UDC 410

M. Petic, PhD.,
L. Raciula, PhD.

COMPUTER BASED IDENTIFICATION OF LINES WITH ROMANIAN CHROMATIC WORDS FROM POEMS

Abstract. *The article presents the computer based mechanism which lead to the identification of lines with Romanian chromatic words as the first step in the process of elaboration of a dictionary of poetic meanings of colors. The elaboration of such a dictionary needs the cooperation of specialists from different fields such as linguists and computer science professionals.*

Keywords: *chromatic words, dictionary, lemma, web service, online Web application, poems, poet, postmodernism, extraction algorithm, validation*

М. В. Петик, канд. техн. наук,
Л. А. Рачула, канд. филол. наук

ИДЕНТИФИКАЦИИ СТРОК С РУМЫНСКИМИ ХРОМАТИЧЕСКИМИ СЛОВАМИ ИЗ СТИХОВ НА БАЗЕ КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Аннотация*. Статья раскрывает механизм, который поможет идентифицировать строки с румынскими хроматическими словами на базе компьютерных технологий, как первый этап выработки словаря поэтических смыслов цветов. В разработке такого словаря необходимо сотрудничество специалистов из разных областей, а именно лингвистов и специалистов по информатике.*

Ключевые слова: *хроматические слова, словарь, лемма, веб-сервис, онлайн веб-приложение, стихи, поэт, постмодернизм, алгоритм извлечения, подтверждение*

М. В. Петік, канд. техн. наук ,
Л. А. Рачула, канд. філол. наук

ІДЕНТИФІКАЦІЇ РЯДКІВ З РУМУНСЬКИМИ ХРОМАТИЧНИМИ СЛОВАМИ З ВІРШІВ НА ОСНОВІ КОМП'ЮТЕРНИХ ТЕХНОЛОГІЙ

Анотація. *Стаття розкриває механізм, який допоможе ідентифікувати рядки з румунськими хроматичними словами на основі комп'ютерних технологій, як перший етап розробки словника поетичних значень кольорів. У розробці такого словника необхідною є співпраця фахівців різних галузей, а саме лінгвістів і фахівців з інформатики.*

Ключові слова: *хроматичні слова, словник, лема, веб-сервіс, онлайн веб-додаток, вірші, поет, постмодернізм, алгоритм вилучення, підтвердження*

1. INTRODUCTION

Our project is placed on the land of chromatic universe, based on the initiative of Michel Pastoureau studies [7]. Color carries meanings and literary texts that attest dynamic layering of meanings which vary from one era to another, from one author to another. For the Romanian historical linguistic heritage, the solution to this problem faces a specific difficulty – the relatively small number and dispersion of deposited resources [1].

A dictionary of poetic meanings of colors is undoubtedly a specialized dictionary, because it describes an important lexical area for shaping the vision, imagination, and authorial sensitivity, or creators of literature go beyond the color itself, looking for its inner beauty.

Thus, the aim of this article is to describe the mechanism that help in computer based identification of lines with Romanian chromatic words using an existing Romanian Part of Speech Tagger web service[1] in order to elaborate a dictionary of poetic meanings of colors.

Being designed as an intensive specialized dictionary (Dubois), it addresses various categories of specialists in language sciences: philologists, linguists, professors, researchers, etc. The elaboration of such a dictionary needs the cooperation of specialists from different fields, such as linguists and computer science professionals.

The linguists will decide which poem texts will be used, which color and similar words will be used as keywords, and what the other experi-

[1] http://nlptools.infoiasi.ro/WebPosRo/ - Simionescu Radu, UAIC Romanian Part of Speech Tagger, 2011

ences in this domain are. Moreover, the expertise of the poem line acquisition is also important to be done by the linguists.

Computer science specialists will decide upon the methods and mechanisms that will automate the procedure of poem line acquisition, the structure of input and output data and the means that are effective to get the point.

There is as well another topic that should be explored by both linguists and computer science specialists – the one concerning the architecture of the dictionary.

Taking all the facts that are stated above, the article is structured as follows:

First of all, we presented a brief state of art from this field with different examples. Next, it comes the description of dictionary architecture. A special section is dedicated to the algorithm presentation and Romanian Part of Speech Tagger web-service possibilities. An example of this running algorithm is given. The article ends with some conclusions and future work presentation.

