**IJCTA,** 9(30), 2016, pp. 21-32


# Application of Statistical Methods for Identifying Fixed Collocations in Texts in the Russian Language to Extract Information Objects Consisting of Two or More Words from News Flows

**Nikolai Valerievich Bradis* and Dmitrii Aleksandrovich Sytnik***

***Abstract :*** In the systems of the processing of information news flows, the priority task is to identify information objects, *i.e.* persons, organizations and geographical locations. The aim of this paper is the development and practical implementation of the method for the identification of information objects, the name of which containsmore than one word. Information objects that are uniquely defined by a single word can be identified using reference books or special dictionaries of geographical names, surnames and organizations of various level of detail. Currently, there is a successfully developing morphological analyzer Pymorphy 2, which, in addition to the options of morphological analysis of a wordwith the correspondent morphological information, makes it possible to receive information on the possible belonging of this word to one of the categories: surname, first name, patronymic, geographical location, organization. For the purpose of identifying information objects and named entities,methods based on rules, dictionaries, context-free grammars, andontologies are often used. The essence of the method discussed in this paperconsists in identifying multiwordinformation objects by means of identifying statistically stable word combinations (collocations), as well as in the identification ofobtained collocations as belonging to aninformation object of one of the listed types. This method is based on the use of statistical measures of inclusion of the types MI, T-score, Dice Factor, etc. Among fixed collocations obtained by this method, it is possible to identify those that potentially describe information objects. The practical implementation of this methodmade it possible to identify not only information objects, but also adjuncts of time, as well as multiword names of various inanimate objects.

***Keywords :*** News flow, statistical measure, information object, statistical method, text analysis, morphological analysis, T-score, MI, Dice Factor, automatic text processing, collocation, threshold, occurrence frequency.

## 1. INTRODUCTION

Description of the object of the study and the source data for the task of identifying information objects.

The extraction ofinformation objects from a flow of unstructured data in natural language is of some interest from the perspective of identifying the main actors and circumstances of the event described in the text, which can be used for clustering and structuring the data flows. The obtained results can also be used in expert systems and decision support systems to analyzethe data of news flows and other sources of textual information. The object of the study of this paper isinformation objects of the type “organization”, “person” and “geographical location”, consisting of more than one word.

As the source datafor the study, the authors used the data of the newsflows from the informationweb portals gazeta.ru, lenta.ru, rbk.ru, ria.ru and vedomosti.ru. A unit of a news flow (news) is understood to bean unstructured

* Complex Systems LLC Russia, 170021, Tver, Skvortsova-Stepanova Street, 83

message, published at a certain point in time (date and time of publication),from a particular source (media). Each news message should describe some event and contain information about the participants in this event. For the analysis, only news from political sections were selected. After downloading, plain text without html tags and any other service information was extracted from each news article. Then the selected text was saved in the database for further processing. All the news messagesfrom different feeds were selected within the same time interval, which makes it possible to conclude that they cover the same political events, which took place during this period.

**Description of the task**

The task of identifying fixed collocations is widely used in the tasks of information retrieval and text analysisfor further indexing, classification and search. Currently, there are a large number of systems for the extraction of keywords and phrases from texts in natural language. As a rule, these systems use either linguistic or statistical methods. Linguistic methods are based on the meaning of words and use ontologies and semantic information aboutwords [1], while statistical methods use numerical data on the frequency of occurrence of words in texts [2]. Linguistic methods are too labor-intensive and aregenerally focusedon a specific subject area, as the development and updating of ontologies takes considerable time.

The aim of this paper is to propose and test a method for extracting from news flowsinformation objects, the name of which consists of more than one word. To solve this problem, it is necessary to carry out the preprocessing of news texts for the purpose of identifying an array of word forms, removing stop words and stop characters, conducting a morphological analysis of the word forms of the text. After the preprocessing, it is required to define the boundaries ofpossible collocations [3] and to identify the most significant ones [4]. The significance of collocations is determined by the frequency of co-occurrence of its words. That is, the priority task is to identify frequent word combinations, the co-occurrence of whichis based on theregular nature of mutual expectation.

