PressAcademia Procedia

2nd World Conference on Technology, Innovation and Entrepreneurship
May 12- 14, 2017, Istanbul, Turkey. Edited by Sefer Şener

STATISTICAL ANALYSES OF SAME CONTENT TEXTS WRITTEN IN DIFFERENT LANGUAGES

DOI: 10.17261/Pressacademia.2017.580
PAP-WCTIE-V.5-2017(18)-p.124-129

Mentor Hamiti[1], Elissa Mollakuqe[2], Asmir Rahmani[2], Florinda Muhaxhiri[3]
[1]South East European University, m.hamiti@seeu.edu.mk
[2]University of Prizren "Ukshin Hoti", elissamollakuqe@gmail.com
[3]State University of Tetova, florinda.muhaxhiri@gmail.com

ABSTRACT

Language is the basic and most consummate way of communication between people. It can be materialized in two ways: spoken and written. Every society has a spoken language, even the primitive ones. Although only civilized societies have a written language with a defined alphabet. The presence of letters in the context of words determines the meaning, while the determined order of these in words presents a work of art. It is, thus, understandable to raise the question: Which letter is used the most and least in different languages? Or maybe there is similarity on their distribution even though it has to do with languages which use different alphabets? Or in general, which are the differences or what could different languages have in common when they interpret the same content?! The answer to this question remains within the scope of this paper

Keywords: Text, English, Albanian, Turkish, Bosnian

1. INTRODUCTION

Computer language software and their presence on the Internet have become a vital part of communication and modern concepts of the so-called "scientific field". Linguists have taken seriously the provocation of the computer era in the field of linguistics, because computer linguistics is the only way for protecting, enriching and advancing every language in the world. The aim of this research is to present the continuous development of the languages, including the statistical research component. With the help of the original program, written in C# programming language, we set the computer in service of different languages, since text with same content were written in English, Albanian, Bosnian and Turkish languages. Linguists can use the gained results for further linguistics research and analyses (Hamiti, 2015).

2. CLASSIFICATION OF ALPHABET LETTERS AND THE SPECIFICS OF COMPUTER BASED PROCESSING

The organization of the letters for the alphabets in question is structured in separate forms for each language. The four languages belong to different families. English belongs to the family of the Germanic languages and uses 26 Latin letters (Murphy, 2008). Albanian is part of Indo-European family of languages with a total of 36 letters, of which 34 are standard (Latin) (Hamiti, 2005). Among these 9 are considered as consisting of 2 letters (digraphs) and 2 are diacritical marks. Bosnian belongs to the family of Slavic languages consisting of 30 letters, of which 24 are standard letters, while 6 are nonstandard, containing also composite ones (Schumann, 2010). Turkish is a language on its own, consisting of 29 letters of which 22 are standard while 7 are nonstandard and does not contain digraphs (Attaoullah, 2012). Because of the stated differences between the languages that are being treated, the standard existing applications, which in most cases are in English, cannot be used for universal processing of texts written in these languages. Therefore, it is necessary to design an original application that takes into account the specifics of the respective languages.

2.1. English Language, the Specifics of Computer Based Processing

English is described as computer language and notably advances in terms of computer processing compared to many other languages. All letters of the English alphabet are found in all standard computer keyboards, therefore the writing of texts and the text analysis written in this language are much more convenient and easier to realize. The English alphabet letters belong to ASCII standard, which remains as the international global standard for the interpretation of commands and texts

based on communication protocols. In addition, this standard has the support of all programming languages like C, C++, C#, etc. Also it allows the possibility of using any of them, without difficulties in processing the written texts in this language. That is not the case with other languages that require UNICODE support for interpretation and processing (wordandphrase, 2015).