2. STATE OF THE ART IN ROMANIAN POETIC LEXICOGRAPHY

A historical insight in poetic Romanian lexicography indicates that Eminescu's poetic language dictionary, coordinated by T. Vianu [9], is a first contribution that establishes a lexicographical tradition of Eminescu poems under the guidance late Professor D. Irimia who would know a wide range of development over nearly three decades [4].

A crucial moment in the poetical Romanian lexicography occurred with the effort of computer science experts was an automatic lemmatizing program developed and supervised by Marian Papahagi in 1999 from "Babes-Bolyai" University of Cluj, who established the concordance with B. Fundoianu's poems [3]. Subsequently, the same model behind the poetic Dictionary Eminescu – Poetry consistencies published during his lifetime (2002), Botosani, Publisher axis, 2 vols, developed at the Faculty of Letters of "Al. Ioan Cuza" University of Iasi, under the guidance of Professor D. Irimia and as Solomon Marcus says,"(...) would be to put a mark on exegesis Eminescu coming decades". Eminescu Concordance Series continued with posthumous poetry [6].

Some few other notes are required here: the consistencies model was taken from the Italian culture that records, in this respect, an impressive number of works coordinated by Giuseppe Savoca – *Concordanza della "Chimera" di Gabriele D'Annunzio. Testo, concordanza, liste di frequenza, indici* (1988); *Concordanza delle poesie di Camillo Sbarbaro. Concordanza, liste di frequenza, indici* (1989*); Concordanze dell' "Isotteo" e delle "Elegieromane" di Gabriele D'Annunzio. Testi, concordanze, liste di frequenza, indici* (1990); *Concordanza delle poesie di Giuseppe Ungaretti.Testo, concordanza, liste di frequenza, indici* (1993); *Concordanza del "Canzoniere* 1921*" di Umberto Saba. Testo, concordanza, liste di frequenza, indici* (1996); *Concordanza delle poesie di Cesare Pavese. Concordanza, liste di frequenza, indici* (1997); *Concordanza delle poesie di Leonardo Sinisgalli. Concordanza, lista di frequenza, indici* (2008); *Concordanza del canzoniere di Francesco Petrarca: testocritico, liste di frequenza, indici-concordanza* (2011) etc.

The second component of Eminescu's Poetic Language Dictionary which is actually the object of our interest is the Poetic Dictionary of Signs and Meanings. Vol I. Arts (2005), vol.II [5]. Primordial elements (2007). Lexicographical reflections appreciated specialists as modern work by opening both the interdisciplinary and structural (Ioana Vasiloiu) dictionary of lexicographical with the following vision "closer to the conception of the Dictionnaire des Symboles, conducted under the direction of Jean Chevalier and Alain Gheerbrant" [4, 8].

The fact that Western culture concordance dictionaries benefit from the most representative authors of the past century and the Romanian academic record from Moldova obvious progresses in this chapter are factors that favor the emergence of a culture poetic lexicography. Our lexicographical initiative is part of a comprehensive lexicographic framework, designed to fill blank spaces in Romanian linguistics. The dictionary would provide multiple benefits: on the one hand, it suggests guidelines and perspectives of interpretation in research postmodern poetry with the advantage of a global vision (by reference to one author to

another); on the other hand, the component dictionary *Concord* can serve as a basis for linguistic research-support in general. The dictionary would thus constitute a source of information that would facilitate the comprehension and interpretation of postmodernist literary phenomenon and linguistic research.

3. THE MACROSTRUCTURE OF POETIC CHROMATIC DICTIONARY

Poetic Chromatic Dictionary will be designed in two variants: as a paper published edition and an online web application. Both variants need the information to be typed and on the first stage – to be saved on a computer. This work is to be done by specialists from philology and computer science departments.

As a dictionary "is not an inventory paper that gathers all the number without system, whithout choice", but is intended to "portray an image of actual reality in its essence and representativeness, essentially sequential term or long-term" (Petru Ursache), the **ultimate goal would be to provide a dictionary that has a color picture of the universe in terms of poetic literary postmodern orientation (postmodern poetry)**.

As the development of such a dictionary approach requires a rigorous methodology that should be adapted to the specific nature of investigation we obviously faced multiple interrogations. The standing things evidently did not appeal understandable reasons, classical method recording of words with color sense and semantic field of their poetic postmodern work. In this respect, we benefited from model and experience, by developing existing poetic dictionaries at the Faculty of Letters, "Al. Ioan Cuza" University of Iazi – the dictionary Eminescu's poetic language – poetic matches and signs and meanings [4]. On the basis of Eminescu poetic model (Iasi), through a mix of the two components (matches and signs and poetic dictionary meanings) and other categories –, we have designed our dictionary's macrostructure, that is more complex including more sections combining concordances with its interpretations. There is not elaborated yet such a dictionary for Romanian language for the online use.

