection of colors in Croatian texts.

Stolac (1994) brings a semantic analysis of color names with an accent on older Croatian lexicography. She further discusses the relationship between denotative and connotative meanings of colors but also its cultural perspective. On the other hand, Štimac Ljubas (2013) gives more attention to the problem of color–naming schemas in more recent (and future) Croatian monolingual dictionaries. Besides the more modern dictionaries published from 2000 to 2013, the author has also used as her research corpus different written and spoken resources ranging from women's magazines, fashion catalogues, daily newspapers, and catalogues dealing with chemistry, pharmaceutics, and cosmetics available either in paper or digital form (internet). The wealth of examples provided in Štimac Ljubas (2013) has proved to be very valuable and has greatly informed the direction of this research. The same can be said for the paper by Brbora (2005), who has approached this subject from a more semantic point of view.

Jelaska and Cvikić (2005) approach the topic of colors from the perspective of lexicon extension based on word–formation, borrowings, and some semantic principles (origin of the color, visual perception, and individual reactions of the speaker). The classification of different colors with regard to word formation provided by these authors was very helpful in building morphological but also semantic grammars that augment the set of basic colors found in the NooJ electronic dictionary. It allowed us to expand the number of recognized forms without greatly affecting the number of basic colors in the dictionary itself. This is becoming more and more important, since the list of new colors is constantly being updated as is emphasized by the same authors (Jelaska and Cvikić 2005) but also, more recently in (Jelaska 2021) with some color names being used by larger communities and industries, and others with only single occurrences. The number of different shades of colors

that an eye can differentiate goes up to several millions, (Štimac Ljubas 2013) and there is an abundance of ways in which people describe them.

Bošnjak Botica and Olujić (2016) discuss the patterns of color–name formations in Croatian and Romanian, but also that of Carasova speech (characteristic of speakers in Karaševo (Romania), which has a population of Croatian speakers), showing both their similarities and differences (suffixation, noun word, Nominative–Genitive NP, adverbs) but stressing the difficulties concerned with the perception of a color. Almost a decade earlier, Gulešić–Machata and Machata (2007) discussed colors used in collocations from the perspective of translation (Croatian–Slovak), but also teaching/learning such collocations when their collocation structure is the same. According to their research, mainly eight out of eleven basic colors are used to build collocations (leaving out gray, pink, and orange) and only two non–basic colors (*zlatna* 'golden' and *modra* 'navy blue'). Bennett (1981) showed that colors found in idiomatic structures present a problem for translators in his research on the translation of idioms that include a color word. He only conducted an investigation of translations from English into French (but a thorough one), yet one can easily expect the same problems to occur in the other direction as well, as Bennett's (1981) narrow reverse inquiry suggests.

The most recent research was reported by Raffaelli et al. (2019), where lexicalization patterns used in Croatian color terms were discussed in line with that of Czech and Polish speakers, with all three languages belonging to the Slavic language family. Their research was a part of a larger cross–linguistic project known as the *Evolution of Semantic Systems* (EoSS) that included over 50 Indo–European languages and was conducted from 2011 to 2014. The compiled resources included 1680 full responses by 20 native speakers of Croatian (cf. Raffaelli 2017 for a more detailed analysis of Croatian color naming patterns in EoSS data) and Czech each, and 1764 responses by 21 native speakers of Polish. Raffaelli, Chromý and Kopecka were interested to learn about strategies used in naming colors, similarities and differences between conventionalized lexicalization patterns, and "the color terms speakers use in the partition of the color spectrum in the three languages" (Raffaelli et al. 2019: 271). Their data shows that Croatian has the least construction types (163) among the three languages included in the research, with simple adjectives as the prevailing type of use (38.2%), followed by the adverb adjective constructions (30.2%), derived adjectives (19.7%), compound adjectives (7.3%), noun genitive–noun (0.6%), and adverb compound adjectives (0.06%). At the same time, Katunar et al. (2019) report on lexicalization patterns of color terms, namely color for object conceptualization strategies, in three languages from different language families: Croatian, Turkish, and Arabic. They based their research on dictionary data focusing on derivatives, compounds, and multi–word units or more precisely for Croatian: 141 derivatives, 40 compounds, and 106 multi–word units.

3. Research

A dictionary of basic colors is easy and quick to build, since there are only a limited number of such colors. However, building a dictionary of all colors would have to include thousands of different color combinations, as proposed in Stolac (1994), or even millions, and we would still risk missing some. We therefore propose a different approach. We propose building morphological and syntactic grammars that use a dictionary of basic color terms as its foundation, which it will expand upon. The results of the proposed algorithms will be shown in the following sections.

3.1. Objectives

The main objectives of this study are:

(1) to investigate different lexicalization patterns of color terms, mainly compounds and multiword units, and to evaluate and expand on existing descriptions of patterns in order to classify them adequately from the computational perspective;
(2) to design an algorithm that will automatically recognize and annotate expressions denoting colors in Croatian texts;
(3) to test the objectives (1) and (2) by applying the designed algorithm on two different types of discourse.

3.2. Methods

After investigating published research on colors and learning about the extent of the paradigms used in producing color names, we have prepared three corpora that were used for testing the proposed algorithm, but also for the research objective (3) – i.e., to learn whether different color–naming constructions are domain–related.

In the first phase of the project, a learning corpus (around 3,000 tokens – i.e., specialized samples of expressions of colors) was prepared from the examples found in the references (included in this paper) and used to design an algorithm for the detection and annotation of different types of color usage. In the second phase, the algorithm was tested on two additional corpora. The first corpus is made of books marked as **children's literature** (referred to in further text as the CLC – *Children's Literature Corpus*): a selection of stories and poems for children in primary school and younger. The second corpus includes titles intended for **older audiences**: a selection of novels for high school students and older (referred to further in this text as the ALC – *Adult Literature Corpus*). This classification was made based on the reading lists for Croatian primary and high school students available on–line.[1]

1 https://lektire.skole.hr/popis–djela–osnovna–skola/ and https://lektire.skole.hr/popis–djela–srednja–skola/

All the texts in the CLC and the ALC are available on–line and have been manually collected and prepared for this study. Since some books exist only as scans, conversion to NooJ text format introduced a few errors in sentence formatting, mostly dealing with wrong sentence breaks, wrong character conversion, or blurred text, most of which we managed to repair.

The CLC comprises 126 different titles and has a total of 2,542,667 tokens, while the ALC has only 30 titles, but almost matches in size the CLC, and contains 2,542,114 tokens. Despite the slight difference in size (553 tokens), both corpora are adequate for independent and comparative analysis – i.e., for objectives (2) and (3), respectively.

We have opted for the NooJ NLP tool (Silberztein 2016) that already has well–established resources for the Croatian language (Vučković et al. 2010; Vučković et al. 2011; Vučković et al. 2013; Kocijan and Librenjak 2016; Kocijan et al. 2016; Šojat et al. 2016; Kocijan et al. 2018; Kocijan and Librenjak 2018; Kocijan et al. 2020 inter alia). Results from this project are used to augment the existing NooJ dictionary with derivational paradigms of color terms, to build a grammar for the recognition of derived color terms at the morphological and syntactic levels, and finally, to test the algorithm and gain more insights from the results. During the linguistic analysis, NooJ applies its resources to a raw text in a cascaded manner, starting with dictionaries, morphological grammars and finally syntactic grammars. NooJ language resources for Croatian, augmented by this research, may be further used as a basis for all future computer–based research on colors found in texts from different domains. This includes text analysis from the perspective of a historical linguistics, translation studies, or comparison of usage between different writers or genre, to name but a few.

4. Analysis

In this section, in order to achieve our first objective, we will define and illustrate different lexicalization patterns of color terms, ranging from the simplest **basic color terms** like red, green, and blue to the more complex[2] color terms, as illustrated in example (1).

(1) *boja južnoameričkog žutozelenkastog dragulja*
the color of a South American yellow–greenish gemstone

4.1. Basic Color Terms

The significance of categorization of color terminology is defined by Taylor (2003). *From the computational perspective, we consider the term to be basic if there*

2 Depending on the research question, it is possible to define within NooJ if concordance should include the longest, the shortest or both matches. Thus, for the example in (1) it would be easy to find only the color term žutozelenkastog as well as the term *boja južnoameričkog žutozelenkastog dragulja,* or both.

is no other way (either at the morphological or syntactic level) we can produce it i.e. that computer can recognize it. There are eleven[3] such color terms in Croatian, and they are also considered to make up the list of basic colors (*crn* 'black', *bijel* 'white', *crven* 'red', *zelen* 'green', *žut* 'yellow', *plav* 'blue', *smeđ* 'brown', *ljubičast*[4] 'purple', *ružičast* 'pink', *narančast* 'orange', *siv* 'gray') as discussed in more detail in Čendo and Jelaska (2013). The same list of basic color terms can be found in other authors like Jelaska and Cvikić (2005), Katunar et al. (2019)[5], Benczes and Tóth–Czifra (2019), Ishikawa (2004) all the way back to Berlin and Kay[6] (1969). All of them are also entries in the NooJ dictionary of adjectives, where they are entered in the masculine singular nominative form for both indefinite and definite alternatives. Colors explained in this section will be marked as `<CT=1>`[7].

Another characteristic of these terms is that they contribute, some more than others, to the formation of more complex colors involving at least two different (basic) color names or *color terms*. Thus, some are used as modifiers (e.g., *bijela kava* 'white coffee', *crno vino* 'red [literally, black] wine') but also for producing complex color terms both with subordinate links (e.g., *plavoljubičasta* 'blue–violet') and without them (e.g., *žuto–zelena* 'yellow–green'). These terms tend to be most widely used in comparison to other lexicalization patterns of color terms. At this time, we have not considered their figurative meanings as described in Bošnjak et al. (2016), (e.g., *biti zelen* 'to be green: to be young'; ***bijela*** *udovica* '**white** widow', ***bijela*** *smrt* '**white** death', ***crveno*** *rođenje* '**red** birth', *u po* ***bijela*** *dana* 'in the middle of a **white** day').

After considering these, from the computational perspective extremely important, features – (a) *it cannot be produced from some other color* and (b) *it is used to produce other colors* – we have decided to enlarge the set of basic colors in the NooJ dictionary with additional colors detected in the learning corpus (as described in section 3.2.) – *antracit* 'anthracite', *bež* 'beige', *bordo* 'bordeaux', *braun* 'brown', *bronč*an 'bronze', *cinober* 'vermilion', *drap* 'beige', *grimizan* 'crimson', *indigo* 'indigo', *kaki* 'khaki', *krem* 'crème', *lila* 'lily', *modar* 'navy blue', *mračan* 'dark', *mrk* 'pitch black',

3 Compare to 13 basic color terms defined within the EOSS project for Croatian that in addition include *modra* 'dark blue' and *tirkizna* 'turquise' (Raffaelli 2017).

4 Raffaelli (2017) and Raffaelli et al. (2019) position *ljubičast* 'purple', *ružičast* 'pink', and *narančast* 'orange' among the complex terms. However, their derivation is too complex and thus too expensive from the computational perspective. It is for this reason that we find it justifiable to position them among the basic colors in the computational approach to the topic. Additionally, as reported in Benczes and Tóth–Czifra (2019), a number of languages interpret 'monolexemic' from Berlin and Kay's first criterion of what constitutes 'basic color term', as semantically simple and not morphologically simple. Such examples are Hungarian and Finnish terms for pink, *vaaleanpunainen* and *rózsaszín* respectively.

5 For Turkish but not for Croatian.

6 While additional evidence for universals in color naming across languages was undertaken by World Color Survey (Kay and Cook, 2015), for criticism on Berlin and Kay's definition of basic color terms see Crawford (1982), Taylor (2003), Wierzbicka (2005, 2008), and Benczes and Tóth–Czifra (2019).

