Quantitative Characteristics of Key Words in Texts of Scientific Genre (on the Material of the Ukrainian Scientific Journal)

Uliana Shandruk[0000-0003-4796-1557]

Lviv Polytechnic National University, 12 Bandera street,
Lviv, Ukraine, 79013, shandruk.uliana@gmail.com

Abstract. The key words in a corpus linguistics statistically deals with the words or word combinations that occur in text more frequently than the others. Each style of the language is characterized by different set of words that makes it stand out among others and easily be referred to. The reason behind this is to convey the relevant information. The scientific style of the Ukrainian language is not the exception in such a respect. It possesses the specific vocabulary that allows conveying the needed information in the most objective, accurate and justified way. In the scientific style, text is most representative language unit as all researches and recent findings are presented in the form of text, so it serves a rich source for a research.

Keywords: frequency, key words, corpus linguistics, scientific texts, system, analysis, research, scientific journal.

1 Introduction

Key words are considered to be the most frequently occurring words in text. Their main function is to indicate the ‘aboutness’ of a particular text or corpus [7] and refer text to a particular style of genre. They could provide a comprehensive understanding of the topic, that is why recently they have been of a high interest of the applied linguistics, corpus linguistics and stylistics.

According to Kintsch and van Dijk [3] any statement that is referred to repetitively should be of a higher importance than the others. Such scholars as Berber [2], Scott [7], Tribble [9], Lazinski [4], Tarasheva [8] study the notion of key words. The researchers believe that if the particular word often repeats in text, it means that it has more importance than the other words. And it is also believed that key words can reveal more information than the other words or word combinations that occur in text. That is why a list of key words has been an inevitable part of any scientific article. It is given right after the abstract aiming to briefly introduce the topic to the reader as well as to facilitate the categorization of articles in general. The object of the research is key words in the scientific texts. The subject is the investigation of the frequency of key words in a scientific paper. The aim is to find out what are the most frequent words used by the Ukrainian scientists and check whether these the most frequent words are included in abstract and the key words section of the scientific paper to properly convey the aboutness of a research. All calculations presented in the paper were made with AntConc [1] and a set of author's programs written in Python.

2 The research corpus

The material of the research is the scientific journal 'Information systems and networks' published by the department of the Lviv Polytechnic National University 'Information systems and networks' from 2009 to 2016. Since 2009 this scientific journal has received a standard structural formatting and division according to subject headings. A total of 391 articles were processed, which were divided into three sections: 'Information systems, networks and technologies' (further in the text ISN) – 235 articles, which generally contain about 742 000 words, 'Computational and mathematical linguistics' (CML) – 101 articles, which generally contain close 352 000 words, 'Project and program management' (PPM) – 55 articles, which in total contain about 171 000 words.

3 Key words in the structural units of the research corpus

The first step towards the organization and preparation of the research corpus, was decided to divide the selection into structural units such as: text, abstract, the list of key words. That was the most natural division as all the researched articles had such a structure. The next step was to identify the key words in each structural unit. The percentage of the most frequent words that have been encountered in the structural units of each group of texts is presented below:

Table 1. The percentage of key words in the structural units of each section.

Structural unit	**ISN**	**CML**	**PPM**
Text	7,3%	8,0%	8,8%
Abstract	15,8%	18,4%	17,0%
Key words	23,9%	26,0%	32,7%

It is no surprise that the highest percentage of the most frequent words is presented in the list of the key words and in abstract. The number of the most frequent words in text itself is at least 3 times lower than in the abovementioned sections. From one side, it can be concluded that the most frequent words are really statistically the most frequent in those sections where they had to be (taking into account the overall number of words presented in all three structural units in general), although it should be noted that the expected results for the list of key words were higher than the actual ones. So, probably the authors, who were publishing the results of their researches in the scientific journal 'Information systems and networks', when it came to the key words list compilation, were not thorough enough to make sure they provided the specific key words where they were meant to be.

The next step of the research was to define the most frequent word for each group of articles in general with the reference to each structural unit.

Table 2. The most frequent word in the structural units of the section Information systems and networks (ISN).

Structural unit	Word	Quantity	Frequency
Text	система (system)	7011	0,96%
Abstract	система (system)	233	2,23%
Key words	система (system)	102	4,02%

For the group of texts ISN the most frequent word is a word система (system). It is the most frequent one in all structural units. It means that the authors accurately reflected the key word of the research both in abstract and key words list.