After identifyingsignificant collocations, it is necessary to determine their belonging to information objects: for this purpose, one cancheckif the words constituting the collocation belong to theseinformation objects. For this purpose, one can use the morphological analyzer Pymorphy2 [5], which makes it possible to receiveinformation on the belonging ofa certain wordto aninformation object; besides, onecan also use words that refer toinformation objects of the type “organization” or “geographical location”. The belonging of a collocation to an informationobject can be determined by checking whether it contains a pointer word which refers to theform of ownership of the organization, or by the presence of one of the following words in the collocationor near it: ministry, fund, bank, department, integrated plant, etc. The belonging of a collocation to an information objectof the type “geographical location” can be determined by the presence of one of the words in the collocation or near it: city, street, peninsula, island, state, region, lane, etc.

## 2. METHOD FOR IDENTIFYING STATISTICALLY STABLE WORD COMBINATIONS (COLLOCATIONS)

General description of the method

To identify collocations, various statistical measures can be used [6]. In this paper, three measures are considered: Measure MI (mutual information) [7], Measure T-score [8], Dice Measure [9].

1. **Measure MI (mutual information) :** David Magerman and Mitchell Marcus were among the first researchers to use the MI measure for processing texts in natural languages. This statistical measure is suitable for identifying specific terms with low frequency of occurrence, fixed collocations and proper names. This measure is based on the ratio ofcontext-dependent frequencies to independent context-free frequencies:

$$\mathrm{MI} = \log_2 \frac{f(a,b) \times \mathrm{N}}{f(a) \times f(b)} \qquad (1)$$

where :

$a, b$ – terms (hereinafter a term is understood to mean a single word or a phrase),
$f(a), f(b)$ – absolute frequencies ofoccurrence of the terms a and b in the text corpus,
$f(a, b)$ – thefrequency ofoccurrence of the term a together with the term *b*,
N – the total number of word forms in the text corpus.

2. **T-score Measure :** The T-score measure, as well as the MI measure,takes into account the frequency of co-occurrence and independent occurrence of the words of the collocation and makes it possible to determine the degree of nonrandomness of the strength of bonds between words in the collocation:

$$\text{T} - \text{score} = \frac{f(a, b) - \frac{f(a) \times f(b)}{\text{N}}}{\sqrt{f(a, b)}}, \quad (2)$$

where :

$a, b$ – terms (hereinafter a term is understood to mean a single word or a phrase),
$f(a), f(b)$ – absolute frequencies ofoccurrence of the terms a and b in the text corpus,
$f(a, b)$ – thefrequency ofoccurrence of the term a together with the term *b*,
N – the total number of word forms in the text corpus.

3. **Dice Measure :** The Dice measure, also known as Sørensen-Dice index, wasinitially intended foridentifying the interdependence of the two charactersbetween species, but later found application for determining the binding of words in collocations:

$$\text{Dice} = \frac{2 \times f(a, b)}{f(a) + f(b)}, \quad (3)$$

where :

$a, b$ – terms (hereinafter a term is understood to mean a single word or a phrase),
$a, b$ – terms (hereinafter a term is understood to mean a single word or a phrase),
$f(a), f(b)$ – absolute frequencies ofoccurrence of the terms a and b in the text corpus,
$f(a,b)$ – the frequency ofoccurrence of the term a together with the term *b*.

The use of the data of statistical measures is based on the fact that for every potential collocation a value of themeasure is calculated, and, if the obtained value exceeds the threshold value, the collocation enters the list of significant collocations.