7 Notation *CT* will be used in the NooJ dictionary and morphological and syntactic grammars as an attribute of a color term with values 1, 1.1, 1.2., 1.3., 2, 2.1, 2.1.1., 2.1.2., 2.1.3., 2.2., 2.2.1., 2.2.2., 2.2.3., 2.2.2.4, 3, 3.1., 3.2., 3.3., 3.3.1., 3.3.2, 3.3.3., 3.3.4., 3.3.5., 3.3.6., 3.3.7., 3.3.8. and 4 denoting the lexicalization pattern of a color term.

oker 'ocher', *pepeljast* 'ashy', *pink* 'pink', *purpuran* 'purple', *riđ* 'rufous', *roz* 'rose', *rumen* 'red', *sinj* 'gray–blue', *srebrn* 'silver', *šaren* 'multicolor', *tirkizan* 'turquoise', *zlatan* 'golden' – that will further allow us to construct more complex color names, such as *brončano–žuta* 'bronze–yellow' or *pepeljasto–siva* 'ash–gray'. Thus, we will be talking about two types of fundamental forms: the **basic** color terms (a list of 11 basic terms) and the **broad** color terms (a list of 136 terms[8]), neither of which can be produced from some other color term and all of which are used to produce other color terms. This classification is based purely on the computational requirements of some future projects dealing with color terms. Tags *+basic* and *+broad* that are added directly to the dictionary entries will aid such research by allowing for an easier distinction (inclusion/exclusion of specific sets) between these two lists.

The Croatian language is no stranger to borrowed terms. Evidence of this is found within the color terminology as well. Thus, terms originating in Latin, Greek, French, German, Italian, Spanish, English, and other languages are found in our test corpus, as many as 41 of them. The majority do not undergo declension as explained in more detail by Jelaska and Cvikić (2005) and Štimac (2019). As such, these adjectives have no description for flection in the NooJ dictionary (e.g., *ambra, antracit, bordo, drap, fuksija, indigo, krem, nugat, mauve, etc.*). However, we have marked them with two semantic tags [*+boja+posuđenica*], so we can clearly select them within the morphological and syntactic grammars. They are also on the list of broad color terms. Additional characteristic of these color terms is that they require the company of the word *boje* 'color' (gen. sing.) after them (e.g., *kuća kivi boje* 'a house of kiwi color / a kiwi–colored house') or another color (e.g., *kivi–zelena* 'kiwi green') and thus cannot be considered single–word color terms by themselves.

4.1.1. Suffixal Derivations of Basic Color Terms

Almost all basic colors have at least one type of derivation. Exceptions to this include two colors from the basic list of colors (*ljubičast*[9], *ružičast*), while the majority are from the broad list of colors (*antracit, bordo, brončan, drap, indigo, kaki, kestenjast, maslinast, mračan, mrk, pepeljast, pink, purpurni, riđi, rumeni, sinji, sur, tirkizan*). So far, there is no research–based evidence that these words can take a derivational suffix in order to describe a new shade of a color.

Those that can, in most cases use the suffixes ***–av, –ast,*** and ***–an*** (Jelaska and Cvikić 2005) with possible additional changes on the root (e.g., deleting the initial 'i' due to the reflex of yat), and adding additional[10] suffix '*–uš–*' before the final suffix *–av* like we observe in the color 'white': *bijel–>**bjeluš**av,* where we observe both

8 Since Croatian adjectives can have finite and infinite forms, some of these color terms are actually doubled, since both forms are represented in the main dictionary (e.g., *bijel* and *bijeli* 'white').

9 Babić (2002: 495) reports on finding only one realization of the term *ljubičastkast* and suggests using 'pale purple' or 'light purple' instead.

10 For more examples see Babić (2002: 495–497).

changes. In order to include all the alternatives within the same POS[11] category, a morphological grammar dealing with both definite and indefinite versions of derivations was designed (*e.g. bijeli–>bjelušavi*; *bijel –> bjelušav*).

There are 28 derivational paradigms that describe all of the indefinite adjective forms of both basic and broad list of colors, while 21 derivational paradigm describes all the definite adjective forms for colors. Derivations are provided for each main dictionary entry which allows us to make a connection between a derived word and its main form (e.g., *plav* and its derivations *plavkast, plavetan, plavičast, plavetnikast, plavčast*) and at the same time to avoid overgeneralization. Thus, at the same time, we are both keeping the full list of main dictionary entries unchanged and recognizing all the known derivations (including all their flective forms regarding gender, number, case, and comparison). The flective paradigm name is provided for each derivation name with the following syntax [+DRV=*derivationName:flection Name*] (see example 2[12]).

(2) plav,A+op+nodr+boja+basic+FLX=SVJEŽ +DRV=PLAVČAST:SVJEŽ
+DRV=PLAVETAN:DIVAN **+DRV=PLAVKAST:SVJEŽ**
+DRV=PLAVETNIKAST:SVJEŽ +DRV=PLAVIČAST:SVJEŽ

With a derivational rule we are transforming the main dictionary entry in order to produce the masculine singular nominative form of a new derived color. Thus, for example (2), by using the derivation **+DRV=PLAVKAST** on the main dictionary entry *plav*, we will produce the form *plavkast,* meaning approximative color blue or 'bluish' in English (more on approximative colors can be found in Štimac, 2019). This is possible via the instruction given in example (3) that virtually makes the entry *plavkast* by adding the suffix *–kast* to the end of the main form (*plav*) and in addition, annotates the new word as an Adjective [A] of descriptive type [+op], indefinite form [+nodr] and with a semantic tag for color [+boja].

(3) PLAVKAST = **kast**/A+op+nodr+boja;

The additional notation **:SVJEŽ** just after the paradigm name (example 2), will be in charge of making all the case, gender, number and comparison variations of newly produced form *plavkast*. Thus, our one–color entry *plav* (example 2) holds descriptions of 1 020 different forms of that color, including its derivations.

Some colors have more derivations than others, most of which (as many as 10) are found for the color *bijel* (*bijel –> bjelasast, bjelkast, bjeličast, bjelušast, bjeluškast, bjelušan, bjelušav, bjelusav, bjelucav, bjelasav*) and the least (only 1) for *cinober* (*cinober –> cinoberast*) and *lila* (*lila –> lilast*).

11 At this time, we did not deal with derivations concerning other POS categories, like nouns or verbs.

12 In the main dictionary, there are no spaces between different descriptions, and each dictionary entry is written as a single line entry. Here, however, we have introduced spaces between different derivation paradigms for better clarity.

A special type of derivation is the so–called elative, used to additionally stress a color. It is derived from a basic color with the suffix *–cat* (e.g., *plav plavcat, crn crncat, žut žutacat*) (Babić 2002: 499). Before we can recognize the two–word combination, we need to recognize the second word at the morphological level and tag it with a semantic attribute *+elativ*. We will use this tag at the syntactic level when looking for the pattern that has two colors next to each other but the second one is marked with *+elativ*. Additionally, both adjectives would have to have the same main form. There are also occurrences of using the same color twice for the same effect (e.g., *bijel bijel, crn crn*).

4.1.2. Prefixal Derivations of Basic Color Terms

Five prefixes have been reported to be used for the derivation of colors: ***o–***, ***na–***, ***pre–***, ***pro–***[13], ***polu–***. To recognize them, we have opted for a morphological type grammar where the algorithm is designed to recognize such cases (e.g., ***o****crvena* 'reddish', ***na****garava* 'sootyish', ***na***žuta 'yellowish', ***na****bijela* 'whiteish', ***pro****zelena* 'greenish', ***pro***žuta 'yellowish', ***polu****siva* 'semi–gray', ***polu****smeđa* 'semi–brown', ***polu****zlatna* 'semi–golden').

Such derivations seem to be rare since we have found only nine such instances in the ACL (***na***žutog 'yellowish', ***polu****mračan* 'semi–dark', ***polu****proziran* 'semi–transparent', ***polu****zelen* 'semi–green', ***polu****zlatan* 'semi–golden', ***polu***žut 'semi–yellow', ***pre****bijel* 'too white, super white', ***pre****blijed* 'super pale', ***pre****crn* 'super black') and four in the CLC (***polu****mračni* 'semi–dark', ***pre****crno* 'super black', ***pre****mračno* 'super dark', ***pre****zeleno* 'super green'), and none had the prefix ***o–***.

4.1.3. Suffixal Derivations of Non–Basic Color Terms

By adding the suffix *–ast* to a singular noun in Nominative (e.g., *limun* 'lemon'), we produce an adjective that may define the color of an entity. This type of a color term is also known as an elaborated color term (Bošnjak Botica and Olujić 2016).

So, for example, if we add *–ast* to the noun lemon (*limun*) we produce the adjective 'lemonish' (*limun****ast***). However, the problem is that although this adjective may denote something that has a color resembling the color of a lemon, it may also denote something resembling the taste or shape of a lemon. The same is true of the suffixes *–en* (*bakar+en–> bakren:* meaning both the color of copper but also some characteristics of copper), *–an* (*čokolada+an –> čokoladan:* color of chocolate or its taste), and *–av* (*čađ+av –> čađav:* the color of soot, or sooty consistency).

Mainly due to the ambiguity of these terms, we will not be dealing with them in any more detail at this time (*cf.* Štimac Ljubas 2013 for more examples and information on their origin).

13 The prefix ***pro–*** marks a property not fully recognized; thus, *prožuta* (pro+yellow) would specify a color not fully yellow.

4.2. Combining colors

In the previous section, we have described the single–color term occurrences. However, the Croatian language allows colors to be mixed in several language–oriented paradigms: as a single word <CT=2.1>, two–word combination <CT=2.2>, and two–word combination with a dash <CT=2.2.1>.

4.2.1. Complex compounds

Besides the basic color terms, another group of color terms are complex compounds that are also written as one word. These colors are made using the **modifier+basic color term** or **modifier+broad color term.** In the following subsections, we will describe different types of modifiers and how they connect to the main color term (a more extensive discussion about these can be found in Jelaska and Cvikić (2005) but also Štimac (2019)). According to the rules of Croatian orthography, depending on the effect we are trying to achieve, both combinations are possible: writing the modifier and a color as one or two words (with or without a dash '–') (see Raffaelli 2017 for more details). At this time, we have not conducted any research to check if corpus data supports this division. First, we will describe the single–word combinations.

4.2.1.1. ColorColor Combination

Single color terms that are found in the broad color list, including their derivations, are used to produce different color combinations. Thus, the color blue (*plav*) from our example (2) can be combined with green (*zelen*) to produce a new color, *plavožuta* 'blue–yellow'. In this combination, the first color used comes always in the form it has in the positive Nominative singular neutral gender form <A+np+Nom+s+n> while the second color changes regarding the comparation, case, number and gender.

This information allowed us to design a morphological grammar that can recognize such combinations of colors and at the same time attach to each all the properties of a second color (Figure 1). Still, since there are colors in the main dictionary that are non–declinable[14] and thus have no flective forms (e.g., *oker, kaki, krem, lila, pink, indigo, drap*), the main graph needed to be enhanced with an additional path (Figure 1 section B) including such cases where the first color is a no–flection color [–FLX].

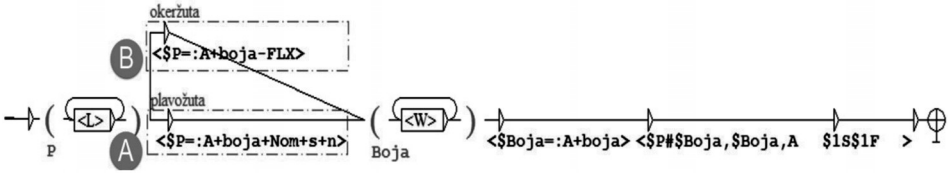

Figure 1. Section of a morphological grammar that recognizes single–word combinations of 2 colors

14 More on non–declinable adjectives can be found in Pudić (2015).

The algorithm takes a string of characters <L> and stores them in a variable $P. It then checks whether that string of letters exists in the dictionary as an adjective in Nominative singular neutral form with a semantic tag for a color [+boja]. If this agreement is true, it moves on to the remaining string of letters and stores them in a variable $Boja. The second variable is also checked against the dictionary entries, but this time it only checks whether the word is marked as an adjective with a semantic marker for a color. If the second agreement is also true, the algorithm recognizes the full string and tags it as an adjective that inherits all the semantic and flective properties of a second recognized string [$1S$1F].