Figure 1: Letters of the English Alphabet

A a	B b	C c	D d	E e	F f	G g	H h	I i	J J	K k	L l	M m
N n	O o	P p	Q q	R r	S s	T t	U u	V v	W W	X x	Y y	Z z

2.2. Albanian Language, the Specifics of Computer Based Processing

Albanian consists of 36 letters, of which 34 can be written easily from the computer keyboard, while two are letters with diacritical signs: "Ë, ë" and "Ç, ç", which means that the process of their writing is more complex. It can be done in one of the following ways: by using keyboard configuration where some additional Albanian language characters are hidden within special keys like parentheses, by installing fonts for text processors, or by using codes for generation of these two symbols. In addition, 9 out of the other 34 letters are treated as compound letters, also known as digraphs, which are formed by linking two Latin characters in one single letter for representing a single phoneme in Albanian language. This in programming is accompanied by another feature, that does not allow each letter to be treated as special character, but takes the meaning of the string, thus additionally complicates the programming and does not allow the use of existing applications for text processing (Academia, 1976)(Hamiti, 2015). Therefore, within the designed application in C#, special algorithms for solving the mentioned problem are used. As an illustration, it is worth mentioning that the MS Word count for Albanian language does not generate the correct result!

Figure 2. Letters of the Albanian Alphabet

A A	B B	C c	Ç ç	D d	DH dh	E e	Ë ë	F f	G g	GJ gj	H h	I i	J j	K K	L l	LL ll	M m
N N	NJ Nj	O o	P p	Q q	R r	RR rr	S s	SH sh	T t	TH th	U u	V v	X x	XH Xh	Y y	Z z	ZH zh

2.3. Bosnian Language, the Specifics of Computer Based Processing

Bosnian has another category of the alphabets. Bosnian uses both alphabets simultaneously: The Latin and the Cyrillic alphabet (Ronelle, 2010). The alphabet contains a total of 30 letters, of which 24 are standard letters, while 5 are considered as nonstandard based on ASCII code, while 2 are composed letters, and one of them is a letter with diacritical mark. Therefore, writing of texts in this language through standard keyboards is very difficult and almost impossible without the installation of special fonts and respective keyboard configuration. But in terms of programming, however, processing of texts is conditioned through the programming languages that enjoy UNICODE support.

Figure 3: Letters of the Bosnian Alphabet

A A	B b	C c	Č č	Ć ć	D d	Dž dž	Đ đ	E e	F f	G g	H h	I i	J j	K K
L L	LJ lJ	M m	N n	Nj nj	O o	P p	R r	S s	Š š	T t	U u	V v	Z z	Ž Ž

2.4. Turkish Language, the Specifics of Computer Based Processing

Turkish has an alphabet suitable for the sound of its own language. The Turkish alphabet consists of 29 letters, which can be categorized into two simple categories, with 22 standard and 7 nonstandard noncomplex letters (Asuman, 2001)(Underhill, 2010). So, in terms of programming, it allows the possibility of using characters, which was not the case with the previous languages, the Albanian and Bosnian.

Figure 4: Letters of the Turkish Alphabet

A A	B b	C c	Ç ç	D d	E e	F f	G g	Ğ ğ	H H	I ı	İ i	J j	K k	
L L	M m	N n	O o	Ö ö	P p	R r	S s	Ş ş	T T	U u	Ü ü	V v	Y y	Z z

3. THE APPLICATION DESIGNED FOR TEXT PROCESSING

Considering the specifics of the analyzed languages, "Microsoft Visual Studio 2013" platform and C# programming language were used within this paper, for designing a specific program, which fulfills the needs for textual processing of English, Albanian, Bosnian and Turkish languages. The source code for this program, used for generating the results, is being fully presented through this paper. The program is called "Analyses of letters".

This program totally answers the needs for analyzing written texts in all four languages, and can be very easily adapted for other languages. It enables reading textual files, sets total number of characters within the textual file (file with .txt extension) (Hamiti, 2015), counts every single letter, punctuation marks and special characters separately, like for example empty line, the use of TAB and jumping on new line, etc. Also, it enables direct calculation of letter appearance frequency within the individual textual file.

