

Chromatic lexicon and entities in Serbian (Working Group 1)

Chromatic expressions are textual descriptions of human visual perception of a color. Although a certain level of universality in naming colors exists in different languages and cultures, there have been a lot of researches that point out specifics in color perception of different nations [1]. These specifics reflect in chromatic lexicons of languages, so each language has its own set of words for naming basic colors [2, 3, and 4]. Beside basic colors terms, there are characteristic structures of multiword expressions used for describing color variants in each language. Hence, it is important to explore chromatic lexicon of particular language for efficient application of text processing that involves chromatic terms. Here, we will present the case of chromatic lexicon and entities in Serbian (also applicable for texts in South Slavic languages closely related to Serbian – Croatian, Bosnian and Montenegrin).

Adjectives are the basic morphological category in chromatic terminology in Serbian [3]. For naming basic colors, single terms are used: *beo*, *žut*, *narandžast*, *crven*, *ružičast*, *ljubičast*, *plav*, *zelen*, *smeđ*, *siv* and *crn* (eng. white, yellow, orange, red, pink, violet, blue, green, brown, grey and black). These adjectives have all characteristics of any other adjective in Serbian (change in case, gender and number and have comparative and superlative forms). There is also a set of single terms originated from foreign languages, commonly used in Serbian for some color variants, such as *roze* (eng. pink), *braon* (eng. brown), *teget* (eng. a variant of dark blue) etc. However, these terms do not change in case, gender or number.

Several suffixes, such as "+ast", "+ičast", "+ikav" etc, are used for alternation of color adjectives. These suffixes only change the intensity of a color, usually by decreasing it (*plavičast* can be considered as "a little bit blue" or "bluish") [5].

Color shades, tones and variants are described with word combinations, either as compound words or multiword expressions. Adjective-adjective combinations are dominant structures for describing color variants. There are three syntactic structures of these combinations with different semantics.

First, for different shades of one color, the two-adjective compounds are used in which the first adjective modifies the saturation or intensity of the second one. The modifiers are adjectives such as *tamno* (dark), *svetlo/svijetlo* (light), *zagasito* (muted) etc. In this case, both adjectives are written together, as one word. The examples are *svetlocrvena* (light red) or *tamnoplava* (dark blue). However, in some corpora, such as web, these terms are usually written separately (*svetlo crvena* or *tamno plava*). Although not being orthographically correct, having in mind importance of the web as corpus, we will not ignore these cases.

The second type of two-adjective compounds is the one in which the first adjective is derived from the noun having a specific color (as in *maslinasto zelen*, meaning green as an olive). Here also the first adjective act as a modifier of the second. In some cases, the noun is used as a modifier, having an attribute function. These noun-adjective combinations are not so common in Serbian. If they are used for naming a color, the words are separated by hyphen, as in *limun-žut* eng. "yellow as a lemon".

Finally, the third adjective-adjective compounds type is the one used for naming color variants placed in between two basic colors in color space, such as *plavozelena* eng. "the color between blue and green". In this case, the first adjective does not act as a modifier of another. Moreover, it is usually allowed to change places of adjectives (*zelenoplava* and *plavozelena* will be two different names of the same color).

For multicolor objects, two or more adjectives are used separated with hyphen (as in *crno-beli*, meaning black and white).

Expressions used for describing a color property of an object are very important in information extraction from texts in domains belonging to so-called "descriptive" sciences,

sciences in which facts about some features are presented in free-form textual descriptions (as in biology, botany, zoology, medicine etc.). In scientific texts from these domains, when referring to possible colors of some objects, it is common to use combinations and range of colors, described with multiword units. For example, the flowers of lat. *Origanum vulgare* are described by different authors as *ljubičasto-ružičasti* eng. "violet-purple", *crvenkasto-ružičasti*, *retko beli* eng. "reddish-purple, rarely white", *beli ili ružičastocrveni* eng. "white or purplish-red", *ružičasto-crveni do svetlo roze* eng. "purple-red to pale pink" etc.

Analysis of chromatic entities in Serbian has been done on domain corpus of botany texts obtained from the web, containing around 750 descriptions of plants' morphology (so called floras). The descriptive phrases from Table 1 have been dominant for describing chromatic properties of plant species.

Table 1. Descriptive phrases of chromatic properties of plants in Serbian

Chromatic entity structure	Examples in Serbian	English translation
Single adjective term	<i>plav, crvenkast, roze</i>	blue, redish, pink
Modified term	<i>svetlocrven, plavozelen, maslinastozelen</i>	light red, between blue and green, olive green
Hyphenated expressions	<i>banana-žut, crno-beli</i>	bannana-yellow, black-white
Ranges "od do" (eng. "from to")	<i>od bele do roze, crvenkasta do roze</i>	from white to pink, redish to pink
Multiple AND/OR expressions	<i>bela, roze i crvenkasta; plavičasti ili zelenkasti</i>	white, pink and redish; bluish or greenish
Mixed expressions	<i>crvenkasto-ružičasti, retko beli</i>	reddish-purple, rarely white

In electronic dictionary of Serbian [6] there are entries for basic color terms and some compounds (such as *jarkocrven*). They are all marked with semantic marker +Col (referring to color). However, there are adjectives derived from nouns that are not marked with this marker, although they are used for naming colors (such as *bakaran* eng "having color as a copper", which is marked as A+Mat only, or adjective *maslinast* eng. "having color as an olive").

For enabling complex semantic analysis, it is crucial to extend a set of semantic markers in electronic dictionary with a new one (+Chr). This marker will be used for all terms describing any chromatic property. These includes basic color names but also adjectives such as *jarko* eng. "intensive", *maslinasto* eng. "olive", *zagasito* eng. "muted" and similar, used as modifiers for intensity, saturation and color variants. Additionally, it will help in resolving certain ambiguities and co-references (expressions *sitni plavi cvetovi* eng. "small blue flowers" and *svetlo plavi cvetovi* eng. "light blue flowers" both have syntactic structure A A N; in the first example only the second adjective describe color, while in the second example both adjectives are describing chromatic properties of flowers).

References

- [1] Berlin, B., & Kay, P. (1999). Basic Color Terms: Their Universality and Evolution. Stanford: CSLI Publications (original published 1969)
- [2] Jakovljević, I., & Zdravković, S. (2014). Color representations in Hungarian -Serbian bilinguals. Perception (Supplement Vol. 43; 145). ECVF, Belgrade, Serbia
- [3] Lazarević, Radmila (2013). Leksičko-semantičko polje boja u italijanskom i srpskom jeziku, Ph.D. thesis, Faculty of Philology, University of Belgrade.
- [4] Pieczulis, Irena (2010). Semantic analysis of English, Lithuanian and Polish idioms with the headwords denoting colour terms. M.Sc. thesis, Vytautas Magnus University
- [5] Klajn, I. (2003). Tvorba reči u savremenom srpskom jeziku II: sufiksacija i konverzija. Beograd: Zavod za udžbenike i nastavna sredstva-Institut za srpski jezik SANU. Novi Sad: Matica srpska
- [6] Krstev, C. (2008). Processing of Serbian. Automata, Texts and Electronic Dictionaries. University of Belgrade, Faculty of Philology.