SECTION IX. APPLIED LINGUISTICS: TRENDS AND ASPECTS OF STUDIES

Illya Danyliuk

УДК 81’33

COLOR MAP OF TARAS SHEVCHENKO’S “KOBZAR” WITH *MATHEMATICA*

The article describes an idea and its realization process for creating Color Map (CM) for the text – in particular “Kobzar” by Taras Shevchenko. CM is a composition, a grid made of colored rectangles (or any other figures) – one for every name of color in the original text. Absolutely objective result shows visually the distribution of particular adjectives. The original article is in Computable Document Format (CDF) and is suitable for random text in Ukrainian considering inflection and even derivation.

Keywords: color name, language model, inflection, Shevchenko, MATHEMATICA.

Today there are a lot of instruments for natural language processing, many of them are highly automated. Modern programming languages make it possible to process in detail the text data on the level of separate symbols, groups of symbols (phrases and sentences), progressively some functions to deal with sense and meaning appear. The amount of built-in string functions grows up, and it makes the developer’s potential wider and helps to save time because of no need for manual creating of such functions and procedures. For example, the procedure for measuring the difference between two sequences – *Levenshtein distance* or the *minimal editing distance* – is to be written by the developer on Python or else as a user defined procedure [Jurafsky 2009]. But in some modern systems this procedure is included as pre-installed one. Computational software program *Mathematica* from Wolfram Research, which is widely used in many scientific, engineering, mathematical and computing fields [Wellin 2013] – and we’ll exploit it for linguistics – has a built-in function *EditDistance [x, y]*. We suppose the major NLP algorithms – such as lemmatization, word forms synthesis, POS-tagging and syntactical analysis or even automated translation, described in [Баранов 2003; Волошин 2004; Дарчук 2008; Карпіловська 2006; Марчук 2000; Партико 2008] – will become the built-in procedures.

This article is originally written in CDF format – *Computable Document Format* for *Mathematica* system, or *Notebook* (*.nb) – and can be downloaded [Данилюк 2014], so some code in printed version will be hidden. To view CDF on your desktop you will need a free viewer from [http://www.wolfram.com/cdf-player/].

So, the **main goal** of the article is to describe the process, instruments, and directly code for automated generating a color map for random text in Ukrainian, and particularly for the “Kobzar” by Taras Shevchenko – his 200-years anniversary approaches. Color map (CM) is a composition, a grid made of colored rectangles (or any other figures) – one for every name of color in the original text. Every usage of adjective meaning color – ‘*білий’ (white), ‘чорний’ (black), ‘червоний’, (red),‘золотий’ (gold)* – in CP will be shown as square of the proper color. So, we can get a full and absolutely objective presentation of lexical data from the particular text and some traits of “world picture”, color concepts for the work of literature. It is a matter of topical interest in Ukrainian linguistics and philology, so the instrument for automated color information retrieval from any text is highly needed as we suppose.

You can think of CM as of automated infographics of language data visualization. The common idea belongs to Tatiana Druzhnyaeva from *Esquire* and some aspects of coding to Roman Osipov from Moscow State University of Fine Chemical Technology.

We divided our research into **several tasks**: 1) to retrieve all possible statistical information from the text of “Kobzar” for further analysis; 2) to create some procedures (in code for

Mathematica Language) to work with certain words and sentences; 3) to build a language model for color names in Ukrainian considering inflection and some aspects of derivation; 4) to use the model for generating CM of "Kobzar", and to describe its perspectives.

The research **object** – text of "Kobzar" by Taras Shevchenko – basically is a txt-format file from [litopys.org.ua], prepared for processing (in Unicode, every token is divided by space). And the **subject** thus is a usage of color names represented in the form of CM.

So let's begin with the first task. The text file – *kobzar.txt* – has to be in the same folder with *notebook* file, and we read data from it to the variable *kobzar*:

```
Short[kobzar=Import[FileNameJoin[{NotebookDirectory[],"kobzar.txt"}]]]
```

When we apply to *kobzar*, we work with whole text and can get basic statistics – how many symbols it contains:

```
StringLength[kobzar]
588995
```

or how many strings there are:

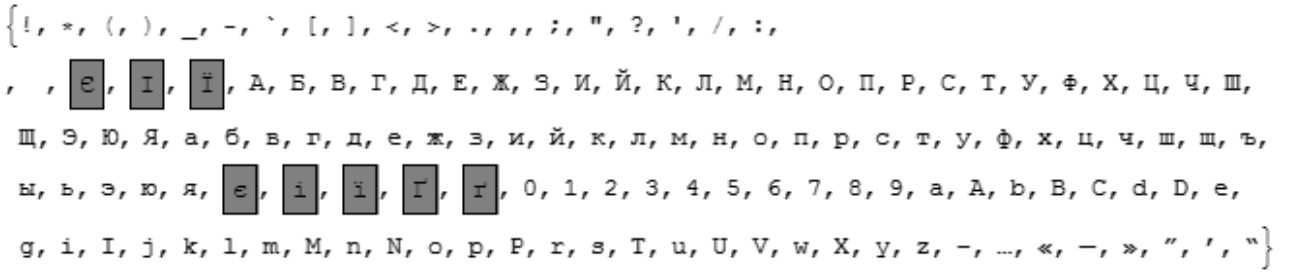
```
StringCount[kobzar,"\n"]
23929
```

or how many sentences approximately it has (assuming the normal sentence can ends with dot, exclamation or question mark):

```
StringCount[kobzar,{».»,»?»,»!»}]
9044
```

or what symbols are used in text, in a sorted view:

```
Union[Characters[kobzar]]
```



Turn your attention to the place where some special Ukrainian letters (framed) have appeared after sorting. The reason is that in standard table of symbols they are located not within common Cyrillic list but on rather accidental positions. This inconvenience you have to take into account using regular expression (RE) patterns like "all Ukrainian letters from A to Я". In fact, RE like /[Є-ґ]+/ is good enough for pattern "all Ukrainian letters", but it returns a couple of non-Ukrainian symbols – Э, ъ, ы, э.

Next step we move to words processing. A variable *allWords* contains all wordforms from *kobzar* without capital/non-capital letters distinction – by replacing using RE:

```
Short[allWords = Sort[Tally[DeleteCases[StringSplit[StringReplace[StringReplace[kobzar,
Thread[Join[CharacterRange["A", "Я"], {"Є", "І", "Ї", "Ґ"}] -> Join[CharacterRange["a", "я"], {"є",
"і", "ї", "ґ"}]]], RegularExpression["[^" <> StringJoin@Join[CharacterRange["a", "я"], {"є", "і",
"ї", "ґ"}] <> "]"] -> " "], " "], ""]], #1[[2]] > #2[[2]] &], 5]
```

The number of tokens in *kobzar* is 110617, and unique wordforms is 17617. 100 of the most frequency ones are:

```
allWords[[1";;" 200]][[";;",1]]
```

{"і", "не", "а", "в", "на", "та", "з", "й", "я", "що", "у", "за", "як", "то", "ти", "и", "до", "де", "так", "чи", "ж", "його", "щоб", "по", "мене", "ні", "ще", "собі", "мов", "б", "все", "мені", "мати", "моя", "бо", "люде", "мій", "тебе", "серце", "с", "аж", "о", "той", "буде", "світі", "вже", "ми", "хоч", "із", "там", "без", "поки", "може", "хто", "того", "нехай", "їх", "нема", "діти", "боже", "над", "він", "було", "те", "неначе", "он", "ви", "мої", "коли", "тяжко", "бога", "тілько", "нас", "же", "про", "лихо", "море", "під", "добре", "таки", "од", "уже", "половина", "тихо",

"сльози", "ой", "старий", "свою", "вона", "как", "отак", "чого", "її", "очі", "друга", "треба", "кого", "п", "колись", "нічого"}

For creating the possibility to look for particular wordform in the text we build a function *wordPosition*, which returns array of symbols positions.

```
replacements=Thread[Join[CharacterRange["а","я"],{"є","і","ї","ґ"}]->Join[CharacterRange["А","Я"],{"Є","І","Ї","Ґ"}]]
```

Here is the output for *тополя*:

```
wordPosition["тополя"]
```

{{19462, 19469}, {40838, 40845}, {56587, 56594}, {134756, 134763}, {248726, 248733}, {308308, 308315}, {320139, 320146}, {517723, 517730}}

If we find positions for dots (and other symbols at the end of the sentence – "!", "?", "…") it will be possible to retrieve a whole sentence (sequence of symbols between two punctuation marks) and put it to the variable *sentence*.

```
Short[dots = #[[1]] & /@ StringPosition[kobzar, {".","?","!","\[Ellipsis]"}], 5];
sentense[{min_,max_}]:=Block[{start=Select[Nearest[dots,min,10],#<min&][[1]]+1,end=Select[Nearest[dots,max,10],#>max&][[1]]},StringTake[kobzar,{start,end}]]
```