Potential collocations are only combinations of words located in close proximity to each other and not separated by punctuation marks, prepositions, and any other words. A potential collocation may include only nouns, adjectives, proper names, or the words not recognized by themorphological analyzer. Verbs and auxiliary parts of speech are not included in potential collocations, since they are extremely rare in the names of organizations andgeographical locations and never occur in firstnames, surnames and patronymics of persons. The restrictions, imposed on collocations, depend on the sources, from which it is required to retrieve the data, and the type of data, which it is necessary to retrieve [10].

**Preprocessing of texts**

Before extracting collocations from the text, it is necessary to divide the analyzed text into words, remove stop-wordsand punctuation marks, carry out themorphological analysis of the remaining word forms, to determine the boundaries of potential collocations and extract single terms, which can be contained in collocations.

The text is divided into words by punctuation marks andspaces; forcompound words (for example, “blue-green”), it is taken into account that the symbol “-” separatestwo words only when there are spaces on both sides.

To remove stop words, a preformed list of stop wordsis used;each word after morphological analysis is checked for inclusion in this list. Stop words arethe words in atext which have no meaning. Stop words or noise words include prepositions, participles, interjections, particles, punctuation marks and introductory words.

After removing the noise words, it is necessary toconduct a morphological analysis of the words in the text for the purpose of identifying each word in the text and counting the frequency of its occurrence, regardless of the form in which that word occurs. There are two types of morphological analysis: stemming and lemmatization.

The essence of stemming is to identify an uninflected word stem by cutting off all changeable parts (endings, prefixes and suffixes). This method is well suited for the English language, but for languages with complex word building, in particular, for the Russian language, it is not very useful. The most famous stemming algorithm was proposed by Martin Porter in 1980[11] and later adapted by him for various Indo-European languages, including Russian. The essence of lemmatization is reducing a word form to the lemma, *i.e*. to a normal dictionary form.

The comparison of the correctness of the work of the algorithms of stemming and lemmatizationwas performed by Mikhail Korobov by comparing Mystem 3.0 and Pymorphy2 [12]. The comparison showed that, depending on the type of source data and words to be parsed, both implementations have both drawbacks and advantages.

The proposed method of morphological analysis uses the morphological analyzer Pymorphy2 [5], which is based on the method of lemmatizationand returns the normal form of theword, as well as the accompanying morphological information on it (part of speech, gender, number, case) and the information on the possible belonging of the word to one of the types of information objects:geographical location, organization, surname, first name or patronymic name. As a rule, Pymorphy2 returns several options of parsing for one and the same word; therefore, all parsing options should be taken into account and,if at least one option makes it possible to include the word in thepotential collocation, this word cannot be excluded from consideration.

In the process of morphological analysis and removing stop words and punctuation marks,tuples of single terms are built. A tuple will be referred to as a continuous sequence of single terms that can be included in the collocation (which are notstop words, punctuation marks, auxiliary parts of speech or verbs). Besides, foreach single term it is counted how many times it occurred in the text corpus (occurrence frequency).

**Identifying collocations**

At the initial stage, from the texts, identified at the preprocessing stage, a list of potential collocations, consisting of two single terms, *i.e*. collocations of the length 2, is formed. Potential collocations of the length 2can be also built at the preliminary stage in the process of forming tuples of single terms, which positively affects the performance of the method. In addition to identifying collocations,the frequency of their occurrence is also counted. For all of the obtained word combinations of the length 2, the value of a statistical measure is calculated, and thosepotential collocations, for which the value of the measureexceeds the threshold value, are added to the list of significant collocations. After that, all potential collocations of the length 2 are merged with the adjacent single terms from the same tuples, and a list of potential collocations of the length 3 is formed (with their occurrence frequencies), which in turn are also tested for significance. The algorithm is repeated until the restriction on the maximum length of a collocation is achieved, or no potential collocations for the achieved length are found.