Introducing this morphological grammar (Figure 1) allows us to expand the number of recognized colors while keeping the same number of entries in the dictionary. Some of the color combinations recognized with it are: *crvenoplava* 'red–blue', *modrozelena* 'blue–green', *roskastocrvena* 'rosey–red'. Two colors written as a single word are considered to be in a subordinate relationship, and they tend to describe one color that has a little bit of both colors. However, if we are describing a two–color concept, where both colors are clearly visible, we tend to write them with a dash, as will be shown in section 4.2.2.1.

4.2.1.2. AdverbColor Combination

Using an adverb to additionally narrow the shade of the color, found in the second position of this complex term, is not unusual in Croatian. Our results show that it is the most used sub–category among complex compounds, in both the ALC and the CLC.

Next to the most commonly used 'light' and 'dark' intensifiers (e.g., ***svijetlo***žuta, ***tamno***žuta), we find related adverbs referring either to brightness or saturation like *zagasito–* 'dull', *jarko–* 'bright', *nježno–* 'soft', *napadno–* 'striking', *sjajno–* 'shiny'. In some instances, the adverb may even relate to an inanimate object such as *čelično–* 'steel–', *baršunasto–* 'velvet–', *olovno–* 'lead–', *krvavo–* 'blood(y)', *smaragdno–* 'emerald–', *nebesko–* 'sky–'. Evidence shows that inversion is also possible; thus, the color term is found in the leading position, as in *sivokestenjast* 'gray–chestnutty'.

Morphological grammar (Figure 1) is thus augmented with an additional path that checks if the first part of the word exists in the main dictionary with an annotation for an adverb, while the second part of the world is any color from the broad color list. If both constraints are satisfied, the term is recognized and annotated so it inherits all the semantic and flective properties of a second part of the word, i.e., the color term.

4.2.1.3. NounColor Combination

Some single–word color terms are formed with a noun in the leading position, depicting usually a plant (e.g., ***limun***žuta 'lemon yellow') or an object (e.g., ***zlato***žut 'golden yellow'). Although they are described in literature, we have detected only

two in the ALC corpus – (***zlato****rumenim* 'golden ruddy' and ***zlato****brončani* 'golden bronze') and none in the CLC.

4.2.2. Two–Word Combinations

4.2.2.1. Combining a Color with another Color

If you are describing the colors of the national flags of Canada and Turkey, you would use two colors: white and red, which in Croatian can be written either as *crveno–bijela zastava* or *bijelo–crvena zastava* (red–white flag or white–red flag). This means that both colors are clearly visible as separate entities. The same rule applies as in the previous section: the first color is always in Nominative neuter singular form, while the second one changes – *e.g., crveno–bijeloj zastavi* (DAT SG DEF), *crveno–bijelim zastavama* (DAT PL), *crveno–bijelu zastavu* (DAT SG NDEF), etc. (cf. to [Adj [o] Adj] pattern in Raffaelli 2017). However, NooJ syntax recognizes the hyphen symbol as a word separator, which means that we have to move the word recognizer to the syntactic level.

Thus, we introduce the second grammar (Figure 2), which uses the same logic as the previous one, but with somewhat different syntax. The first node recognizes any adjective with a semantic tag for color provided that it is either a color term without flection `[A+boja–FLX]`, or a color term in nominative singular neutral form `[A+boja+Nom+s+n]`. The second node recognizes a dash that has no space before or after it `[#–#]`. The third node again recognizes any adjective with a semantic tag for color, but this time, the node is placed inside a variable `$B` so we can copy its gender, case, number and degree to the newly recognized complex color term.

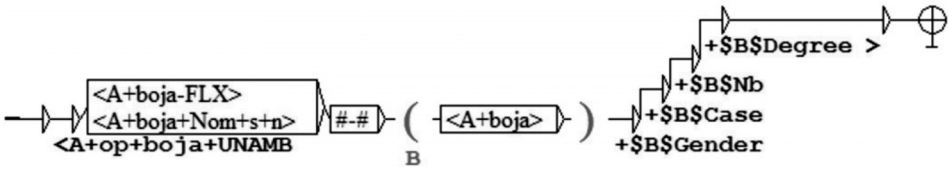

Figure 2. Syntactic grammar that recognizes two word combinations connected with a dash '–'

Examples that the algorithm in Figure 2 recognizes are, among others, the colors on the Canadian and Turkey flags, which are now marked as a single–color reference.

4.2.2.2. Combining an Adverb with a Color

Besides adjectiveColor–adjectiveColor combinations, there are color terms that have an adverb in the first position. These are the same patterns as found in single–word complex compounds (color type 2.1.2) but this time, the words are

written with a space between them (e.g., *tamno plava* 'dark blue', *svijetlo zelen* 'light green', *blijedo žuta* 'pale yellow', *jarko narančasta* 'bright orange', *lagano ružičasta* 'slightly pink') (cf. to [Adv Adj] pattern in Raffaelli 2017).

Since there is evidence of writing this word combination as a single– and a two–word term, both morphological and syntactic grammars (Figure 3) were augmented to accommodate this category.[15] The first node recognizes an adverb (a list of eligible adverbs is written above the note for illustration purposes) that is followed by the second node, in which an adjective with a semantic tag for color must be found `<A+boja>`. Again, the entire expression is marked as an adjective describing a color with gender, case, number, and degree data inherited from an adjective inside the variable `$B`.

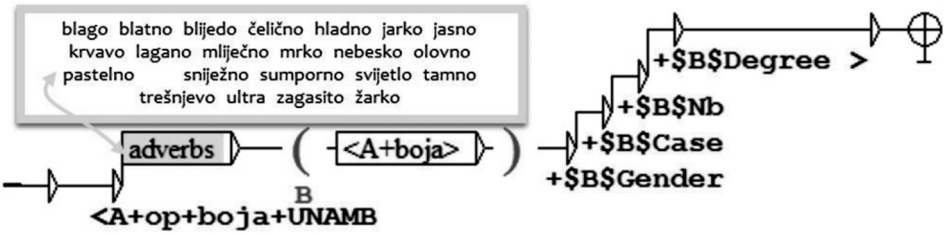

Figure 3. Syntactic grammar that recognizes Adverb–Adjective color terms

4.2.2.3. Combining an Adjective with a Color

A number of adjectives may precede the color term. In our test corpus there were examples with a descriptive adjective preceding the color (cf. Brbora 2005; Štimac 2019), as in examples of different shades for red – *crven* (4), yellow – *žut* (5), and green – *zelen* (6).

(4) *topla crvena* 'warm red', *divlje crvena* 'wild red', *plameno crvena* 'flame red'
(5) *jasno žuta* 'clear yellow', *nježna prozirnožuteljiva* 'gentle transparent yellowish'
(6) *otvorena zelena* 'open green', *zagasito zelena* 'extinguished green'.

There are also examples of a possessive adjective (see more in Štimac 2019) derived from the name of a person[16] (7) or place, like the examples of different shades of blue in (8).

(7) *šagalovski plava* 'Chagall's blue', *hukerovski zelena* 'Hooker's green', *bizmarkovski crvena* 'Bismarck's red', *brojgelovski smeđa* 'Brueghel's brown'
(8) *berlinski plava* 'Berlin blue', *mediteranski plava* 'Mediterranean blue', *pariško plava* 'Paris blue', *pruski plava* 'Prussian blue', *karipski plava* 'Caribbean blue'.

15 Figure 3 gives only a view of a syntactic grammar, since the morphological grammar follows the same logic but uses a different syntax.

16 Notice that Van Gogh's yellow has two variations: *vangogovski žuta* but also *Van Goghova žuta*

4.2.2.4. Combining Nouns with Adjectives

In some cases, a noun is used to paint the shade of a color, as in 'mustard yellow' (*senf žuta*). A noun and a color term are, in this case, written either as two separate words (with space between them) or they are connected with a dash, and we can describe this rule with a syntactic grammar, as shown in Figure 4.

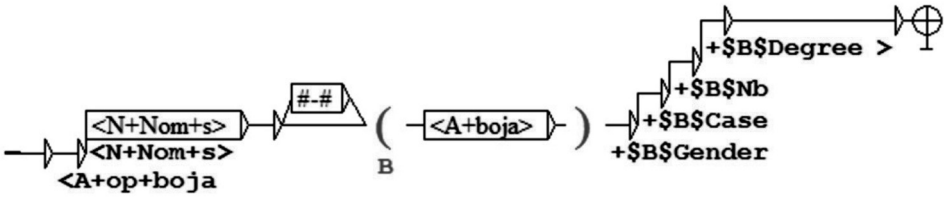

Figure 4. Syntactic grammar that recognizes Noun–Adjective color terms

In each of these cases, the noun does not inflect, and it remains in constant form as a singular noun in the nominative (gender is not considered here since it depends on the noun itself). Also, both common nouns (*limun žuta* 'lemon yellow', *jantar žuta* 'amber yellow', *senf žuta* 'mustard yellow', *smaragd zelena* 'emerald green', *kadmij žuta* 'cadmium yellow') and proper nouns (*Berlin plava* 'Berlin blue') are found as examples for this category. Literature gives evidence of writing such combinations as one word, as well (e.g., *limunžuta* 'lemon–yellow'), but such occurrences are described with a morphological grammar.

A special way of using nouns as color determiners is when we find combinations of two nouns where the natural color of the first noun translates to the second noun. In such cases, the first noun behaves like a non–flective adjective. For example, the expression *lavanda zid* 'lavender wall' means that the wall is not made of lavender but it is the color of that flower. Another example is a *ciklama shirt* 'cyclamen shirt' describing a shirt that is the color of the cyclamen flower. However, at this time, we will not be dealing with instances like these.

4.3. Colors as Multiword expressions

Sag et al. (2002: 3) define multiword expressions – MWEs – as "idiosyncratic interpretations that cross word boundaries (or spaces)". Such expressions are quite common and may make up even more than half of the speaker's lexicon making them very interesting for both linguistics and natural language processing (see Hüning and Schlücker 2015, Gantar *et al.* 2018). Thus, it is not surprising to find them in the terminology of colors as well. From the computational perspective, they are also quite problematic to deal with, since it is difficult to predict if the string of words should be observed as one semantic unit or not. Gantar *et al.* emphasize that MWEs "which language users consider to be individual lexical units based on their idiosyncratic meaning, but which basically follow regular language rules, can

represent a greater challenge for NLP than lexically and syntactically idiomatic combinations" (2018: 3). Furthermore, NooJ perceives a multiword expression as "potentially discontinuous sequence of word forms" (Silberztein 2016: 243) that are associated with relevant linguistic information and can be described either via dictionaries or syntactic grammars, but mostly with both.

The following types of color expressions are by nature multiword expressions. They either start with the word 'color' (*boja*) followed by a noun phrase, or they end with a noun 'color' (*boja*), or they start with a color term followed by a comparison word such as, among others, 'as' or 'like' (*kao* or *poput*) followed by a noun phrase.

4.3.1. *Boja* and NP

In the history of the Croatian language, the Croatian word for 'color' – *mast* – had been replaced with the word *boja* (of Turkish origin) in more recent language usage, although the word *farba* (of German origin) is even today also used in colloquial speech in some regions of Croatia (Jelaska and Cvikić 2005; Štimac Ljubas 2013). There are more terms related to the word 'color' (e.g., *kolor*, *kolur*), but also words that are derived from them (e.g., *kolorist* 'colorist', *kolorit* 'coloring', *kolorirati* 'to color') or are in some way related to them (e.g., *ličiti* 'to paint [a room or house]', *ličenje* '[house/interior] painting') and are in more details explained in Štimac Ljubas (2013).

Our main interest is colors that start with the word *boja* followed by a possessive genitive noun phrase – i.e., a noun phrase in the genitive case (see more in Štimac Ljubas 2013, Katunar *et al*. 2019; cf. to [N + Ngen] pattern in Raffaelli 2017). Such examples include colors like *boja mesa* 'the color of meat', *boja leda* '…of ice', *boja bijele kave* '…of white coffee', *boja lješnjaka* '…of hazelnuts', *boja lososa* '…of salmon', *boja jantara* '…of amber', *boja cigle* '…of brick', *boja senfa* '…of mustard', *boja jorgovana* '…of lilacs', *boja užarenog pijeska* '…of hot sand'. Naturally, we have chosen a syntactic grammar in NooJ to describe such occurrences (Figure 5).