Unable to find more professional texts with similar content translated into these four different languages, in the frame of this paper, only 205 pages of text were analyzed. In order to obtain convincing results, we categorized them into ten groups: Basic expressions; greetings and invitations; communications and notifications; merchandise, money and numbers; transportation; food and beverage; guidance; documents (regulations); health and emergencies; and in the tenth group the other cases, containing some specific data such as messages, conversations, etc. In the paper we generated the results of averages of specific groups in order to avoid eventual errors and the possibility of deviation of results for specific content. Therefore, the possibility for some deviation remains open for the obtained values if considering bigger dataset such is the case with the EU languages, where you can find text with the same content in dozens of different languages, like the case with JRC-Acquis dataset. But the Albanian, Turkish and Bosnian languages do not enjoy this privilege, because they are not yet EU members.

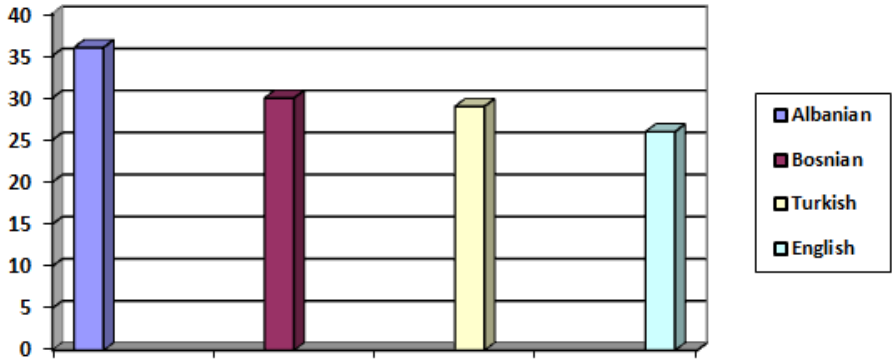
Figure 5: Analysis of letters

4. SAME CONTENT TEXT ANALYSEZ RESULTS

The number of characters, punctuation, standard and nonstandard letters, differ between the four analyzed languages, the English, Albanian, Bosnian and Turkish. The results are quite interesting, on how the same content is presented in four different languages. Figure 6. shows the total number of characters for all four languages, where the first appears to be the Albanian with a total of 703.192 characters, which has 36 letters in the alphabet, followed by the Bosnian language with 665.840 characters, which has 30 letters in the alphabet, where as third appears the Turkish with a total of 625.037

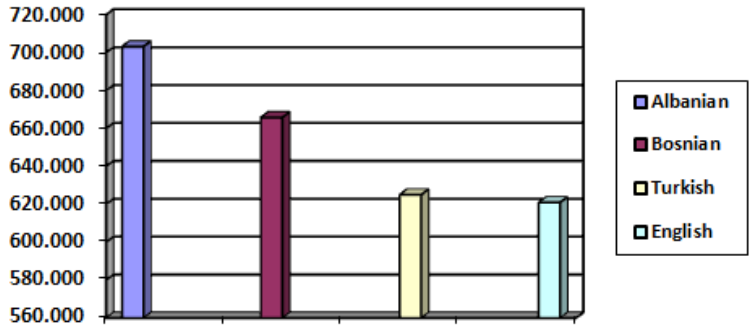
characters, and which has 29 letters in the alphabet, and in the end appears to be the English language with 620.920 characters with an alphabet consisting of 26 letters.