The dictionary should include the following sections which we tried to outline below:

- **A list of authors and titles of texts with serial number**, operating as a clue to identify the origin of the lyrics that illustrates lemmas;
- **An encyclopedic appendix column** (on the left of the page) that contains:
  a – locket with the author's photo;
  b – short bibliographical data;
  c – ethnographic information.
- **Illustration** – information provided will be accompanied by illustrative paintings - faces of color. As an illustration of a dictionary is not a simple accessory, requiring treatment similar to that which we operate where text, selection of paintings will be in accordance with the principle of relevance that color artistic universe of the creator (paintings more valuable than color as an expression of artistic vision). For example, the volume would insert Black Soulages's paintings that art can be called representative exponent of black. Illustrate your answer function: item description sui generis, completion and enhancement of aesthetic emotion;
- The second column (on the right) **interpretation itself** (description, semiotic analysis, contextual hermeneutic method psychocriticism etc.).
- **Concordances** (partial model Concordia dictionary Iasi). Concordances will be established by extraction from poems of the postmodernist poets. As it is a great number of poems, the natural way is to automate this stage. The algorithm of extraction of the specific/distinctive lines with color range of words is discussed in the following section.

In the next section we will speak about the algorithm of extraction of the concordances from the poems of the postmodernist poets. The main condition to apply the algorithm is to have the text of the poems in an electronic version.

There are several ways to get it, namely copying from Internet, typing from a paper book and scanning and OCR-izing the text.

First two variants are clear. Copying from Internet needs the existence of the correct texts. Typing from the books needs to spend more

time, and then verifying whether no mistakes were made.

The process of scanning and OCR-izing is more complicated. First of all, it needs a good scanner and a trained OCR program. More about the process of OCR is written in [1].

4. ALGORITHM EXTRACTION

The first step in the elaboration of the dictionary concern the process of establishing Chromatic concordances (*Concordanțe cromatice)*, the processing will consist in the extraction of those lines which contain a word that expresses a chromatic range. That is why, the lines will be grouped and counted by keywords. In order to achieve it timely, automatic mechanisms may be used. The necessary condition to use these mechanisms is the *input* available of structured data – the poem should be in an electronic readable format and the list of keywords that express the chromatic set of colors. As a result the *output* we will get a file like:

nr1 nr2 verse

where nr1 represents the line number in the poem, nr2 is the numer that indicate to the first occurence in the verse of the keyword and verse represent the text of the verse line of the poem.

The natural way for automatic processing of these data in order to get the desired result is the following (Fig. 1):

1) read from input a line to be processed;

2) search for the word that matches the keyword from chromatic range;

3) write the line that contains the keyword in a special file with the lines that contain keywords from corresponding chromatic range.

Below we will describe how each of the three steps listed above will be achieved.

Reading of a line for processing is simply achieved by using the standard procedures of reading from a file and text string processing.

Search in verse of a keyword occurs in several stages:

a – ***Verse division*** in separate words;

b – ***Lemmatizing*** of each word (getting lemma of the words);

b – ***Matching*** the lemmatized word with keyword.

To divide the verse into separate words, we will use special standard routines or those defined by the programmer to separate words in a verse.