To improve the performance of the method, in the process of the formation of the list of potential collocations of the length *n*,only significant collocations of thelength *n*-1 can be used. However, in this case a part of significantcollocations of the length *n* can be lost, since according to the above formulas of various statistical measures, the value of a measurefor a collocation does not depend directly on the value of measures forsmaller collocations which are included in it. Let us show it using the example of the Dice measure.

$$\text{Dice}\,(ab, c) = \frac{2 \times f(ab, c)}{f(a, b) + f(c)}, \tag{4}$$

From (3), it follows that:

$$f(a, b) = \frac{1}{2} \times \text{Dice}\,(a, b) \times (f(a) + f(b)). \tag{5}$$

By substituting the expression from (5) into (4), we obtain:

$$\text{Dice}\,(ab, c) = \frac{2 \times f(ab, c)}{\frac{1}{2} \times \text{Dice}\,(a, b) \times (f(a) + f(b)) + f(c)} \tag{6}$$

From (6), it follows that the measure of the collocation of the length *n* is inversely proportional to the measure of its constituent collocation of smaller length.

To identify significant collocations, the implemented method uses all three of the above-mentioned statistical measures. Depending on the settings of the method, a collocation may be considered significant, if any of its measuresmeets the corresponding threshold value, or if thevalue of a particular measuremeets the threshold value of this measure, orif the values of all measures meet the corresponding threshold values. By means of experiments, a threshold value was selected for each of the measures. For MI and T-score measures, thethreshold value was determined to be equal to 1, and for the Dice measure –1/2. The issue of the selection of the threshold value for each of the measures is still open, and there are various hypotheses [13]. The use of the T-score measurehas a disadvantage consisting in identifying collocations with very frequent words, which often include functionalwords (prepositions, conjunctions, etc.), but this problem is solved in the proposed method by theprevious removal of the stop words.The main disadvantage of the Dice measure is that it does not take into account the volume of the text corpus.

In addition to using measures of statistical significance, the method implies imposing a restriction on theoccurrence frequencyfor a collocation, in order tocut off rare and insignificant collocations. For this purpose,it is possible to specify a minimum threshold of the frequency of occurrence of the collocation in the texts of theprocessed news articles. During the experiments, conducted in the course of theimplementation of this method, thecollocations, occurring two or less times in the texts of all articles to be processed, were eliminated.

## 3. RESULTS

The result of the work is the software implementation of the described method and its validation on real news flows; besides, the comparative evaluation of the statistical measures, used in this method, was carried out. For the implementation of the method, the Python programming language was selected. For this programming language, there is a package of libraries NLTK for the symbolic and statistical processing of texts in natural languages[14].

The method was tested on various amount of news (from 5 to 1000) on political topics for the period from October 2015to March 2016fromnews flows gazeta.ru, lenta.ru, rbk.ru, ria.ru and vedomosti.ru. To test the performance of the method,100 news for October-November 2015 were selected and processed; as a result, 502 significant collocationsconsisting of 2 or 3 words were identified, from which the system selected 103 collocations, related to various information objects, and 14 most significant dates, for example, September 30, 2015(the beginning of the Russian military operation in Syria). The small size of the test sample is associated with the fact that this amount of articles can be analyzed by human with the purpose of the manual validation of the correctness of the work of the method. Table 1 shows the first 16 significant collocations, which were revealed by the program, in descending order of occurrence frequency.

**Table 1. The results of applying the method, sorted by the occurrence frequency.**

| *Collocation* | *Frequency* | *Dice* | *MI* | *T-score* |
|---|---|---|---|---|
| Islamic State | 32 | 0.61 | 5.73 | 5.55 |
| Vladimir Putin | 31 | 5.23 | 5.42 | 5.23 |
| RIA News | 26 | 6.69 | 5.04 | 6.69 |
| Bashar al-Assad | 22 | 6.64 | 4.64 | 6.64 |
| September 30 | 21 | 6.033 | 4.51 | 6.03 |
| Head of state | 19 | 0.25 | 3.70 | 4.02 |
| Dmitry Peskov | 18 | 0.53 | 0.68 | 1.59 |
| Terrorist group | 12 | 0.50 | 6.35 | 3.42 |
| Coalition forces | 12 | 0.41 | 0.78 | 1.45 |