Figure 6: Number of Letters in Alphabets



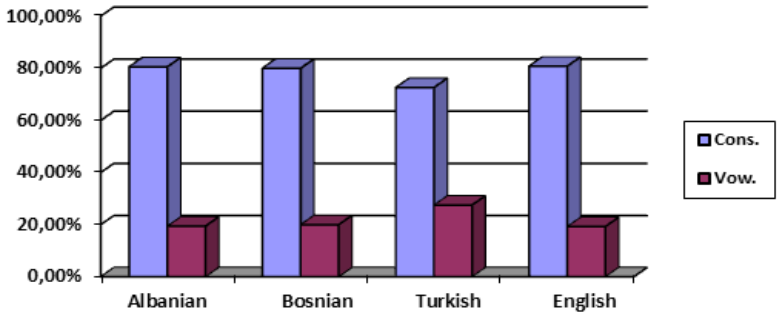
While analyzing these four languages and their distribution of characters, it is important to note that the study also reviews punctuation marks, spaces and other symbols (numbers, special signs, etc.). In this case, the study shows that the distribution of these symbols varies significantly between the four languages. Albanian in the examined text contains 2.44% of punctuation marks, Turkish has 2.62%, Bosnian with the highest percentage of punctuation marks with 2.77%, and English with least, around 2.21%. Furthermore, regarding spaces used the results are as follows: Albanian uses 16.99%, Bosnian luses 16.53%, Turkish uses 16.15%, and English uses approximately 16.02% of space. This helps us to understand that for the same content the Albanian uses more words, with Bosnian as second, followed by Turkish, and English in the end, by what we conclude that this sequence is the same as the one with the number of letters in the alphabets.

Figure 7: Number of Used Characters



During the analysis of these four alphabets, it is noted that in terms of using the vowels and consonants, they are not very compatible. Albanian, English and Bosnian languages have almost the same vowels "a, e, i, o, u", while Turkish has additionally vowels "ı , ö, ü". However, the distribution of these components is similar, wherein: the Albanian in the alphabet has 80.56% consonants and 19.44% vowels, Bosnian has 20% vowels and 80% consonants, Turkish has 27.48% vowels and 72.58% consonants, and the English has 80.77% consonants and 19.23% vowels, which are shown in the figure below.

Figure 8: Dispersion of Vowels and Consonants



Vowels are mostly used in Albanian text with 42.32% and the most used vowels appears to be “e, ë, I”. Second is Turkish with 41.02% vowels in text, where the most used vowels are: “e, a, o, I”. Third is English with 40.52% vowels in text, where most commonly are used "e, i". Fourth is Bosnian with 38.72% vowels in the text, where most commonly used vowels are: “a, i, e”. In this case it is noted that the Albanian, English and Turkish more frequently use vowel “e”, while Bosnian does not. The highest frequency in Albanian and English have letters E and T, with about 48.52%, in Bosnian vowels A and I have the highest frequency, around 25.02%, while E and A have the highest frequency in Turkish with 24.28%. In the four graphs below we show the frequency of letters for the four languages as well as the features provided by the application.