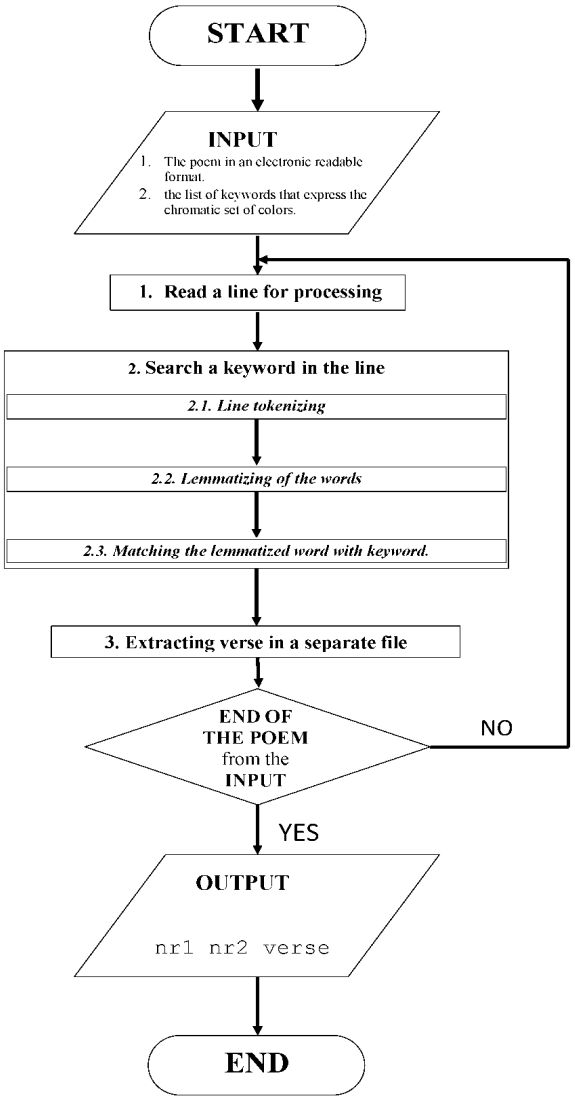

Fig.1. Algorithm scheme

Lemmatizing is probably the most difficult task in this entire algorithm. To do this, you will have to use the online services that were already developed by other researchers for different purposes. In our case, the best solution would be to use the PoS Tagger for Romanian[2], developed within the natural language processing research group at the Faculty of Computer Science of the "A. I. Cuza" University of Iasi. In this case, the procedures of verse division and lemmatizing will be merged because this web application offers the possibility to give the original data of a text and get an annotated (labeled) text with grammatical categories and word lemma. The existent WSDL service allows us to use this service in a program for our aims.

[2] http://nlptools.infoiasi.ro/WebPosRo/ – Simionescu Radu, UAIC Romanian Part of Speech Tagger, 2011

The matching will take place with special subprograms that are available in the programming language that will be used.

Extracting verse in a separate file will be made by adding the containing keywords to the list of verses. For each keyword we will create a separate file.

Even we have automated the process of extraction of those lines with consist of word that expresses a chromatic range, we need a stage of validation of all those work done, just because the Romanian part of speech tagger and lemmatizer is not perfect and works only with an accuracy of 96,6 %. That is why the expertise of the poem line acquisition is also important to be done by the linguists.

5. ALGORITHM EXAMPLE

In order to validate the mechanisms described above, we will try to see how the algorithm will work on a specific text.

In our case we took the first part of the poem Dragostea (in English Love) by Mircea Cărtărescu [2]:

1 deasupra mamei răsărise un curcubeu negru.
2 pe-atunci mama era doar o fetiţă
3 şi părul ei era din şuvite de diamant.
4 ea se opri din jocul ei
5 îşi netezi rochiţa
6 şi privi înspre curcubeu.
7 era un curcubeu negru şi de atâta negru scânteietor
8 pe câmp muşeţelul se întunecase.
9 mama privi înspre curcubeu.
10 era atât de jos, că aproape-l putea atinge cu buzele
11 iar părul ei electrizat, transparent
12 se lipea de tuburile moi, ca de orgă.
13 roşul curcubeului era negru.
14 şi oranjul lui era negru.
15 şi galbenul lui era negru.
16 şi verdele lui era negru.
17 şi albastrul lui era negru.
18 şi indigoul lui era negru.
19 doar violetul lui rămânea violet.
20 violetul lui era un şuvoi care se pierdea în mare.
21 mama ştiu atunci că este pierdută.
22 avea să cunoască dragostea.
23 şi brusc umbra ei pe safaltul cald din floreasca
24 se împodobi cu vene şi oase.
25 mama sui într-o corabie din pânză de păianjen
26 şi o porni în sus pe curcubeu.
27 umbra ei, cu vertebre şi intestine
28 îi continua joaca.

In this text we try to find lines containing the word *black* (in Romanian – *negru*, with its flections – *negri*, *negre*, etc.), *dark* (in Romanian – întuneric, întunecat).

First line: *deasupra mamei răsărise un curcubeu negru.* is read by the program from the file and divided into separate words and lemmatized using the Romanian PoS Tagger service[3]:

deasupra – adverb, lema=deasupra
mamei – substantiv, lema=mamă
răsărise – verb, lema=răsări
un – articol, lema=un
curcubeu – substantiv, lema=curcubeu
negru – adjectiv, lema=negru.

As all lemmas were found, among them the keyword *negru* was found (in English *black*) that begins with character 38 of the line. In this respect, the first line will be written in the file negru.txt in the following way:

1 38 deasupra mamei răsărise un curcubeu negru.