| *Collocation* | *Frequency* | *Dice* | *MI* | *T-score* |
|---|---|---|---|---|
| Russian leader | 11 | 0.14 | 3.13 | 2.93 |
| State Duma deputy | 11 | 0.28 | 4.82 | 3.19 |
| Air operation | 10 | 0.33 | 5.67 | 3.10 |
| President of Syria | 10 | 0.09 | 1.96 | 2.35 |
| A321 airbus | 10 | 0.90 | 8.32 | 3.15 |
| Federation Council | 10 | 0.42 | 6.19 | 3.11 |
| Sinai Peninsula | 9 | 0.51 | 7.22 | 2.97 |

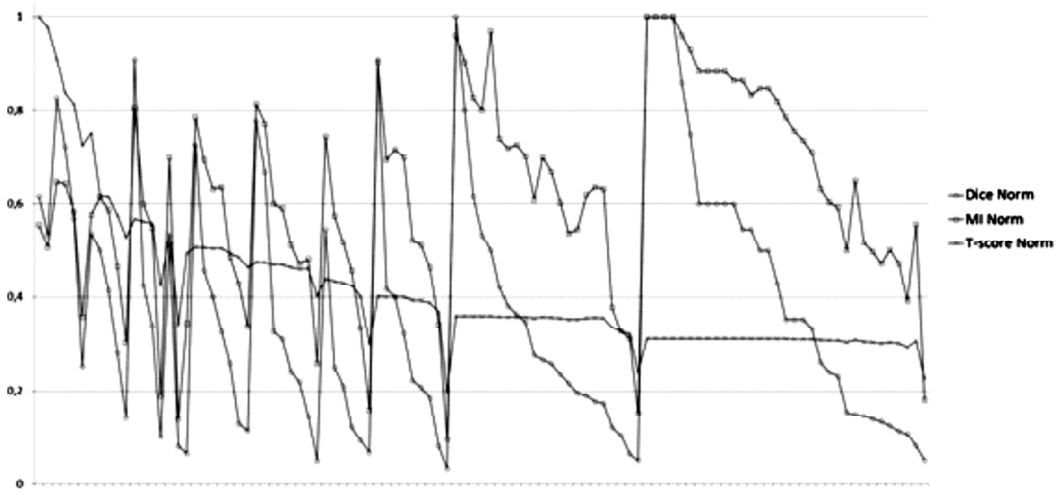


**Fig. 1. Graphs of the values of statistical measures calculatedfor the identified information objects in the descending order of their occurrence frequency.**

In the covered period,the Russian military operationin Syria began;therefore, the result of applying the method, which identified among the main information objects the Islamic State, Vladimir Putin, Bashar Al-Assad, the Federation Council, the Coalition forces and Dmitry Peskov, is quite logical. In the same period, the crash of the aircraftwith Russian tourists over theSinai Peninsula took place, which was bound to be reflected in mass media and, consequently, in the result of the work of the algorithm. The events in the East of Ukraine and the Minsk agreements were also reflected in the results of the work of the algorithm.

This method is aimedto identify only those information objects that consist of more than one word, whichexplains the relatively small amountof identified objects.

Figure 1 shows that all three statistical measures have similar ascending/descending domains, but the T-score has the minimum dispersion, which confirms its main disadvantage consisting in that the value of the T-score measure is strongly influenced by the occurrence frequency of a collocation. From the formulas (1), (2) and (3), it may be concluded that all three measures directly depend on the occurrence frequency of a collocation and are inversely dependent on the values of the occurrence frequencies of its constituents. By means of simple calculations, it is possible to express theT-Score measurethrough the MI measure.