Figure 9: Frequency of Letters in Albanian

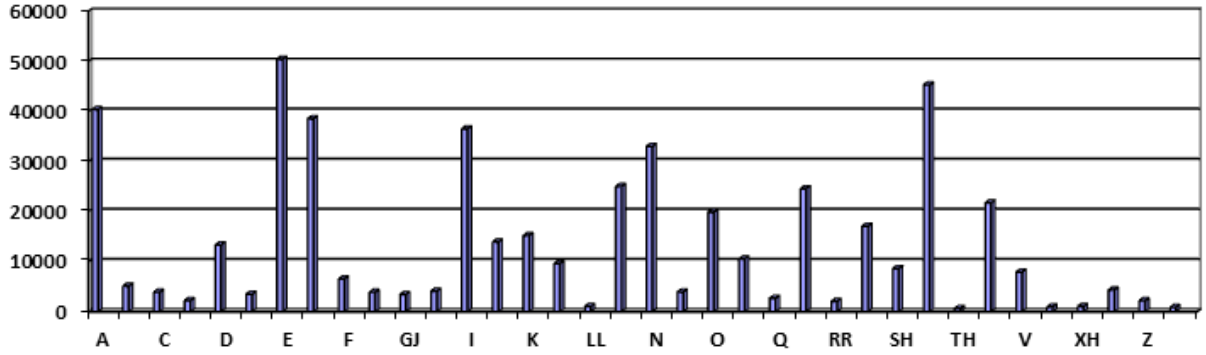


Figure 10: Frequency of Letters in English

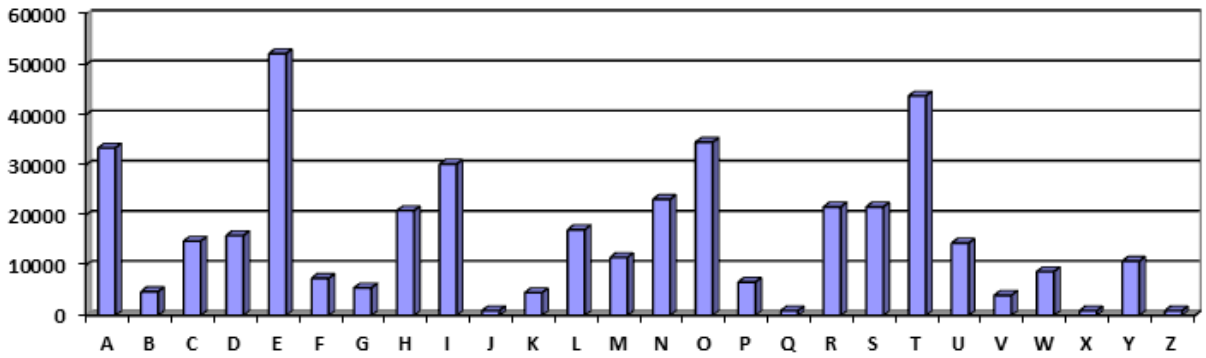


Figure 11: Frequency of Letters in Bosnian

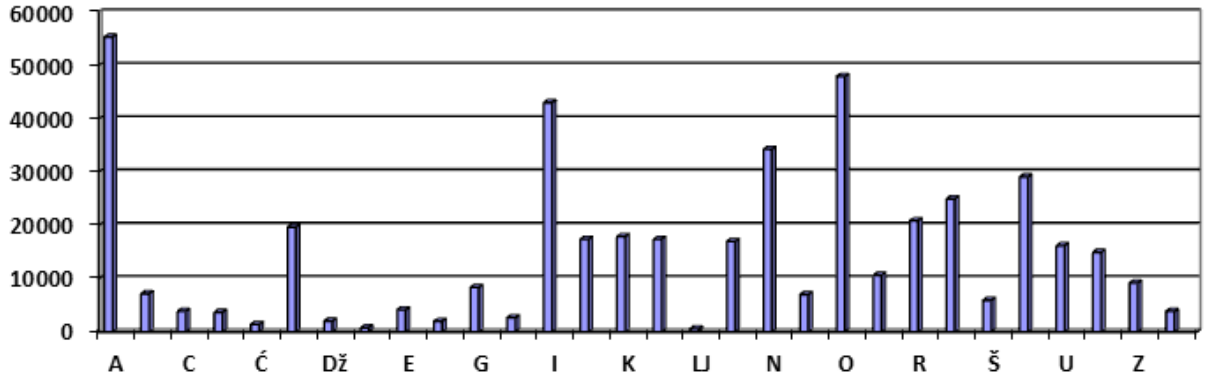
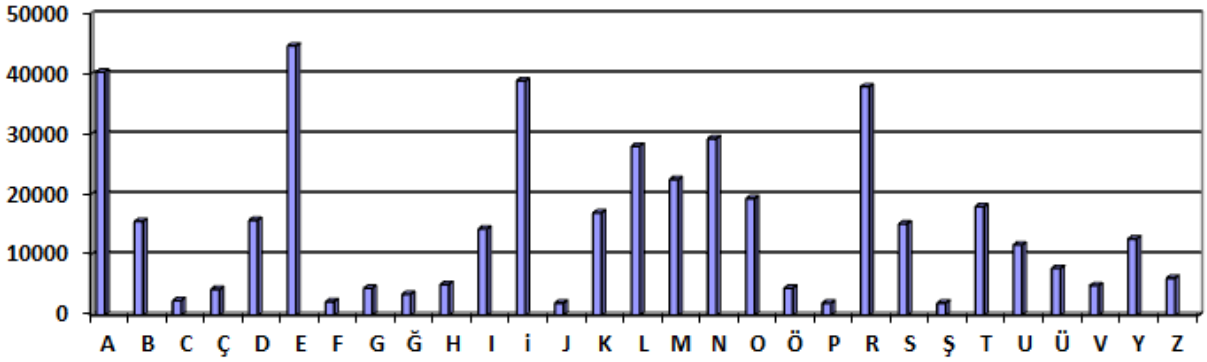