Number 1 indicates that the first line of the poem was included, 38 – indicates with which character begins the keyword. Further, we proceed in the same way.

The next line containing the keyword is one of the 7-th: *era un curcubeu negru şi de atâta negru scânteietor*.

We noticed that the verse contains the keyword twice. The proposal is to write 2 times this line:

1 38 deasupra mamei răsărise un curcubeu negru.
7 17 era un curcubeu negru şi de atâta negru scânteietor
7 35 era un curcubeu negru şi de atâta negru scânteietor

In verse 8 (*pe câmp muşeţelul se întunecase*.), we find the word *întunecase* that being lemmatized is a verb întuneca (in English dark). In our case, it does not match the word întunecat (in English – *dark*).

The following lines are analog processed to yield the file negru.txt:

1 38 *deasupra mamei răsărise un curcubeu negru.*
7 17 era un curcubeu negru şi de atâta negru scânteietor
7 35 era un curcubeu negru şi de atâta negru scânteietor
13 23 roşul curcubeului era negru.
14 20 şi oranjul lui era negru.
15 21 şi galbenul lui era negru.
16 20 şi verdele lui era negru.
17 22 şi albastrul lui era negru.
18 21 şi indigoul lui era negru.

This new created file should be saved and integrated in the database with other files for the future web application that would contain all the

[3] http://nlptools.infoiasi.ro/WebPosRo/ – Simionescu Radu, UAIC Romanian Part of Speech Tagger, 2011

information stated in the section concerned to the architecture of future system.

6. CONCLUSIONS AND FUTURE WORK

In the future this dictionary of postmodern poetic chromatic language would treat poetic poetry postmodern senses color by color: Black, Blue, Red, Yellow, Green, Orange, Purple, and White. The architecture of the whole data series may be set as 8 volumes.

Dictionary would be designed and completed in the sense of cultural heritage preservation. That is why future public-access Web application is intended to be an important and useful interoperable computational linguistic resources tool that would contain poems from several postmodern poets.

Since a dictionary generally is built as a tool for knowledge, the lexicographer Jean-Louis Trouillon suggests a study in lexicography in a specialized Approached Language whose development process can be partially taught to students, master students in vocational training in the form of collaborative work.

References

1. Boian E., Cojocaru S., Ciubotaru C., Colesnicov A., Malahov L., and Petic M. Electronic Linguistic Resources for Historical Standard Romanian. *In: Proceedings of the 9th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language",* May 16-17, 2013, Miclăuşeni-Iaşi, România, pp. 35 – 50.

2. Cărtărescu M., *O seară la operă*, Chișinău, Î.E.P. Știința, 2009, 148 p.

3. Papahagi M., *Concordanța Poeziilor Lui B. Fundoianu*, *Coord.,* Cluj-Napoca, Echinox, 1999, 700 p. Dumitru Irimia, and Botoşani Axa, *Dicționarul Limbajului Poetic Eminescian: Concordanțele Poeziilor Antume,* 2002, Vol.2, 1048 p.

4. Dumitru Irimia, Irina Andone, Odette Arhip., Iași, Editura *Dicționarul Limbajului Poetic Eminescian: Semne și Sensuri Poetice, Universitătii „Alexandru Ioan Cuza",* 2005,
Vol. 2, 246 p.

5. Solomon M. *Un nou Dicţionar Eminescu în România literară*, 2007, No. 4. – http://www.romlit.ro/un_nou_dicionar_eminescu

6. Pastoureau M., *Negru. Istoria Unei Culori*. Traducere de Emilian Galaicu-Păun, Chișinău, Cartier, 2012, 252 p.

7. Vasiloiu I., *Eminescu: Semne și Sensuri în Caiete Critice*, No. 1-2-3, 2003, pp.35 – 37.

8. Vianu T., *Dicţionarul Limbii Poetice a Lui Eminescu*. Bucureşti, 1968. 646 p.

Received 28.02.2014

Mircea Petic,
Lecturer, PhD. in Computer Science, Departament of Mathematics and Computer Science, Alecu Russo Balti State University Republic of Moldova,
38 Puskin street, Balti, MD-3100.
Tel.: +373 231 52488.
E-mail: petic.mircea@gmail.com

Lilia Raciula,
Associate Profesor, PhD. in Philology, Departament of Romanian Language, Alecu Russo Balti State University
Republic of Moldova,
38 Puskin street, Balti, MD-3100.
Tel.: +37253190343.
E-mail: raciula@mail.ru