From (1), it follows that : $$2^{\mathrm{MI}} = \frac{f(a,b)\times \mathrm{N}}{f(a)\times f(b)} \tag{7}$$

$$\mathrm{T-score} = \sqrt{f(a,b)} - \frac{f(a)\times f(b)}{\mathrm{N}\times\sqrt{f(a,b)}} = \sqrt{f(a,b)} - \sqrt{f(a,b)}\times\frac{f(a)\times f(b)}{\mathrm{N}\times f(a,b)} \tag{8}$$

From (2), it follows that: $$\mathrm{T-score} = \sqrt{f(a,b)}\times\left(1-\frac{1}{2^{\mathrm{MI}}}\right) \tag{9}$$

From (7) and (8), it follows that :

$$\text{T} - \text{score} \;=\; \sqrt{f(a,b)} \times \left(1 - \frac{1}{2^{\text{MI}}}\right) \tag{9}$$

**Table 2. The results of applying the method, sorted by thevalue of the Dice measure.**

| *Collocation* | *Frequency* | *Dice* | *MI* | *T-score* |
|---|---|---|---|---|
| National Unity Day | 4 | 1.00 | 9.92 | 1.99 |
| Caspian Flotilla | 3 | 1.00 | 10.34 | 1.73 |
| South China Sea | 3 | 1.00 | 10.34 | 1.73 |
| Saddam Hussein | 3 | 1.00 | 10.34 | 1.73 |
| Mustafa Dzhemilev | 3 | 1.00 | 10.34 | 1.73 |
| Airbus A321 | 10 | 0.90 | 8.32 | 3.15 |
| Right Sector | 5 | 0.90 | 9.34 | 2.23 |
| German Chancellor | 3 | 0.85 | 9.92 | 1.73 |
| RIA News | 26 | 0.82 | 6.69 | 5.04 |
| Lada Vesta | 4 | 0.80 | 9.34 | 1.99 |
| Valdai Discussion Club | 7 | 0.77 | 8.41 | 2.63 |
| Investigative Committee of Russia | 3 | 0.75 | 9.60 | 1.72 |
| Bashar al-Assad | 22 | 0.72 | 6.64 | 4.64 |
| Minsk agreements | 7 | 0.66 | 7.97 | 2.63 |
| Verkhovna Rada | 4 | 0.61 | 8.53 | 1.99 |
| Islamic State | 32 | 0.61 | 5.73 | 5.55 |