Figure 5. Syntactic grammar for recognizing 'boja' followed by an NP pattern

Also, in literature we find slight variation of this rule, where a noun in the nominative precedes the noun *boja*, but in this case, they are connected with a dash (e.g., *lješnjak–boja* 'hazelnut–color') or the noun *boja* is preceded by a possessive adjective (e.g., *lješnjakova boja* 'hazelnut's color'). These two examples have the same meaning as the expression *boja lješnjaka*, and the difference is present only at the level of color term formation.

4.3.2. Adjectives and *Boja*

In some occurrences, the word *boja* is found at the end of a string that starts with an adjective mostly expressing some visual aspect (*kromatske* 'chromatic', *hladne* 'cold', *tople* 'warm') or psychological state (*vesele* 'cheerful', *tmurne* 'gloomy', *smirujuće* 'calming'), or even some auditive (*vrišteće* 'screaming') or tactile (*zemljane* 'earthy', *pješčane* 'sandy') quality (Brbora 2005). These adjectives are marked in the NooJ dictionary as descriptive adjectives `[A+op]`. However, some possessive adjectives `[A+po]` can also be found in the same position (e.g., 'the rainbow's colors' – *dugine boje*, 'rust's color' – *rđina boja*).

4.3.3. Comparisons

Any color term can be additionally compared to something in order to amplify its nuance. So, if you say that someone's shirt is red, that red can be any shade, from very light to very dark. But if you say that the shirt is "red as blood," you would clarify more precisely the type of red the shirt is.

As we have observed from our learning corpus (described in Section 3.2.), there are several ways we can express the shade of a color by comparing it to a specific plant (e.g., *zelen kao* ***trava*** 'green as grass'), animal (e.g., *crn kao* ***gavran*** 'black as a raven'), or phenomenon (e.g., *bijel kao* ***snijeg*** 'white as snow'). These phrases of comparison – i.e., phrasemes – also differ in the pattern in which they are expressed (see Štimac, 2013 for more detail). We have found eight such lexicalization patterns, which are visible in Fig. 6 (each lexicalization pattern is described with one path marked with a letter from A through H). We will describe them briefly in the following subsections.

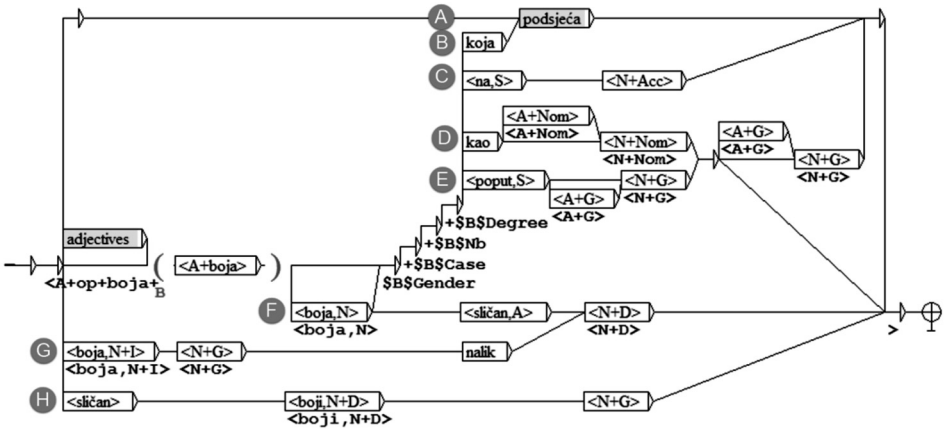

Figure 6. Syntactic grammar for colors using comparison

4.3.3.1. Suggesting a Color – *PODSJEĆA*

Pattern A (Figure 6) recognizes examples such as (9) to (11), in which the noun *bojom* is optional. Interestingly, we have not found any occurrences of this lexicalization pattern in either of the corpora.

(9) {bojom} podsjeća na rozu [podsjeća na A+boja]
'{in color} reminiscent of a rose'
(10) *{bojom} podsjeća na boju breskvina ploda* [podsjeća na boju NP+Genitive]
'{in color} reminiscent of the color of a peach'
(11) *{bojom} podsjeća na zrelu breskvu* [podsjeća na NP+Accusative]
'{in color} reminiscent of a ripe peach'.

4.3.3.2. A Color Suggesting Another Color

Pattern B (Figure 6) is a variation of pattern A in a way that recognizes expressions where another color is directly specified before it and connected with *koja* 'which, that' to the color it is compared to. Again, no such occurrences were detected in either of the corpora.

(12) *svjetloružičasta {boja}* ***koja*** *podsjeća na rozu*
'light pink that reminds one of a rose'
(13) *svjetloružičasta {boja}* ***koja*** *podsjeća na boju breskivna ploda*
'light pink that reminds one of the color of a peach'
(14) *svjetloružičasta {boja}* ***koja*** *podsjeća na zrelu breskvu*
'light pink that reminds one of a ripe peach'.

4.3.3.3. Color Term + *na* + NP

Although rare, there are occurrences when a color is compared to an entity using the preposition *na* followed by an NP in accusative, e.g. (15) and (16). This lexicalization pattern is described with pattern C in Figure 6.

(15) *zelena {boja}* ***na*** *bocu* [color term NA NP+Accusative]
'green as bottle'
(16) *zeleno* ***na*** *vodu* [color NA NP+Accusative]
'green as water'.

4.3.3.4. Color Term + *kao* + NP

Pattern D (Figure 6) describes the lexicalization patterns in which a shade of a color is compared to the color of another entity by using the word *kao* 'like, as'. In such cases, another NP in the nominative case follows, usually a single noun or an adjective and a noun (example 17), or an NP in the nominative followed by an NP in the genitive (example 18).

(17) *svjetloružičasta {boja}* ***kao*** *zrela breskva* [color term KAO NP+Nominative]
'light pink like a ripe peach'
(18) *svjetloružičasta {boja}* ***kao*** *stablo breskve* [color term KAO NP+Nominative + NP+Genitive]
'light pink like a peach tree'.

The word {*boja*} is optional in both examples (17 and 18); thus, for example, *plave boje kao more* ('blue color like the sea') or *plavo kao more* ('blue like the sea') would both be recognized.

4.3.3.5. The Color Term + *poput* + NP

In Croatian, you can use the comparison word *poput* followed by a genitive noun phrase (10). This pattern is described with path E in Figure 6. It is also possible to have another noun in genitive after the NP (11), like e.g., *roza poput trešnjinog stabla* ('pink like a cherry tree') where 'cherry' is used as an adjective, or *roza poput stabla trešnje* ('pink like the tree of a cherry'), where 'cherry' is used as a noun in the genitive case.

(19) *žuta {boja}* ***poput*** *šafranova cvijeta* [color POPUT NP+Genitive]
'yellow like a crocus's flower'
(20) *otvorena crvena {boja}* ***poput*** *cvijeta divljega maka* [color POPUT NP+Genitive + NP+Genitive]
'open red like the flower of a wild poppy'.

As in the previous lexicalization patterns, the word {*boja*} within the first part of the expression is optional.

4.3.3.6. Color Term + *nalik* + NP

Although rare, yet another possibility occurs in texts, when the word *nalik* is used to connect a shade of color that 'looks like' some object (pattern G in Figure 6). More precisely, the expression containing *nalik* has two sections: one before and one after it. The first part of the expression is the noun *boja* in its instrumental form followed by another noun in the genitive (denoting the entity whose color we are describing). The second part is the NP in the dative case, as for example in (21). The noun *boja* is not optional in this lexicalization pattern.

(21) *bojom kose* ***nalik*** *lisici* [boja NP+Genitive NALIK NP+Dative]
'with hair color like that of a fox'.

4.3.3.7. Color Term + *sličan* + NP

This lexicalization pattern includes the word for 'similar to' – *sličan* – that is found in two patterns: F and H in Figure 6, describing examples (22) and (23), respectively.

(22) *roza boja slična breskvi* [color term SLIČAN NP+Dative]
'pink color similar to a peach'
(23) *slično boji breskvine* [SLIČAN boji NP+Genitive]
'similar to the color of a peach'.

4.3.3.8. Color Term + *baca na* + NP

The last lexicalization pattern within the subcategory 3.3. uses the expression *koji baca na* in the sense of 'looks like' with the same syntax as sub–subcategories 3.3.1. and 3.3.2. (i.e., patterns A and B in Figure 6). Example (24) is found in our learning corpus, while the test corpus shows no evidence of this pattern.

(24) *bijela* ***koja baca na*** *lila* [color KOJA BACA NA color]
'white that looks like lily'.

4.4. Shades of color

The last category of color terms belongs to patterns that explicitly describe the shades of a color. Lexicalization patterns 4.a and 4.b describe patterns in which the expression starts with the noun *nijansa* 'nuance' and in which the same noun is found later on in the expression. Examples recognized with the algorithm are given in (25) a–d for the 4.a. lexicalization pattern and (26) a–c for the 4.b. lexicalization pattern.

(25) *a. pastelna nijansa ljubičastoplave* 'pastel nuance of violet–blue'
b. nijansa zagasite ljubičaste boje sa sivom primjesom 'nuance of extinguished violet with gray admixture'
c. nijansa modrozelene boje što podsjeća na boju mora 'nuance of blue–green that reminds one of the color of the sea'
d. nijanse tirkizne boje bliže plavim tonovima 'nuance of turquoise closer to blue tones'
(26) *a. otvorena plava nijansa* 'open blue nuance'
b. plava sa zelenim tonovima 'blue with green tones'
c. plava prema zelenoj 'blue towards green'.

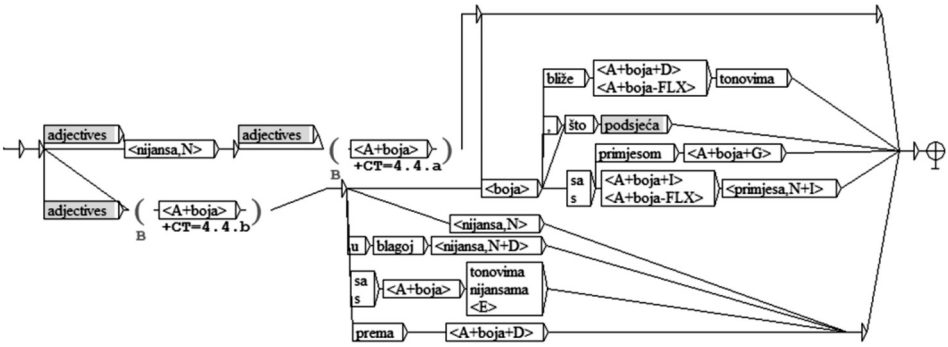

Figure 7. Syntactic grammar for colors explicitly describing shades

4.5 Other uses of Colors

Besides as adjectives, colors are found in some verbs (e.g., *zelen–> zelenjeti, **po**zelenjeti, **za**zelenjeti* 'turn green'; *plav –> plaviti, poplaviti, zaplaviti* 'turn blue'*; crven –> crvenjeti, zacrveniti, pocrveniti* 'turn red'; but *tirkizna* 'turquoise', for example, has no verbal forms connected with it), nouns (e.g., ***crveno*** 'red', ***plavo**oki* 'blue–eyed', ***žuto**kljunac* 'yellow–bill'), and multiword nouns (e.g., ***žuti** tisak* 'yellow journalism', ***crno** vino* 'black (i.e., red) wine', ***zeleni** čaj* 'green tea', ***ljubičasti** luk* 'purple onion', ***plava** riba* 'blue fish', ***smeđi*** šećer 'brown sugar', ***modre** usne* 'blue lips', ***bijela** tehnika* 'white goods', ***crvena** krvna zrnca* 'red blood cells', ***zlatna** kiša* 'golden rain') (cf. Katunar *et al.* 2019 for additional examples).

Colors are also used in phraseology (e.g., *gledati kroz **ružičaste** naočale* 'see through rosy glasses', *imati nešto **crno** na **bijelo*** 'have something in black and white', *princ na **bijelom** konju* 'prince on a white horse'). However, at this time we will not be dealing with any such phenomenon. Still, the colors as separate entities within the phrases will be recognized and marked as a color. These annotations will help us in some future projects to detect the full phrases.