Figure 12: Frequency of Letters in Turkish



5. CONCLUSION

After conducting many statistical analyses on the same text, written in four different languages, Albanian, English, Bosnian and Turkish, we generated the following conclusions:

- Fewer characters were used in the text written in English i.e. the language that has fewer letters in the alphabet, 26 in total, followed by Turkish with 29 letters, Bosnian with 30 letters, and Albanian with 36 letters in the alphabet and that uses most characters.
- Punctuations in Turkish and Bosnian are almost equivalent with about 2.75%, and have direct relationship compared with the total number of characters.
- Spaces used in written Albanian occupy 16.99% of the text, which leaves us to understand the texts in Albanian use more words to interpret the same content, followed by the Bosnian 16.53%, Turkish with 16.15% and finally English with 16.02%.
- The ratio between vowels and consonants in Albanian is equivalent to 80.56% with 19.44%, in Bosnian is 80% to 20%, Turkish has a ratio of 72.58% to 27.42% and English language has a ratio of 80.77% by 19.23 %.
- In Albanian, English and Turkish, vowel “e” is used more frequently, whereas in Bosnian language most frequent vowel is “a”.
- All four languages have a single letter with a very low frequency of usage. In Albanian it is “x”, in Bosnian "đ", in Turkish “o” and in English the letter “z”.
- The application realized in C# meets the requirements for the languages set in the study and it is easy for use and modification from other researchers based on the requirements for specific languages.
- Also the application is offered by the authors as an open source for all concerned in continuing the research in the field of computational linguistics for educational purposes.

REFERENCES

Alexander, R. (2010). Bosnian, Croatian, Serbian, a Textbook: with Exercises and Basic Gramma. Sarajevo.

Attaoullah, F. A. (2010). Beginner's Turkish (Beginner's (Foreign Language)). Izmir.

Dodi, A. (2011). Fonetika dhe fonologjia e gjuhës shqipe. Tiranë: Akademia e shkencave në Shqipëri.

Hamiti, A. (2005). Fonetika dhe fonologjia e gjuhës standarde shqipe. Shkup: Interdiskont.

Hamiti, M. (2015). Text Analyzer.

Letërsisë., I. i. (1976). Fonetika dhe gramatika e gjuhës së sotme letrare shqipe. Tiranë: Akademia e Shkencave e Republikës së Shqipërisë.

Mentor, H. (2015). Analyses of Same Content Texts written in Different Languages.

Murphy, R. (2008.). English Grammar In Use. Third edition. Cambridge University Press.

PHRASE, W. A. (n.d.). Frequency of Letters in English Language . Retrieved 12 23, 2015, from //www.wordandphrase.info/frequencyList.asp

Pollard, A. Ç. (2001). Teach Yourself Turkish1. Istambul.

Schumann, J. (2010). Bosnian For Beginners: A Book In 2 Languages. Sarajevo: Bilingual.

Underhill, R. (2010). Turkish Grammar (Turk dili grameri, dil, Turk dili, Turkce grameri) (English and Turkish Edition).