Combining *wordPosition* and *sentence* is a concordance for particular wordform:

```
Grid[Transpose@{StringReplace[sentense /@ wordPosition["тополя"],
"\n" -> ""]}, Background -> {None, {{Orange, LightGray}}},
ItemStyle -> Directive[16, Bold, FontFamily -> "Arial"],
Alignment -> Left, Dividers -> All]
```

Як тополя , стала в полі При битій дорозі ; Як роса та до схід сонця , Покапали сльози , За сльозами яа гіркими І світа не бачить , Тілько сина пригортає , Цілує та плаче .
–Петербург] Тополя По діброві вітер виє , Гуляє по полю , Край дороги гне тополю До самого долу .
Дивлюся , сміюся , дрібні утираю , — Я не одинокий , є з ким в світі жить ; У моїй хатині , як в степу безкраїм , Козацтво гуляє , байрак гомонить ; У моїй хатині синє море грає , Могила сумує , тополя шумить , Тихесенько Гриця дівчина співає , — Я не одинокий , є з ким вік дожить .
Та й виросла Ганна кароока , Як тополя серед поля , Гнучка та висока .
21 октября 1845 , Марьинское Наймичка Пролог У неділю вранці– рано Поле крилося туманом ; У тумані , на могилі , Як тополя , похилилась Молодиця молодая .
Не прийнялись три ясени , Тополя всихала , Повсихали три явори , Калина зов'яла .
Мов тополя , виростає Світові на диво .
— І похилилась , мов тополя Од вітру хилиться в яру .

Now we have to define a common color model for using to build CM. Its relevance and deepness in Ukrainian – including lemmatization accuracy and color concept detail – are crucial for CM quality. The basic approach and the first step we use are to find adjectives, which signify colors directly and consist of one wordform. We put those adjectives to array variable *tc* (*table of colors*) and describe them with RGB model (the piece of code is quite long and is suitable for viewing in digital version). The adjectives are: *білий, червоний, зелений, синій, жовтий, чорний, сірий, рожевий, коричневий, блакитний, пурпурний, пурпуровий, оранжевий, помаранчевий, фіолетовий, амарантовий, бурштиновий, аметистовий, абрикосовий, аквамариновий, арсеновий, спаржевий, бежевий, латунний, бронзовий, брунатний, карміновий, морквяний, лазуровий, каштановий, шоколадний, цинамоновий, кобальтовий, мідний, кораловий, кукурудзяний, блаватний, кремовий, малиновий, джинсовий, смарагдовий, баклажановий, лляний, золотий, індиго, нефритовий, хакі, лавандний, лимонний, бузковий, малахітовий, гірчичний, оливковий, помаранчевий, ліловий, персиковий, грушевий, барвінковий, сливовий,*

бурий, іржавий, шафрановий, сапфіровий, багряний, срібний, болотний, мандариновий, будяковий, бірюзовий, ультрамариновий, фіолетовий, пшеничний.

And here is the fragment of the *tc* presentation:

білий	GrayLevel[1]
червоний	RGBColor[1, 0, 0]
зелений	RGBColor[0, 1, 0]
синій	RGBColor[0, 0, 1]
жовтий	RGBColor[1, 1, 0]
чорний	
сірий	GrayLevel[0.5]
рожевий	RGBColor[1, 0.5, 0.5]
коричневий	RGBColor[0.6, 0.4, 0.2]
блакитний	RGBColor[0, 1, 1]
пурпурний, пурпуровий	RGBColor[1, 0, 1]
оранжевий, помаранчевий	RGBColor[1, 0.5, 0]
фіолетовий	RGBColor[0.5, 0, 0.5]
амарантовий	RGBColor[229/255, 43/255, 16/51]

The color model also includes rules for retrieving different wordforms and finding their lemmas, so *червоний, червоного, червона, червонії* etc. will be represented by the lemma *червоний* and the red square. For that purpose we use sequentially a wordforms synthesis with the dictionary of the stems and endings (generate all possible wordforms for every adjective in the table of colors), then we find the positions for these wordforms in *kobzar*, assign these positions to particular lemmas and finally range the lemmas as they appear in the text and replace them with color squares.