Before we proceed with the presentation of our results, it is important to clarify that, although we have found evidence of such occurrences, at this time, we have not included **verbs** denoting a change in the color, e.g., to turn something white (*pobijeliti*) or blue (*poplaviti*). Also, the **personal names**, that is, the names of characters in the novels that are the same as the color term, e.g., *Bijela* (White), *Rumen* (Pink), *Riđi* (Red) and *Roza* (Rose), are also excluded from the study.

5. Results

Just a quick glance at Figure 8 gives us the impression that usage of color terms differs between authors writing for the younger or older readers. If this is something they are doing intentionally or not is not the subject of this research. We will

concentrate more on the numbers, starting from the wider picture down to the detailed analysis for each lexicalization pattern of color terms.

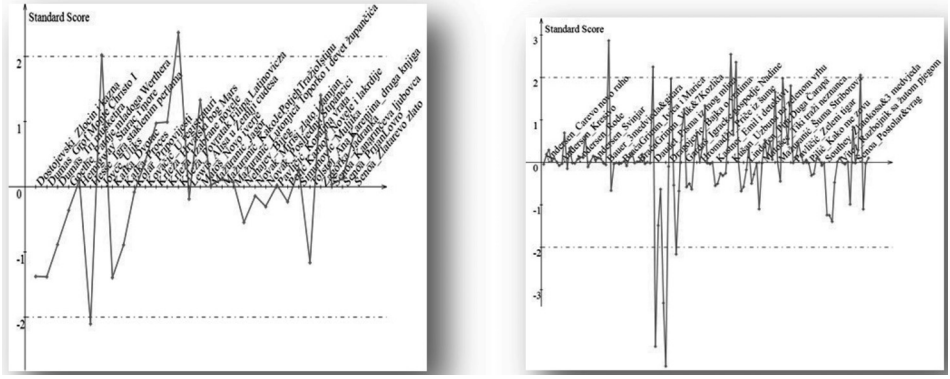

Figure 8. Usage of color terms across the ALC (left) and the CLC (right)

The left side of Figure 8 shows that some authors in the ALC corpus use color terms more than others, Krleža's *Povratak Filipa Latinovicza* [The return of Filip Latinovicz][17] dominating with a 2.37 standard score and Joyce's *Uliks* [Ulysses] with a standard score just barely over 2.02. Hesse's *Igra staklenim perlama* [The glass bead game] is just below the average score with the lowest standard score of –2.10. Other than these 3 books, all others are using colors within the –2 to +2 threshold.

On the other hand, children's books in the CLC corpus show a somewhat different pattern (right graph in Figure 8). Although the number of words per book is much smaller compared to the ACL books, the number of color terms is higher in the CLC corpus. Above the threshold are Baum's *Čarobnjak iz Oza* [The Wizard of Oz] (2.87), Daudet's *Pisma iz mog mlina* [Letters from my windmill] (2.25), Kušan's *Uzbuna na Zelenom vrhu* [The Mystery of Green Hill] (2.55), and Šenoa's *Čuvaj se senjske ruke* [Beware of the hands of Senj] (2.06). There are some children's books in the CLC corpus in which the percent of color terms is below the threshold: Defoe's *Robinson Crusoe* (–4.33), Dickens's *Oliver Twist* (–3.31), Dostojevski's *Zločin i kazna* [Crime and Punishment] (–4.80), and *Dnevnik Anne Frank* [The Diary of Anne Frank) (–2.16).

If we group the subtypes in 4 major sections – 1: basic CT, 2: combined colors, 3: MWEs and 4: shades – we learn that the ALC uses the color term types in the order 2, 3, 1, 4, and the CLC uses them in the order 1[18], 2, 3, 4, both descending from higher to lower usage (Figure 9). It is also quite clear that the difference is not only in the order but also in the quantity of used types in each section (for comparison

17 For a detailed analysis of color terms in this particular novel, see Čendo and Jelaska (2013).

18 Acquaviva et al. (2020) report that Croatian basic color terms are the most frequently used and thus considered to be highly conventionalized based on the frequency data obtained from a 1.72 billion–token Web corpus. This is in line only with our data for CLC.

to Croatian section of data collected within the EoSS project, cf. Raffaelli 2017). We will now look into each section separately to learn more about them.

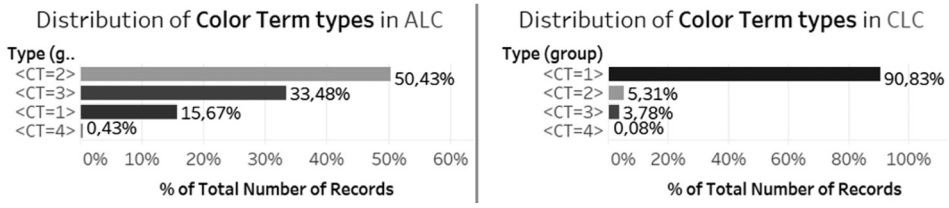

Figure 9. Distribution of color term types in the ALC (left) and the CLC (right)

From the results in Figure 10, we learn that our ALC corpus does not show evidence of all complex types of color terms. The missing types are 3.3.1, 3.3.2, 3.3.6, and 3.3.8. In the CLC corpus, the situation is similar, with missing examples for types 3.3.1, 3.3.2, 3.3.7, and 3.3.8 and with additional (*a)* realization of type 4. According to these results, the CLC corpus uses a larger variety of color terms but in smaller doses. Results for types 2.2.1 and 2.2.2 correspond to results obtained within EOSS data for [Adj [o] Adj] and [Adv Adj] patterns respectively (Raffaelli 2017). However, we have detected more occurrences of type 3.1 than it would be expected given the results reported for EOSS data where [N + Ngen] pattern was marked as 'borderline example of conventionalization' due to its low frequency and productivity (Raffaelli 2017: 183).

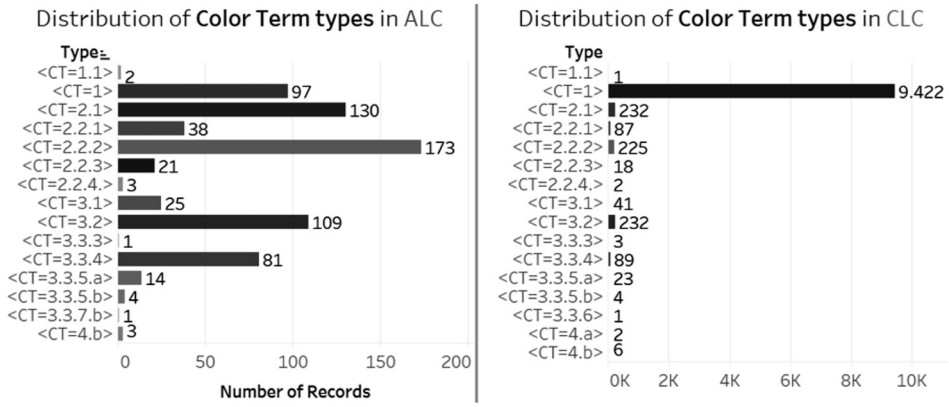

Figure 10. Distribution of color term types accross the ALC (left) and the CLC (right)

Complex compounds [CT=2.1]

Although the number of occurrences in category 2.1. is much greater in the ALC than in the CLC, the distribution of two subtypes (2.1.1. and 2.1.2.) is similar in two corpora (ALC: 43.70% and 56.30%; CLC: 39.55% and 60.45%). Only 19 color terms are found in both corpora (*blijedoružičast, blijedožut, jarkocrven, krvavocrven, sivozelen, smaragdnozelen, snježnobijel, svijetlomodar, svijetlozelen, svijetložut, tamnocrven, tamnoljubičast, tamnomodar, tamnosmeđ, tamnozelen, tamnožut, žućkastosiv, žutobijel* and *žutocrven*).

Others are corpus–specific, such as the following for ALC: *blatnosiv, blijedomodar, električnoplav, gnojnožut, gubavosiv, hrastovoblijed, jantarnozlatan, medenožut, musavosiv, narančastorumen, olujnomrk, otrovnožut, pamučnobijel, sumračnosiv, sunčanozlatan, zlatnoblijed*, and for CLC: *baršunastocrn, indijskožut*, *mrkozelenkast, nježnomodar, plamenocrven, riđekestenjast, smeđastozelenkast*, *sumpornožut, zagrebačkoplav, zelenkastoljubičast, zlatnoružičast*, among others.

Two–word combinations [CT=2.2]

Within the color terms built with two–word combinations (2.2), the order of preference is the same for both corpora, with adverbs in the first position dominantly leading in both (Figure 10).

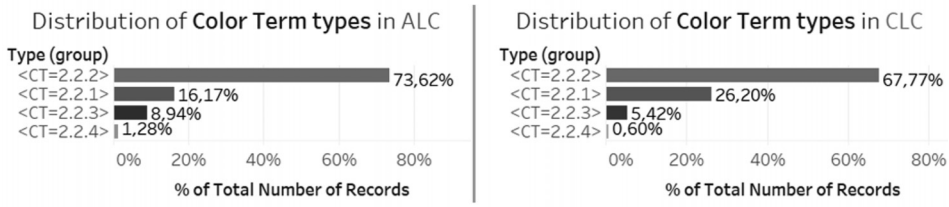

Figure 11. Distribution of two–word combinations in the ALC (left) and the CLC (right)

Multiword expressions [CT=3]

Category number 3 (see section 4.3) is the fourth most used category in the ALC and fifth in the CLC. It has 3 main subcategories, among which the first (3.1.) is used the least and the second (3.2) is used the most in both corpora. However, their distribution differs, mainly between the 3.2 and 3.3 groups (Figure 11). And while the usage of 3.2 and 3.3 is almost the same in the ALC, the CLC is dominated by category 3.2.

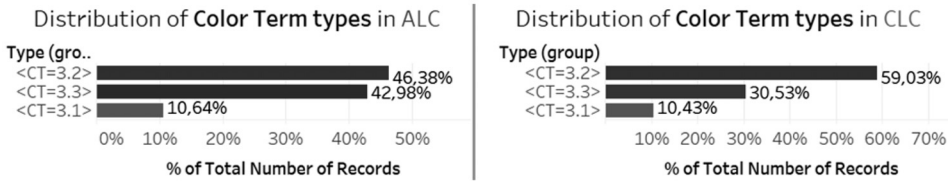

Figure 12. Distribution of different subtypes within the category 3 of color terms in the ALC and the CLC

According to Štimac Ljubas (2013), color terms produced with a possessive adjective followed by the noun *boja* (e.g., *smaragdna boja* 'emerald color') – type 3.2 in our classification – are less characteristic of the Croatian language[19] than color terms starting with the noun *boja* followed by a noun in the genitive case (e.g., *boja smaragda*) – type 3.1 in our classification. However, we have observed quite the opposite in the ALC corpus, with 3.2 occurring 4.36 times more often than type 3.1 color terms. The difference is even bigger in the CLC corpus, where type 3.2 was used 5.66 times more often. The full lists of color terms belonging to type 3.1 and 3.2 are provided in Tables 1 and 2 respectively, in the section *Dictionary of Color Terms*.

The most used color terms of type 3.1 are *boja kože* 'color of skin' and *boja rubina* 'color of ruby', both occurring four times in the CLC (highlighted in gray in Table 1). From the comparative analysis of detected terms belonging to this category, we learn that there are only three expressions used in both the ALC and the CLC (marked in bold face in Table 1): 'color of skin' – *boja kože,* 'color of the face' – *boja lica,* and 'color of the sea' – *boja mora*. The terms listed in Table 1 are in alphabetical order and are marked with the number of occurrences in each corpus separately.