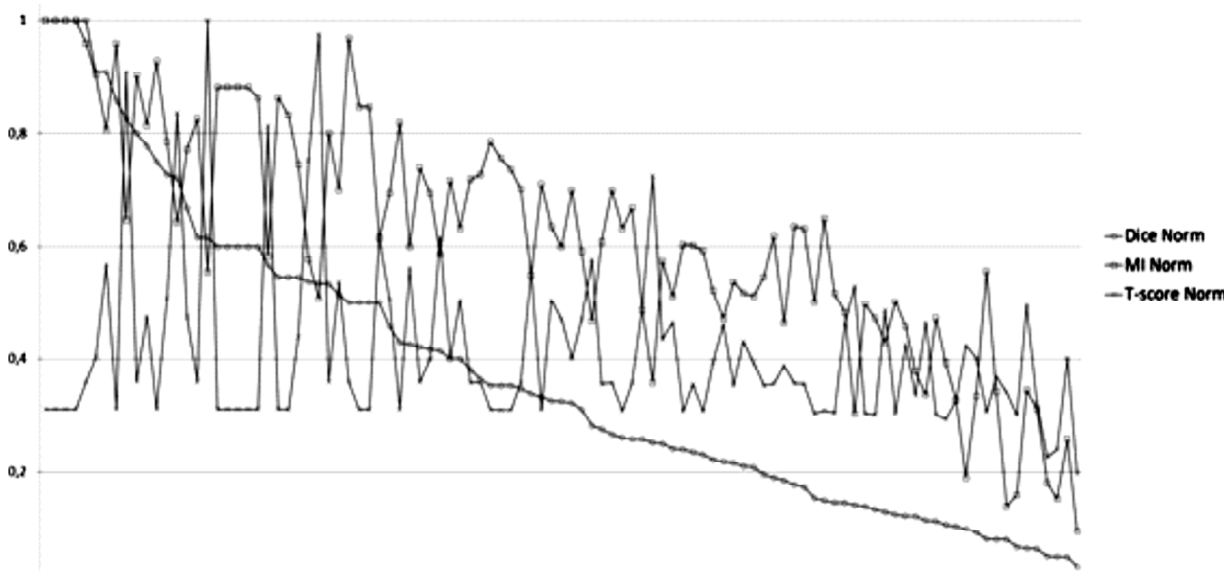


**Fig. 2. Graphs ofthe values of statistical measures calculatedfor the identified information objects in the descending order of the value of theDice measure.**

Table 2 shows the results for the first 16 significant collocations in the descending order of the value of theDice measure. From Table 2, it can be seen thatthe Dice measure is mainly influenced by therarity of the occurrence of the words, included in the collocation, outside this collocation.

**Table 3. The results of applying the method sorted by thevalue of the MI measure.**

| *Collocation* | *Frequency* | *Dice* | *MI* | *T-score* |
|---|---|---|---|---|
| Caspian Flotilla | 3 | 1.00 | 10.34 | 1.73 |
| South China Sea | 3 | 1.00 | 10.34 | 1.73 |
| Saddam Hussein | 3 | 1.00 | 10.34 | 1.73 |
| Mustafa Dzhemilev | 3 | 1.00 | 10.34 | 1.73 |
| Unrecognized Crimean Tatar Majlis | 4 | 0.50 | 10.01 | 1.99 |
| National Unity Day | 4 | 1.00 | 9.92 | 1.99 |
| German Chancellor | 3 | 0.85 | 9.92 | 1.73 |
| Investigative Committee of Russia | 3 | 0.75 | 9.60 | 1.72 |
| Right Sector | 5 | 0.90 | 9.34 | 2.23 |
| Lada Vesta | 4 | 0.80 | 9.34 | 1.99 |
| Angela Merkel | 3 | 0.60 | 9.11 | 1.72 |
| Natalia Poklonskaya | 3 | 0.60 | 9.11 | 1.72 |
| Verkhovna Rada | 3 | 0.60 | 9.11 | 1.72 |
| Nikolai Arefyev | 3 | 0.60 | 9.11 | 1.72 |
| Ivan Nikitchuk | 3 | 0.60 | 8.92 | 1.72 |
| Russia 24 TV channel | 3 | 0.54 | 8.92 | 1.72 |