Adjective inflection in Ukrainian includes hard (*червоний*) and soft (*синій*) groups, contracted (*червона*) and non-contracted forms (*червоная*). The variables *ColorEdningsTv* and *ColorEdningsMk* contain accordingly ending for hard contracted and non-contracted forms and soft contracted and non-contracted forms:

ColorEdningsTv={"ий","ого","ому","им","ім","е","а","ої","ій","у","ою","і","их","им", "ими", "еє","ая","ую","ії"};

ColorEdningsMk={"ій","ього","ьому","ім","є","я","ьої","ю","ьою","і","іх","іми","еє", "єє","яя", "юю","ії"};

A variable *colorRules* includes a generated array of all possible wordforms for naming colors (combination of stems from *tc* and endings from *ColorEdningsTv* and *ColorEdningsMk*) with built-in probable modification of the adjective using affixes (*білий – біленький*). The fragment od code:

```
colorRules=Flatten[{Thread[Flatten[Table[{"біл",
"біленьк"}[[v]]<>ColorEdningsTv[[u]],{v,2},{u,Length[ColorEdningsTv]}]]->White],
Thread[Table["син"<>ColorEdningsMk[[u]],{u,Length[ColorEdningsMk]}]->Blue],…
Thread[Table["пшеничн"<>ColorEdningsTv[[u]],{u,Length[ColorEdningsTv]}]-
>RGBColor[245/255,222/255,179/255]]}]
```

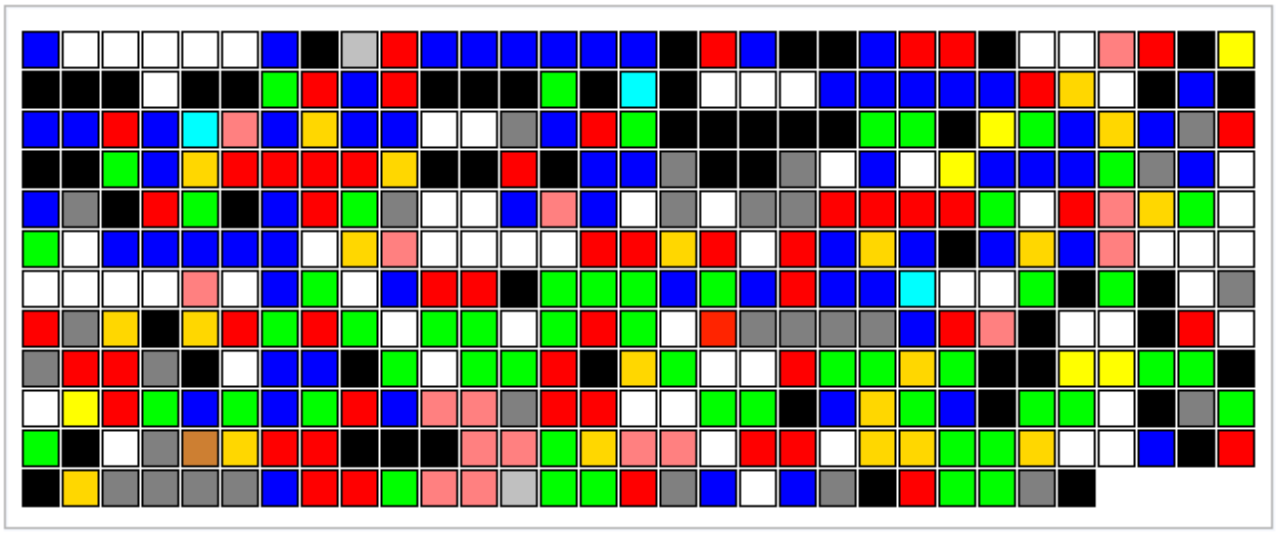
The positions of generated wordforms are retrieved to variable *colorInformationPre* and then are sorted in *colorInformation*:

```
colorInformationPre={#,wordPosition[#]}&/@colorRules[[;;,1]];
colorInformation=Sort[Flatten[Partition[Riffle[#[[2]],#[[1]],{2,-1,2}],2]&/
@colorInformationPre,1],Mean[#1[[1]]]<Mean[#2[[1]]]&]
```

And finally we build CM:

Panel[Grid[Partition[Graphics[{#,EdgeForm[Black],Rectangle[]},ImageSize->20]&/@(colorInformation[[;;,2]]/.colorRules),30,30,1,""],Spacings->{0,0}]]

Here is the result:



In digital version for this article a possibility to see the real sentence by clicking on the particular square is added.