A list of color terms of type 3.2 in Table 2 is sorted alphabetically within the color palette and it also provides the number of occurrences. The color term *bijela boja* 'white color' is the most used expression (detected 12 times in the CLC and 6 times in the ALC) and 35 expressions are used in both corpora. Most variations within the same color palette are found for the colors gray (*blijedosiva boja* 'pale–gray color', *kamenosiva boja* 'stone–gray color', *olovnosiva boja* 'lead–gray color', *olovna boja* 'leaden color', *siva boja* 'gray color', *sivkasta boja* 'grayish color', *žutosiva boja* 'yellow–gray color') and red (*crvena boja* 'red color', *crvenkasta boja* 'reddish color', *grimiznocrvena boja* 'crimson–red color', *jarkocrvena boja* 'bright red color', *svijetlocrvena boja* 'light–red color', *tamnocrvena boja* 'dark–red color', *žarkocrvena boja* 'glowing–red color'). There are also 110 variations of color terms that have not been marked for a color palette since no specific color is mentioned within the expression.

19 Štimac Ljubas also states that constructions produced with a possessive adjective followed by the noun *boja* were used more in the 2nd period of standard Croatian language history and she based her findings on the corpus built from the fashion magazines published from 1918 to 1941 (Štimac Ljubas 2013).

A list of the expressions from the last type in the third category is given in Table 3. The terms are given in alphabetical order within the color palette. The most used expression is *blijed kao krpa* (5 in the ALC and 8 in the CLC) followed by *bijel kao snijeg* (2 in the ALC and 7 in the CLC) and *blijed kao smrt* (4 in the ALC and 4 in the CLC). There are only 13 out of 182 expressions in this category that are found in both corpora. The top four color palettes with the most variations are red with 29 expressions, white with 25, and black and pale with 20 each. We have not detected evidence for the subcategories 3.3.1, 3.3.2 and 3.3.8. Within category 3, the most used pattern is the one using *kao* between the color term and its comparison, marked as 3.3.4 subgroup (Figure 12). These results are quite similar within the ALC and CLC, with the only difference in the list of sub–subcategories being in 1 occurrence of 3.3.7.b in the ALC and 1 occurrence of 3.3.6 in the CLC.

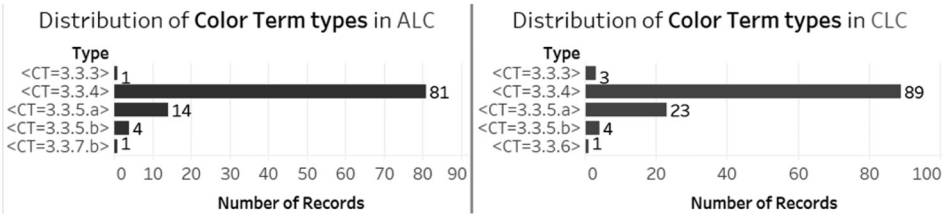

Figure 13. Distribution of different subtypes within the 3.3 subgroup of color terms in the ALC and the CLC

Shades of color [CT=4]

The last category has but 11 matches in both corpora: in ALC only 3 matching the 4.b pattern, and in CLC 6 matching the 4.b pattern and 2 matching the 4.a pattern.

Color Term vs Color Palette

We were interested to learn what the most used color term is, but also what the most used color palette is. What we consider to be the same palette is the main color in the term (we did not take into consideration color terms that do not include any color *per se,* like in the examples *boja užarenog pijeska* 'the color of glowing sand' <CT=3.1> or *vrišteće boje* 'screaming color' <CT=3.2>).

Figure 14 presents the top 16 listings of color terms (the two upper graphs) and of color palettes (the two lower graphs) in the ALC and the CLC, respectively. In both corpora, the first two positions are occupied by the same two colors, black and white (this is in line with the results obtained only for Krleža's novel *Povratak F.*

Latinovicza as reported in Čendo and Jelaska (2013)[20] and with findings reported in Katunar *et al.* 2019), while from that point onward, the colors have different orders depending on the corpus. If we consider only the frequency of basic colors, they have lined up in the following sequence in the ALC: black – white – red – green – yellow – blue – **grey**, and in the CLC: black – white – **green – red – blue – yellow – grey**.

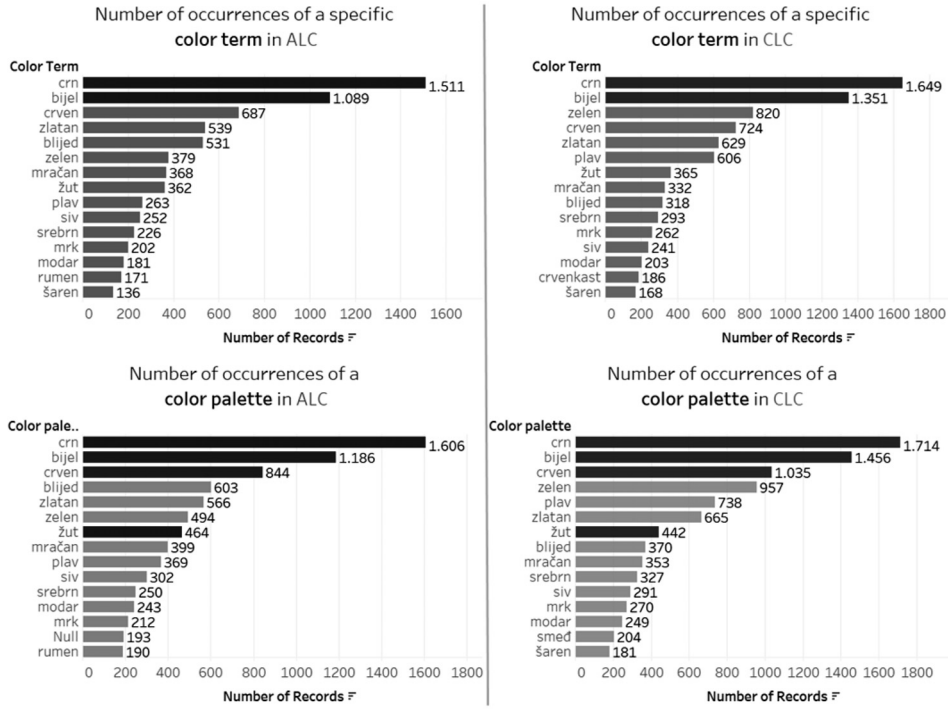

Figure 14. Top 15 color terms and color palettes in ALC and CLC

After merging the colors from the same palette, the first two positions remain unchanged, while the third one is occupied by the color 'red' in both corpora. The frequency of basic colors remains the same for the ALC and changes slightly in the CLC to: black – white – red – green – **blue – yellow** – **grey – brown**. If we compare these results to Berlin and Kay's (1969) evolutionary sequence (black – white – red – yellow | green – blue – brown – grey | orange | purple | pink), we see the correlation in the ALC (both in single and palette color frequency) for all the colors except for grey that occurs more often than brown in the ALC. The sequence is broken in several places in the CLC for single color frequency (green & red, blue & yellow, and grey & brown swapped positions), and in the CLC for palette color frequency (blue & yellow, and grey & brown swapped positions).

20 As a comparison, the same is observed in the LNC – *Lawrence Novels Corpus* (app. 1.1 million tokens) but not in the English Novels Corpus (app. 4.6 million tokens including children's stories and novels by T. Hardy, V. Woolf and J. Joyce) where white and red were the most frequently used color terms. Both corpora were compiled by Ishikawa (2004).

6. Conclusion

Our analysis shows that color terms exist in distinct lexicalization patterns, each of which requires a different computational approach. Detection of such lexicalization patterns fulfilled the first objective we have set for this project, and it was also used for the second objective – i.e., for the design of algorithms that recognize and annotate language units denoting colors, as both single and multiword expressions. We have used the power of a morphological transducer within the NooJ NLP environment to build an algorithm that identifies different lexicalization patterns of single words denoting colors. Using that model and the NooJ electronic dictionary, we have utilized the syntactic transducer to detect color terms as multiword expressions.

To test both algorithms and gain additional insights on the possibilities of color term recognition and annotation, we have prepared the corpus of adult (ALC) and children's (CLC) books. The results show that there is a difference in pattern usage between these two corpora with the CLC strongly dominating in basic, i.e., single–word, usage – almost 90% of all used color terms, while the ALC books use the two–word Adverb–ColorTerm combinations the most – almost 25% of the time. In our future work we will experiment with other domains (general language corpus but also specialized corpora such as medical and political) to see if the detected lexicalization patterns of color terms show similar patterns of usage.

Acknowledgements

Author thanks the anonymous reviewers for their critical reading of earlier versions of this paper and their many insightful comments and suggestions that helped improve and clarify this manuscript.

7. References

Acquaviva, Paolo, Alessandro Lenci, Carita Paradis, and Ida Raffaelli (2020). Models of lexical meaning. Pirrelli, Vito, Ingo Plag, and Wolfgang U. Dressler, eds. *Word Knowledge and Word Usage: A Cross–Disciplinary Guide to the Mental Lexicon*. Frankfurt: De Gruyter Mouton, 353–404, https://doi.org/10.1515/9783110440577–010

Babić, Stjepan (2002). *Tvorba riječi u hrvatskome književnome jeziku*. Treće, poboljšano izdanje. Hrvatska akademija znanosti i umjetnosti, Nakladni zavod Globus

Benczes, Réka, and Erzsébet Tóth–Czifra (2019). Rethinking the category of 'basic color term': Evidence from Hungarian lexicalization patterns. Raffaelli, Ida, Daniela Katunar, and Barbara Kerovec, eds. *Lexicalization patterns in color naming: a cross–linguistics perspective*. Amsterdam; Philadelphia: John Benjamins Publishing Company, 23–44

Bennett, Thomas J. A. (1981). Translating Colour Collocations. *Meta: Translators' Journal,* 26(3): 272–281, https://doi.org/10.7202/002057ar

Berlin, Brent, and Paul Kay (1969). *Basic Color Terms: Their Universality and Evolution*, University of California Press, Berkeley and Los Angeles, California

Bošnjak Botica, Tomislava, and Ivana Olujić (2016). O stvaranju naziva za boje u hrvatskom i rumunjskom jeziku te govoru Karaševa. *Romanoslavica* LII (2): 7–22

Brbora, Sonja (2005). Što je zajedničko marelici i lososu? (O nazivima za boje). Granić, Jagoda, ed. *Semantika prirodnog jezika i metajezik semantike*. Zagreb; Split: Hrvatsko društvo za primijenjenu lingvistiku: 111–121

Brown, Roger W., and Eric H. Lenneberg (1954). A study in language and cognition. *The Journal of Abnormal and Social Psychology* 49(3): 454–462, https://doi.org/10.1037/h0057814

Crawford, T. D. (1982). Defining „Basic Color Term". *Anthropological Linguistics* 24(3): 338–343, https://www.jstor.org/stable/30027848

Čendo, Kristina, and Zrinka Jelaska (2013). Paleta boja u romanu Povratak Filipa Latinovicza, *Croatica*, sv. 37: 213–253

Gantar, Polona, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso (2018). Multiword expressions: between lexicography and NLP. *International Journal of Lexicography* 32(2): 1–25, https:/doi.org/10.1093/ijl/ecy012

Gulešić–Machata, Milvia, and Martin Machata (2007). Boje u hrvatskim i slovačkim kolokacijama. *Riječ: časopis za slavensku filologiju* 13(2): 99–107

Hüning, Matthias, and Barbara Schlücker (2015) Multi–word expressions. Müller, Peter O., Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, eds. *Word–Formation*: *An International Handbook of the Languages of Europe*, Vol. 1. De Gruyter Mouton, 450–467, https://doi.org/10.1515/9783110246254–026

Ishikawa, Shin'ichiro (2004). A Corpus–Based Approach to Basic Colour Terms in the Novels of D.H. Lawrence. Nakamura, Junsaku, Nagayuki Inoue, and Tomoji Tabata, eds. *English Corpora under Japanese Eyes*. Leiden, The Netherlands: Brill, 185–212, https://doi.org/10.1163/9789004333758_011

Jelaska, Zrinka (2021). Boje na rubovima hrvatskog jezika. Bońkowski, Robert, Milica Lukić, Krešimir Mićanović, Paulina Pycia–Košćak, and Sanja Zubčić, eds. *Periferno u hrvatskom jeziku, kulturi i društvu. Peryferie w języku chorwackim, kulturze i społeczeństwie.* Katowice, Sosnowiec: Institut za slavenske filologije, Šlesko sveučilište u Katowicama, 46–66, https://doi.org/10.31261/PN.4038.05