Table 3. The most frequent word in the structural units of the section Computational and mathematical linguistics (CML).

Structural unit	Word	Quantity	Frequency
Text	слово (word)	2287	0,66%
Abstract	метод (method)	67	1,72%
Key words	інформаційний (informational)	32	2,68%

In the second group of texts, the most frequent in the word слово (word) for the structural unit text, метод (method) for the structural unit abstract and інформаційний (informational) for the structural unit key words. Hence, it can be concluded that the texts under the topic 'Computational and mathematical linguistics', the key word is not that obvious as in the first group of texts. It differs in different structural units. In this case it is also interesting to trace the aboutness of the articles of this group as computational and structural linguistics is about words, methods and information.

Table 4. The most frequent word in the structural units of the section Project and program management (PPM).

Structural unit	Word	Quatity	Frequency
Text	дані (data)	1237	0,74%
Abstract	інформаційний (informational)	42	1,40%
Key words	інформаційний (informational)	29	4,50%

In the third group of texts, the tendency is quite different as we can see that the key word in the structural unit text is дані (data) while this word cannot be observed in such structural units as abstract and key words what can lead to the conclusions that the authors do not convey an accurate analysis of their key words and their abstracts and key words lists may lack accurate information.

4 The intersections of the sets of the most frequent words

The next step was to find out the most frequent words in each group of texts and how they coincide following structural units:

- intersection of text with abstract (BiAi)
- intersection of text with key words section (BiKi)
- intersection of abstract with key words section (AiKi)
- intersection of text with abstract with key words section (BiAiKi)

The aim of this research was to find out if the words that experimentally proved to be the most frequent words are the key words in the group of texts ‘Information Systems and Networks’ (ISN group), ‘Computer Science’ (CML group) and ‘Project and program management’ (PPM group).

The experiment was carried out in the following way:

1. The most frequent words for each group of texts were counted.
2. Each of them was assigned a rank from 1 to 25. Ranks and frequencies of words are inversely proportional, meaning that the most frequent word has a rank 1 while the least frequent one has rank 25.
3. 25 the most frequent words were taken to illustrate the experiment.
4. The intersection of 25 most frequent words in the sets of text and abstract (BiAi), text and key words (BiKi), abstract and key words (AiKi), and text and abstract and key words were found (BiAiKi).
5. Such manipulation was done for three group of texts.

The results of the first group of texts is presented in the table below:

Table 5. The intersection table for the ISN group.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
BiAi	🟨	🟨	🟨			🟨		🟨	🟨	🟨				🟨		🟨	🟨				🟨		🟨		
	🟨	🟨		🟨	🟨	🟨	🟨		🟨		🟨		🟨	🟨								🟨			🟨
BiKi	🟩	🟩	🟩			🟩		🟩	🟩					🟩		🟩									🟩
	🟩	🟩	🟩	🟩	🟩		🟩						🟩		🟩										🟩
AiKi	🟦	🟦	🟦	🟦	🟦	🟦				🟦	🟦		🟦				🟦		🟦				🟦		
	🟦	🟦		🟦	🟦	🟦	🟦		🟦			🟦	🟦	🟦	🟦									🟦	
BiAiKi	🟧	🟧	🟧			🟧		🟧	🟧					🟧		🟧									
	🟧	🟧		🟧	🟧	🟧					🟧		🟧				🟧								
	🟧	🟧	🟧	🟧	🟧	🟧							🟧		🟧										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

From the results above it can be concluded that not all the most frequent words that were found experimentally are really the most frequent in a scientific text.

For example, only 12 words among 25 the most frequent both occur in text and abstract (in other words, the intersection BiAi is the set that contains all elements of B that also belong to A) and they have different ranks meaning that they have different frequency. These words are presented in the table in terms of their ranks.

The ideal picture would be if all cells from 1 to 12 are colored. Instead it can be observed that those words that occur in a intersection of sets (be it text and abstract or abstract and key words or others) are not necessarily the most frequent words in the selection.

The same experiment was done for all three group of text and the following results were obtained:

Table 6. The intersection table of the whole selection.