As a common conclusion: the procedure of creating CM for Ukrainian text in *Mathematica* is rather complicated because of excluding of some symbols (*є, і, ї, ґ*) from "normal" list in Unicode, necessity to pass over the fact that Cyrillic letters are not included to the list of correct word symbols in RE as Latin ones and digits. Other way creating some explicit functions helps to solve that problem effectively. And the *Mathematica* system itself is a powerful environment for computational linguistics studies with rich high-level programming language, and can be recommended for studying by the students of philological departments in Ukrainian high schools.

As a particular conclusion: Taras Shevchenko's "Kobzar" in a color conceptual point of view is rich in color names, main of them are white and black, blue and red, green, and gray. Comparing CM for "Kobzar" with CMs for other works of literature and folklore must be significant, but it's a subject for further studying. As it is maintained about Ukrainian in [Ковтун 2009: 51], "color feature appeared in the language in diachronic sequence. In folklore initially the dominant is represented by definitions of white and black, followed by red (triad white – black – red), later – green and yellow, then – blue and brown." It seems even superficially that Taras Shevchenko's work is close to national folklore and is a piece of common Ukrainian discourse.

The described method to build a CM form linguistics position is just a basic one. It can be improved by adding to the color model derivative adjectives (*білявий, чорнющий*), verbs (*біліти*), and nouns (*чорнота*), some Ukrainian names of secondary colors (*ясно-зелений*), closely related lexicon – coat colors, etc.

References

Баранов 2003: Баранов, А.Н. Введение в прикладную лингвистику [Текст] / А. Н. Баранов. – М. : Едиториал УРСС, 2003. – 360 с. – ISBN: 5-8360-0196-0.

Волошин 2004: Волошин, В. Г. Комп'ютерна лінгвістика : Навчальний посібник [Текст] / В. Г. Волошин. – Суми : ВТД «Університетська книга», 2004. – 382 с. – ISBN: 966-680-134-5.

Данилюк 2014: Данилюк, І.Г. Аналіз тексту "Кобзаря" Тараса Шевченка в середовищі Mathematica : символи, слова і кольори [Текст] / І. Г. Данилюк. – Режим доступу : https://app.box.com/kobzar. – Назва з екрана.

Дарчук 2008: Дарчук, Н.П. Комп'ютерна лінгвістика : Автоматичне опрацювання тексту [Текст] / Н. П. Дарчук. – К. : Видавничо-поліграфічний центр «Київський університет», 2008. – 351 с. – ISBN 978-966-439-079-5.

Карпіловська 2006: Карпіловська, Є. А. Вступ до прикладної лінгвістики : комп'ютерна лінгвістика. Підручник [Текст] / Є. А. Карпіловська. – Донецьк : ТОВ «Юго-Восток, Лтд», 2006. – 188 с. – ISBN 966-374-078-7.

Ковтун 2009: Ковтун, Л. Український колористичний код світотворення [Текст] / Л. Ковтун // Вісник Київського національного університету імені Тараса Шевченка. Українознавство. – Вип. 13 / КНУ імені Тараса Шевченка. – Київ : ВПЦ "Київський університет", 2009. – ISSN 1728-2330.

Марчук 2000: Марчук, Ю.Н. Основы компьютерной лингвистики [Текст] / Ю. Н. Марчук. – М. : Народный учитель, 2000. – 320 с. – ISBN 5-17-039480-2.

Партико 2008: Партико, З.В. Прикладна і комп'ютерна лінгвістика : Вступ до спеціальності [Текст] / З. В. Партико. – Львів : Афіша, 2008. – 224 с. – ISBN 978-966-325-092-2.

Jurafsky 2009: Jurafsky, D., Martin, J.H. Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition [Text] / D. Jurafsky, J. H. Martin. – Prentice Hall, 2009. – 988 pp. – ISBN 0-13-095069-6.

Wellin 2013: Wellin, Paul R. Programming with Mathematica, An Introduction [Text] / P. R. Wellin. – Cambridge, 2013. – 711 pp. – ISBN: 9781107009462

У статті описано ідею і реалізацію процесу створення кольорової мапи (КМ) для довільного тексту взагалі, й зокрема для «Кобзаря» Тараса Шевченка. КМ – це множина, сітка кольорових прямокутників (або інших фігур), кожен із яких стосується певної кольороназви (лексеми на позначення кольору) у вихідному тексті. Цілковито об'єктивний результат демонструє дистрибуцію відповідних прикметників. Оригінал статті створено у новітньому CDF – форматі обчислюваного документа, і може бути застосований до довільного тексту українською мовою з урахуванням словозміни й словотвору.

Ключові слова: кольороназва, мовна модель, словозміна, Шевченко, MATHEMATICA.

Available 12 September 2013.