Jelaska, Zrinka, and Lidija Cvikić (2005). The words for colors in Croatian: Different Means of Lexicon Extension. *Zbornik radova*. Sveučilište u Zadru, Stručni odjel za izobrazbu učitelja i odgojitelja predškolske djece, 5: 7–25

Katunar, Daniela, Barbara Kerovec, and Nawar Ghanim Murad (2019). From object to color and back: Seeing the world in color in Croatian, Turkish, and Arabic. Raffaelli, Ida, Daniela Katunar and Barbara Kerovec, eds. *Lexicalization patterns in color naming: a cross–linguistics perspective*. Amsterdam; Philadelphia: John Benjamins Publishing Company, 379–400

Kay, Paul, and Richard S. Cook (2015). World Color Survey. *Encyclopedia of Color Science and Technology*. Springer Science+Business Media, New York, https://doi.org/10.1007/978–3–642–27851–8_113–10

Kocijan, Kristina, Marijana Janjić, and Sara Librenjak (2016). Recognizing diminutive and augmentative Croatian nouns. Barone, Linda, Mario Monteleone, and Max Silberztein, eds. *Automatic processing of natural–language electronic texts with NooJ: revised selected papers*. Cham: Springer International Publishing, 23–36

Kocijan, Kristina, Silvia Kurolt, and Linda Mijić (2020). Building Croatian medical dictionary from medical corpus. *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 46(2): 765–782, https://doi.org/10.31724/rihjj.46.2.17

Kocijan, Kristina, and Sara Librenjak (2016). Comparative idioms in Croatian: MWU approach. Corpas Pastor, G. ed. *Computerised and corpus–based approaches to phraseology: monolingual and multilingual perspectivesI*. Geneva: Tradulex, 523–532

Kocijan, Kristina, and Sara Librenjak (2018). The quest for Croatian idioms as multiword units. Ruslan, Mitkov, Johanna Monti, Gloria Corpas Pastor, and Violeta Seretan, eds. *Multiword units in machine translation and translation technology*. Amsterdam: John Benjamins Publishing Company, 202–221, https://doi.org/10.1075/cilt.341.10koc

Kocijan, Kristina, Krešimir Šojat, and Dario Poljak (2018). Designing a Croatian aspectual derivatives dictionary: preliminary stages. Barreiro, Anabela, Kristina Kocijan, Peter Machonis, and Max Silberztein, eds. *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing (LR4NLP–2018)*. Stroudsburg, PA, SAD: Association for Computational Linguistics, 28–37

Pudić, Jelena (2015). *Otvorena pitanja pridjeva u hrvatskome standardnom jeziku*. MA paper at University of Pula, https://urn.nsk.hr/urn:nbn:hr:137:006571 (13.11.2020)

Raffaelli, Ida (2017). Conventionalized patterns of colour naming in Croatian. Cergol Kovačević, Kristina, and Udier, Sanda Lucija, eds. *Applied Linguistics Research and Methodology; Proceedings from the 2015 CALS conference*. Frankfurt am Main; Bern; Bruxelles: Peter Lang Verlag, 171–186

Raffaelli, Ida, Jan Chromý, and Anetta Kopecka (2019). Lexicalization patterns in color naming in Croatian, Czech, and Polish. Raffaelli, Ida, Daniela Katunar, and Barbara Kerovec, eds. *Lexicalization patterns in color naming: a cross–linguistics perspective*. Amsterdam; Philadelphia: John Benjamins Publishing Company, 269–286

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. (2002). Multiword expressions: A pain in the neck for NLP. Gelbukh A., ed. *Computational Linguistics and Intelligent Text Processing*. CICLing 2002. Lecture Notes in Computer Science, vol 2276. Springer, Berlin, Heidelberg, 1–15,

https://doi.org/10.1007/3–540–45715–1_1

Silberztein, Max (2016). *Formalizing Natural Languages: The NooJ Approach*. Cognitive science series, Wiley–ISTE, London, UK

Steinvall, Anders (2002). *English Colour Terms in Context*. PhD Thesis, Umeå Universitet.

http://www.diva–portal.org/smash/get/diva2:144764/FULLTEXT01.pdf

(12.11.2020.)

Stolac, Diana (1994). Boje u starijoj hrvatskoj frazeologiji i leksikografiji, *Filologija* 22–23: 259–267

Šojat, Krešimir, Božo Bekavac, and Kristina Kocijan (2016). Detection of verb frames with NooJ. Barone L., Mario Monteleone, and Max Silberztein, eds. *Automatic processing*

of natural–language electronic texts with NooJ: revised selected papers, Cham: Springer, 157–168

Štimac, Anamarija (2019). *Komparacija pridjeva u suvremenome hrvatskom jeziku*, MA Thesis, University of Zagreb, Faculty of Humanities and Social Sciences, https://bib.irb.hr/datoteka/1022591.Anamarija_Stimac_–_Diplomski_rad_–_2019.pdf (17.11.2020.)

Štimac Ljubas, Vlatka (2013). Podrijetlo, sintaktička struktura i leksikografska obradba naziva za boje, *Studia Lexicographica* 7 (1(12)): 91–115, https://hrcak.srce.hr/128748

Taylor, John R. (2003). *Linguistic Categorization*. Third Edition. Oxford University Press.

Vučković, Kristina, Marko Tadić, and Božo Bekavac (2010). Croatian Language Resources for NooJ. *CIT. Journal of computing and information technology* 18: 295–301, https://doi.org/10.2498/cit.1001914

Vučković, Kristina, Sara Librenjak, and Zdravko Dovedan (2011). Deriving Nouns from Numerals. Gavriilidou, Zoe, Elina Chatzipapa, Lena Papadopoulou, and Max Silberztein, eds. *Proceedings of the NooJ 2010 International Conference and Workshop*, Komotini, 84–95

Vučković, Kristina, Sara Librenjak, and Zdravko Dovedan Han (2013). Derivation of Adjectives from Proper Names. Donabédian, Anaïd, Victoria Khurshudian, and Max Silberztein, eds. *Formalising Natural Languages with NooJ*, Newcastle upon Tyne: Cambridge Scholars Publishing, 57–68

Wierzbicka, Anna (2005). There Are No "Color Universals" but There Are Universals of Visual Semantics. *Anthropological Linguistics* 47(2): 217–244, https://www.jstor.org/stable/25132327 (04.03.2022.)

Wierzbicka, Anna (2008). Why There Are No "Colour Universals" in Language and Thought. *The Journal of the Royal Anthropological Institute* 14(2): 407–425, https://www.jstor.org/stable/20203637 (04.03.2022.)

8. Dictionary of Color Terms

CT=3.1. Multiword expressions of type *boja* + NP

Table 1. List of color terms of the type 3.1 recognized in ALC and CLC and the number of occurrences

Color Term	Number of records	
	ALC	CLC
boja bjelokosti	1	
boja bliskog brijega		1
boja cimeta		1
boja čiste posteljine		1
boja dine		1

Color Term	Number of records	
	ALC	CLC
boja duge	1	
boja duše	1	
boja dvorske žalosti	1	
boja hladnog soka		1
boja jednog kraljevića		1
boja kestena		1
boja koralja	1	
boja kože	**2**	**4**
boja krvi	1	
boja lavande	1	
boja lica	**2**	**1**
boja lipova cvijeta		1
boja lješnjaka		1
boja ljudskoga glasa		1
boja meda	1	
boja mesa		1
boja mora	**1**	**1**
boja mrkve		1
boja opeke		1
boja plamena	1	
boja plijesni		1
boja pljesnivih jagoda	1	
boja pogleda	1	
boja predvečerja		1
boja prizme		1
boja purpura		1
boja pustinje		1
boja rđe		2
boja rijeke		1
boja rubina		4
boja safira	2	
boja senfa	2	

Color Term	Number of records	
	ALC	CLC
boja starog zlata		2
boja strasti	1	
boja sumraka	1	
boja svetog srca	1	
boja šarenih pločica		1
boja šljivova cvijeta	1	
boja tamne crvene ruže	1	
boja tamne opeke	1	
boja tapete		1
boja tinte	1	
boja večeri		1
boja vina		1
boja vinskog taloga		1
boja vještica		1
boja zemlje		1
boja zrela lipova cvijeta		1
boja živih ljudi	1	

CT=3.2. Multiword expressions of type Adjective + boja

Table 2. List of color terms of the type 3.2 recognized in ALC and CLC

Color Palette	Color Term	No. of records	
		ALC	CLC
	ažurna boja		1
	bjelokosne boje	1	
	bogata boja		1
	ciganskih boja	2	
	čarobna boja		1
	čokoladne boje	1	
	čudesna boja		1
	dekorativne boje	1	
	diskretne boje	1	
	divne boje		2

Color Palette	Color Term	No. of records	
		ALC	CLC
	djetinjska boja	1	
	drečave boje		1
	drugačije boje	1	
	druge boje	2	5
	drukčije boje		1
	duginih boja	5	
	fantastične boje	1	
	grofovskim bojama	1	
	gušće boje		1
	holandskih boja	1	
	idealnim bojama	1	
	istančane boje	1	
	iste boje	**10**	**2**
	ista boja kože		1
	izabrana boja		1
	jake boje pale	2	
	jakim bojama blizine	1	
	jarke boje		4
	jedina boja	1	
	jedna boja	**2**	**3**
	jorgovanove boje	**1**	**1**
	jutarnja boja		1
	kraljeve boje	1	
	krasne boje		1
	lijepu boju	**2**	**2**
	lješnjakove boje	2	
	međunarodne boje		1
	mirnu boju	1	
	mladićke boje	**1**	**1**
	moguće boje		1
	mrtvačku boju	1	
	mrtve boje	1	

Color Palette	Color Term	No. of records	
		ALC	CLC
	nadnaravnu boju	1	
	najljepše boje		1
	najljepšom bojom indiga	1	
	neizdiferencirane boje	1	
	nejasnu boju	1	
	neobične boje	2	
	neopisive boje	1	
	neosušene boje	**1**	**1**
	nezdrava boja		1
	nježne boje		1
	nova boja		2
	obična boja		2
	odvratna boja		1
	omiljena boja		1
	otmjena boja		1
	oštre boje	1	
	pastelne boje	1	
	plamene boje	2	
	pojedinih boja	1	
	pokisle boje	1	
	posebna boja		1
	potencijalna boja	1	
	prava boja	**2**	**2**
	predivna boja		2
	predvečernjih boja	1	
	primjetnim bojama	**1**	**1**
	prirodne boje	**1**	**2**
	proljetnim bojama	1	
	prvim bojama jeseni		1
	prvim bojama noći	1	
	raskošne boje		1
	ratnim bojama	1	
	različitih boja	**3**	**7**
	raznih boja	**2**	**2**

Color Palette	Color Term	No. of records	
		ALC	CLC
	sasušene boje		1
	sjajnih boja	**1**	**2**
	slabija boja	1	
	slične boje		1
	sportskim bojama	1	
	sretna boja	1	
	staru boju	1	
	stolarske boje		1
	strašne boje		1
	svake boje	**4**	**1**
	svakojake boje	1	
	svijetlu boju	**1**	**1**
	šafranove boje	1	
	šljivove boje	2	
	tamnom bojom	**2**	**1**
	toplih boja mora	1	
	treperave boje		1
	trešnjeve boje	1	
	uljena boja	**6**	**3**
	umjetnička boja	3	
	uobičajena boja		1
	upadljive boje	1	
	višnjeve boje	2	
	vodenim bojama	1	
	zagasita boja		1
	zamamna boja		1
	zaštitnoj boji	1	
	zatvorene boje	1	
	zemljana boja		1
	zimskih boja	1	
	žarka boja		4
	živih boja	**2**	**6**
	živih boja perja		1
	živopisnih boja		1