	B∩A	**B∩K**	**A∩K**	**B∩A∩K**
ISN	48%	36%	48%	32%
CML	36%	36%	48%	24%
PPM	52%	44%	60%	40%

From the results, it can be concluded that for the 1st group of texts ISN, approximately half of the most frequent words occur in both text and abstract, if to be more precise only 48%. Only 36% of the most frequent words occur in both text and key words section, only 48% – in abstract and key words sections. And only 32% of all the most frequent words fount experimentally are really the most frequent in text, abstract and key words section. Quite the same tendencies are observed in the 2d group of texts CML and 3d group PPM.

5 The most frequent words and their collocations

After finding out the most frequent words in each group of texts and in all intersections described above, it was decided to look at the collocations these the most frequent words occur in the texts. The most frequent words from the whole selection is presented below:

Table 7. 10 the most frequent words from the whole selection.

Word	**Frequency**
система (system)	0,76%
дані (data)	0,74%
аналіз (analysis)	0,37%
інформація (information)	0,37%
модель (model)	0,36%

контент (content)	0,32%
час (time)	0,28%
використання (use)	0,25%
кількість (quantity)	0,25%
значення (meaning)	0,23%

From the list of the most frequent words in the whole selection it is seen that the word система (system) and дані (data) come to the fore. They have relatively high occurrence comparing to the other most frequent words what shows that the researchers publish their findings in the 'Information systems and networks' describe some system and data, then they conduct the analysis, work with information, model or content. To go further and discover what collocations they use with the most frequent words, it was decided to find the total number of collocations with the most frequent words. The final stage of the research was to identify the collocations that frequently occur with the top 5 frequent words of the selection. The results can be seen in the tables below:

Table 8. The total number of collocations for 5 the most frequent words.

Word	The number of collocations
система (system)	20570
дані (data)	17416
аналіз (analysis)	8225
інформація (information)	9958
модель (model)	7588

To obtain the most comprehensive results, all the collocations (occurring both on the left and right sides) of the given word was taken into account and shown:

Table 9. 10 the most frequent collocations with the word SYSTEM.

Rank	**Wordform**	**Quantity**			**%**
		total	**left**	**right**	
1	електронної (electronic)	330	6	324	1,60%
2	управління (management)	289	9	280	1,40%
3	інформаційної (information)	286	286	0	1,39%
4	інформаційних (information)	251	249	2	1,22%
5	пошукових (search)	159	158	1	0,77%

6	опрацювання (processing)	145	3	142	0,70%
7	функціонування (functioning)	144	138	6	0,70%
8	роботи (work)	132	130	2	0,64%
9	підтримки (support)	116	1	115	0,56%
10	технічних (technical)	104	103	1	0,51%

Table 10. 10 the most frequent collocations with the word DATA.

Rank	Wordform	Quantity			%
		total	left	right	
1	бази (bases)	499	484	15	2,87%
2	баз (base)	290	288	2	1,67%
3	сховища (repository)	193	180	13	1,11%
4	база (basis)	160	154	6	0,92%
5	базі (basis)	151	151	0	0,87%
6	аналізу (analysis)	149	128	21	0,86%
7	опрацювання (processing)	142	136	6	0,82%
8	потоків (flow)	132	132	0	0,76%
9	вхідних (input)	121	120	1	0,69%
10	сховищ (reporsitory)	117	110	7	0,67%

Table 11. 10 the most frequent collocations with the word INFORMATION.

Rank	Wordform	Quantity			%
		total	left	right	
1	текстової (text)	96	96	0	0,96%
2	опрацювання (processing)	88	86	2	0,88%
3	містить (contain)	88	86	2	0,88%
4	захисту (protection)	77	76	1	0,77%
5	пошуку	62	61	1	0,62%

	(search)				
6	отримання (receive)	61	58	2	0,61%
7	необхідної (needed)	60	51	9	0,60%
8	подання (presentation)	46	45	1	0,46%
9	обміну (exchange)	46	46	0	0,46%
10	джерел (sources)	44	43	1	0,44%

Table 12. 10 the most frequent collocations with the word MODEL.

Rank	**Wordform**	**Quantity**			**%**
		total	**left**	**right**	
1	даних (data)	125	20	105	1,65%
2	інформаційної (informational)	59	39	20	0,78%
3	системи (systems)	58	2	56	0,76%
4	математичні (mathematical, pl)	46	45	1	0,61%
5	математичної (mathematical, sing)	39	39	0	0,51%
6	концептуальної (conceptual)	39	39	0	0,51%
7	оцінки (assessment)	38	12	26	0,50%
8	математичну (mathematical)	37	37	0	0,49%
9	побудови (building)	37	33	4	0,49%
10	предметної (subject)	37	0	37	0,49%

Table 13. 10 the most frequent collocations with the word ANALYSIS.

Rank	**Wordform**	**Quantity**			**%**
		total	**left**	**right**	
1	останніх (recent)	231	0	231	2,81%
2	даних (data)	200	25	175	2,43%

3	отриманих (received)	107	0	107	1,30%
4	контенту (content)	102	33	69	1,24%
5	основі (basis)	96	96	0	1,17%
6	результатів (results)	95	26	69	1,16%
7	кластерного (cluster)	74	74	0	0,90%
8	системи (system)	74	64	10	0,90%
9	тексту (text)	60	4	56	0,73%
10	морфологічного (morphological)	58	57	1	0,71%

The most frequent collocations are the following: electronic system, management system, information system, database, data repository, text information, data model, recent analysis and received analysis.

6 Conclusions

The main question the paper aims to answer is if the most frequent words in text could be considered as key words revealing the aboutness of this text, and whether the authors use these words in abstract and key words sections. The results showed that the list of key words they compile for their articles do not accurately reflect the aboutness of their researches, because these are not the most frequent words of their texts. Quite the same situation was observed with abstract. There are some discrepancies in terms of the most frequent words between the abstracts, key words lists and text itself. It can be concluded that when the author writes a scientific paper, he or she, of course, uses some words more frequently than others. This is a natural process as they outline some narrow issue. When they write the abstract and compile the list of key words, they do not necessarily remember about the importance to use the most frequent words from their researches in abstract and key words sections.

Generally, abstract and key words section must convey the meaning of the article, their function is to convey paper’s aboutness, this is why it is of high importance to include the most frequent words there, and therefore consider these words as key words. Such tendency was only observed partially allowing to conclude that only every 2d or 3d word in the abstract and list of key words are really paper’s the most frequent.

Hence, the modern Ukrainian researchers when describing their results, use slightly different key words in abstracts, the list of key words and text itself. They are still within the same topic, but these are different words what leads to the assumption that when dealing with abstract or the list of key words, the authors does not carefully analyze their text and make decision of which lexical units to use in abstract or the list of key words based on their feelings of the subject area.

Another thing raised in the paper is what are the most frequent words of the analyzed scientific journal and what is the hidden meaning to use such words. The most frequent word used in the whole selection is the word система (system). According to the definition [6], the word система (system) means order, or a set of principles or procedures according to which somethings is done; an organized methods or scheme. So, it can be concluded that the Ukrainian researchers who work in the fields of information systems, computational and mathematical linguistics and project and program management, tend to order and systematize things they investigate.

It is also interesting to mention that the second most frequent word is дані (data) assuming that the object of their researches in the most cases is data. So, further to conclude is that the modern Ukrainian scientists basically work with data. The word аналіз (analysis) is the third most frequent word allowing concluding that the preliminary goal of the scientists is to carry out an analysis of something. Among the most frequent words are also such as інформація (information), модель (model), контент (content), час (time), використання (use), кількість (quantity), значення (meaning). Although the results are already representative, the further work is definitely to be done on a larger selection and with a reference corpus.

7 References

1. AntConc: http://www.laurenceanthony.net/software/antconc/releases/AntConc343/help.pdf.
2. Berber Sardinha, T.: Wordsets, keywords, and text contents: an investigation of text topic on the computer, Delta 141-149. (1999).
3. Kintsch, W., van Dijk, T.: Toward a model of text comprehension and production. In: Psychological Review, vol. 85(5), 363-394. (1978).
4. Lazinski, M.: Key words in semantics and statistics. In: Biuletyn Polskiego Towarzystwa Językoznawczego, vol. 62, 57-68. (2006).
5. Medvedev M., Pashchenko I.: Teoriia imovirnostei ta metematychna statystyka, Lira, Kyiv, 536. (2008).
6. Merriam-Webster Dictionary: https://www.merriam-webster.com/dictionary/system.
7. Scott, M.: PC Analysis of key words-and key key words. In: System, vol. 25, 233-245. (1997).
8. Tarasheva, E.: Repetitions of Word Forms in Texts, Cambridge Scholars Publishing. (2011).
9. Tribble, Ch., Scott, M.: Textual Patterns: Key Words and Corpus Analysis. In: Language Education, vol. 11. (2007).