Color Palette	Color Term	No. of records	
		ALC	CLC
bakren	*bakrene boje*		2
bezbojan	*bezbojna boja*		1
bijel	***bijela boja***	**6**	**12**
	bjelkasta boja		1
	crvenobijela boja		1
	mliječnobijelom bojom	1	
blijed	***blijeda boja***	**1**	**3**
blistav	*blistava boja*		2
brončan	*brončana boja*		2
crn	***crna boja***	**5**	**5**
crven	***crvene boje***	**7**	**10**
	crvenkastu boju	**1**	**3**
	grimiznocrvenu boju		1
	jarkocrvenu boju		1
	svijetlocrvenu boju		1
	tamnocrvene boje	1	
	žarkocrvene boje	**1**	**3**
čađav	*čađavu boju*	1	
dugin	*dugine boje*		10
grimizan	***grimizne boje***	**1**	**3**
ljubičast	*blijedoljubičasta boja*	1	
	ljubičasta boja	**1**	**4**
	svijetloljubičasta boja	1	
maslinast	***maslinaste boje***	**2**	**1**
	maslinove boje	1	
modar	***modre boje***	**1**	**2**
mrk	*mrke boje*		1
narančast	***narančaste boje***	**2**	**2**
pepeljast	*pepeljaste boje*		1
plav	*električnoplave boje*	1	
	jarkoplavom bojom	1	
	plava boja	**3**	**14**
	plava boja potočnice		1
	svijetloplave boje	2	

Color Palette	Color Term	No. of records	
		ALC	CLC
rat	*ratne boje*		3
	ratničke boje		4
riđ	*riđe boje*		2
ružičast	*ružičaste boje*	2	
	blijedoružičaste boje		1
siv	*blijedosive boje*		1
	kamenosive boje	1	
	olovnosiva boja	1	
	olovnu boju	1	
	siva boja	**2**	**2**
	sivkasta boja		1
	žutosiva boja		1
smeđ	*smećkastu boju*	1	
	smeđe boje	3	
	svijetlosmeđe boje	1	
	žutosmeđe boje	1	
srebrn	***srebrne boje***	**2**	**2**
šaren	*šarene boje*		1
	šarolike boje	**2**	**1**
	šarolike boje utvara	1	
zelen	*fluorescentnozelene boje*		1
	sjajnozelena boja		1
	zelene boje	**2**	**8**
	zelenkasta boja		1
	zelena boja maštanja		1
	zlatnozelenkasta boja		1
zlatan	***zlatne boje***	**3**	**3**
žut	*blijedožuta boja*		2
	tamnožuta boja	**1**	**1**
	zlatnožuta boja		1
	žuta boja		4

CT=3.3. Multiword expressions of type *boja* + comparison

Table 3. List of color terms of the type 3.3 recognized in ALC and CLC

Color palette	Color term	CT= 3.3.3	CT= 3.3.4	CT= 3.3.5.a	CT= 3.3.5.b	CT= 3.3.6	CT= 3.3.7.b
	boje ko obrazi djevojčeta		1A				
	bojom kose nalik lisici					1C	
	višnjeve boje kao glavno odijelo		1A				
	slična boji podloge zrcala						1A
akvamarin	akvamarin kao packe		1A				
bijel	bijel kao alabaster		1A				
	bijel kao kost		1A				
	bijel kao latice zelenkade		1A				
	bijel kao mramor		1A				
	bijel kao najfiniji papir		1A				
	bijel kao papir		2C				
	bijel kao snijeg		**2A + 7C**				
	bijel kao svježi hljeb		1A				
	bijel kao vila		1C				
	bijel kao vosak		1A				
	bijel kao zid		2C				
	bijel kano tek		1A				
	bijel kano mramor		1A				
	bijel ko snijeg		1C				
	bijelo na jaja	1C					
	bijel poput bisera			2A			
	bijel poput bjelokost			1C			
	bijel poput slonove kosti			2A			
	bijel poput snijega			**1A + 2C**			
	bijel poput srebra			1A			
	bijel poput papira			1A			
	nježnobijel poput voska			1A			
	poput mlijeka bijela ramena				1A		
	poput mlijeka bijeli zubi				1A		
	poput zmija bijeli puteljci				1C		

Color palette	Color term	CT= 3.3.3	CT= 3.3.4	CT= 3.3.5.a	CT= 3.3.5.b	CT= 3.3.6	CT= 3.3.7.b
blijed	blijed kao anđeo		1A				
	blijed kao bizantinska ikona		1A				
	blijed kao dunja		1A				
	blijed kao gljiva		1A				
	blijed kao gnjilo meso		1A				
	blijed kao kamen		1C				
	blijed kao kreda		1C				
	blijed kao krpa		**5A + 8C**				
	b**lijed kao mramor**		**2A + 1C**				
	blijed kao mrtvac		1C				
	blijed kao mumija		1C				
	blijed kao osuđenik		1A				
	blijedo kao papir		1A				
	blijed kao sablast		1A				
	blijed kao smrt		**4A + 4C**				
	blijed kao snijeg		1A				
	blijed kao stijena		1C				
	blijed kao svjetlost		1C				
	blijed ko krpa/e		2C				
	blijed na smrt	**3A + 2C**					
blistav	blistav kao nebesko oko		1A				
	blistav kao rijeka		1C				
	blistav kao trezor		1A				
	blistav ko dan		1C				
crn	crn kao crnac		1A				
	crn kao ebanovina		3C				
	crn kao dim strijele		1A				
	crn kao gavran		1C				
	crn kao grob		1A				
	crn kao gromada crnog kamena		1C				
	crn kao jezera		1C				
	crn kao krtica		3A				
	crn kao kupina		1A				
	crn kao noć		2C				

Color palette	Color term	CT= 3.3.3	CT= 3.3.4	CT= 3.3.5.a	CT= 3.3.5.b	CT= 3.3.6	CT= 3.3.7.b
	crn kao okviri		1C				
	crn kao packa		1A				
	crn kao smrt		1A				
	crn kao ugljen		**1A + 1C**				
	crn kao vrana		1C				
	crn ko blato		1C				
	crni ko pustahije		1A				
	crn poput ebanovine			1A			
	crn poput vraga			1A			
	crni poput kurije đavla			1A			
crven	**crven kao bakar**		**1A + 1C**				
	crven kao božur		**1A + 1C**				
	crven kao crvena paprika		1C				
	crveni kao crvendaći		2A				
	crven kao dobro kuhani rak		1C				
	crven kao koralj		1A				
	crven kao krv		**1A + 1C**				
	crven kao krvavi most		1A				
	crven kao mak		1A				
	crven kao mrkva		**1A + 1C**				
	crven kao plamenovi		1C				
	crven kao prije		1C				
	crven kao purpur		1A				
	crven kao rak		3A				
	crven kao rđa		1C				
	crven kao rubin		1C				
	crven kao slova		1C				
	crven kao ujak		1A				
	crven kao zrele trešnje		1A				
	crven ko jagoda		1C				
	crven ko krv		1A				
	crven ko rak		1C				
	crven ko skrlet		1A				
	crven poput kosovskog božura			1A			

Color palette	Color term	CT= 3.3.3	CT= 3.3.4	CT= 3.3.5.a	CT= 3.3.5.b	CT= 3.3.6	CT= 3.3.7.b
	crven poput krvi			1A			
	crven poput makova cvijeta			1A			
	crven poput rđe			2C			
	crven poput rubina			1C			
	crven poput želea			1C			
grimizan	grimizno kao vatra		1A				
ljubičast	ljubičast kao jorgovan		1C				
modar	modar kao najdublje more		1C				
	modar kao patlidžan		1C				
	modar kao sumporni plamen		1C				
mračan	**mračan kao oblak**		**1A + 1C**				
mrk	mrk kano oblak		1A				
	mrk ko oblak		1A				
plav	ljubičastoplav kao jezero suza		1C				
	plav kao kučina		1A				
	plav kao more		1C				
	plav kao nebo		2C				
	plav kao potočnice		1C				
	plav kao proljetne vode		1A				
	plav kao proljetno nebo		1A				
	plav kao šljiva		1A				
	plavičast kao tamjanov kad		1A				
	plavetan kao nebo		1C				
	plav poput potočnica			2C			
	plav poput safira			1A			
	poput neba plavom				1C		
proziran	proziran kao čisti zrak		1C				
	proziran kao staklo		1C				
	prozirna poput ružine latice			1C			
	prozirna poput tarantule			1A			
rumen	lijepo rumen kao ružica		1A				
	rumen kao krv		1C				
	rumen kao krvavo srce		1A				

Color palette	Color term	CT= 3.3.3	CT= 3.3.4	CT= 3.3.5.a	CT= 3.3.5.b	CT= 3.3.6	CT= 3.3.7.b
	rumen kao ruža		1C				
	rumen kao sunce		1C				
	rumen kano božićnice		1A				
	rumen ko trešnja		1A				
	rumen poput mrkve			1C			
	rumen poput ružice			1C			
	rumen poput ružine latice			1C			
	poput ruže rumenu				1C		
ružičast	ružičast kao grančice mlade kajsije		1A				
	ružičast poput breskve			1C			
siv	siv kao krpa		1C				
	siv kao miš		1A				
	siv kao običan golub		1C				
	siv kao pepeo		1C				
	siv kao fotografski negativ		1A				
	siv poput kamena			1C			
	siv poput pješaka			1C			
smeđ	smeđu boju poput blata			1C			
srebren	srebrni kao ljiljan		1A				
	srebrnkaste boje kao podloga zrcala		1A				
	poput lepeze srebrno				1A		
sur	sur kao tvorić		1A				
zelen	zeleni kao zelembaći		2A				
	zelen kao blatna voda		1A				
	zelena kao tanana svila		1A				
	zelen poput mahovine			1C			
	zelen poput trave			1C			
zlatan	poput kapi zlatne svjetlosti				1A		
žut	blijedožuta kao mrtvac		1A				
	žut kao badem		1C				
	žut kao cekin		1A				
	žut kao dinja		1C				
	žut kao guščje salo		1A				

Color palette	Color term	CT= 3.3.3	CT= 3.3.4	CT= 3.3.5.a	CT= 3.3.5.b	CT= 3.3.6	CT= 3.3.7.b
	žut kao lišće		1A				
	žut kao nizovi biserja		1A				
	žut kao vosak		**1A + 1C**				
	žut kao voštanica		1A				
	žut kao zlato		1C				
	žut ko pražetina		1A				
	žut poput jaglaca			1C			

Kako bojimo svijet riječima

U radu je dan sveobuhvatan prikaz različitih obrazaca koji se koriste u terminologiji boja u hrvatskom jeziku i koji su do sada opisani kroz objavljena istraživanja u ovom području. U fokusu je prikaz iz računalnog pristupa automatskom otkrivanju leksičkih obrazaca. Svrha predstavljenog istraživanja je definirati postojeće modele za izgradnju izraza o boji u hrvatskom jeziku, s posebnim naglaskom na složenice i višerječne izraze te implementacija prepoznatih modela u računalnoj obradi jezika.

Analiza i definiranje različitih modela na osnovu postojeće literature za boje u hrvatskom jeziku imala je za cilj njihovu klasifikaciju i pripremu za uporabu u računalnoj obradi jezika. U ovoj su fazi definirana 4 osnovna uzorka sa svojim pod–klasama. Ovako definirani leksikalizirani obrasci korišteni su unutar NooJ alata za obradu jezika gdje su omogućili izradu (a) digitalnog rječnika s popisom osnovnih boja i opisom njihovih derivacija te (b) računalnog algoritma za automatsko prepoznavanje i označavanje boja u hrvatskom jeziku i pripadajućih oznaka klase.

U radu je dodatno predstavljena usporedna analiza različitih klasa izraza za boje pronađenih u korpusu izgrađenom iz književnih djela namijenjenih mlađoj (CLC) i starijoj (ALC) populaciji kako bi se dobili dodatni uvidi o korištenju određenog obrasca ovisno o uzorku teksta nad kojim se radi analiza. Podaci istraživanja dani su i kroz tablični prikaz tri tipa izraza za boju u klasi višerječnih izraza. Pripremljeni resursi otvaraju mogućnost dodatnih analiza tekstova iz drugih domena i s novim istraživačkim interesima koji uključuju boje u računalnoj obradi jezika.

Keywords: color terms, lexicalization patterns, multiword expressions, natural language processing, digital humanities, Croatian, NooJ

Ključne riječi: izrazi za boje, leksikalizirani obrasci, višerječni izrazi, računalna obrada prirodnog jezika, digitalna humanistika, hrvatski jezik, NooJ