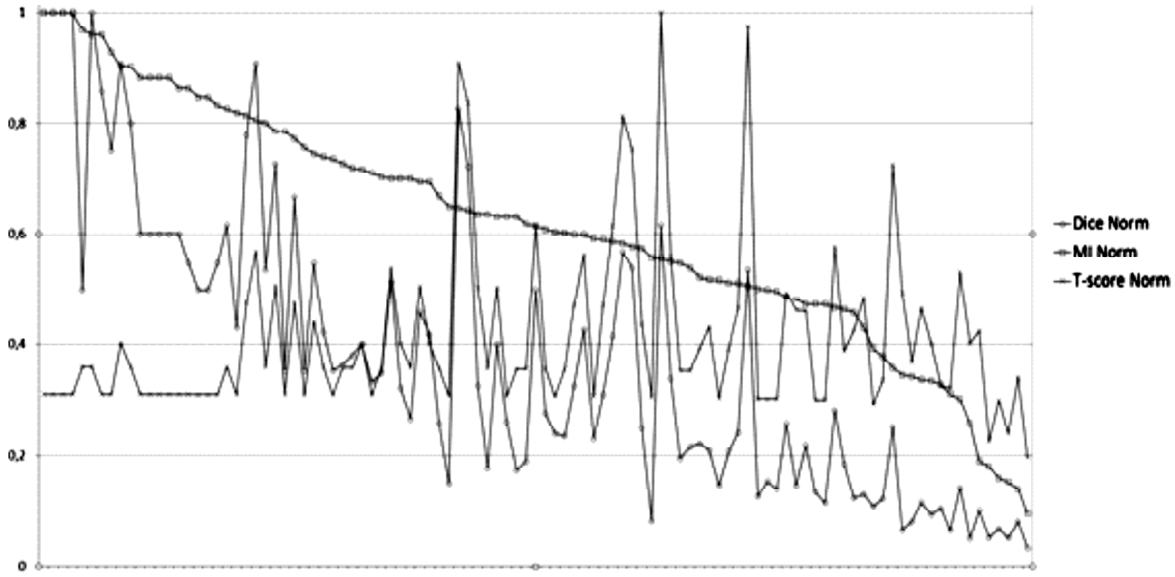


**Fig. 3. Graphs ofthe values of statistical measures calculatedfor the identified information objects in the descending order of the value of theMI measure**

Table 3 shows the first 16 significant collocations in the descending order of the value of theMI measure. On the basis of the analysis of the data of Tables 2 and3 and the graphs presented in Figures 2 and 3, it may be concluded that the Dice and MI measures are less influenced by theoccurrence frequency of collocations as compared with the T-score measure, and the results of the use of the Dice and MI measures are closer to each other than to the results of the use of the T-score measure.

The impact of the occurrence frequency of a collocation on the value of the T-score measure and the impact of the occurrence frequencies of the words, included in the collocation, on the MI and Dice measures can be

confirmed by the example of the collocation "Vladimir Putin". This collocation issecond in terms of the occurrence frequency andsecond in terms of the value of theT-score measure; at the same time,the occurrence frequency of the word "Putin" separately from the word "Vladimir" is very high, therefore, thecollocation "Vladimir Putin" ranks 28th in terms of the value of the Dice measure and 74th in terms of the value of the MI measure.

For the purpose of more correct evaluation of the method performance,163 news for the period of June 2-16, 2016were selected and processed; as a result, 687 significant collocationsof the length 2 and 3 were identified, from which the system selected 134 collocations, related to various information objects. Since the active phase of the Russian operation in Syria was already over in the selected period, it is not surprising thatthe articles, published within the specified period, payed more attention toforeign policy events not directly related to Syria and internal political events in the country. In the news flows,the following events were covered: the election race in the United States involving Donald Trump and Hillary Clinton, the controversy surrounding Germany's recognition of the Armenian genocide by Turkeya century ago, the expansion of the North Atlantic Treaty Organization (NATO) in the Black Sea and East Europe, the rioting during the European Football Championship in France, negotiations between Turkey and the European Union on the abolition of visa regime, the resolution of the crisis in the East of Ukraine and the negotiations of the Normandy Format, disputes in the society around the idea ofrenaming a bridge over one ofSt. Petersburg canalsafter Akhmad Kadyrov, the prank call on the Russian ballerina Anastasia Volochkovaand other events. These events were also reflected in the results of the application of the statistical method of identifying information objects, consisting of more than one word, which is demonstrated in Tables 4, 5 and 6.

**Table 4. The first 16 significant collocations in the descending order of the value of the T-score measure, extracted from news flows for the period of June 2-16, 2016.**

| *Information object* | *Frequency* | *T-score* |
|---|---|---|
| Vladimir Putin | 47 | 6.70 |
| Dmitry Peskov | 30 | 5.37 |
| Federation Council | 28 | 5.22 |
| RIA News | 24 | 4.86 |
| Fair Russia | 18 | 3.88 |
| State Duma Committee | 17 | 3.76 |
| Franz Klintsevich | 11 | 3.30 |
| Lower Chamber of the Parliament | 11 | 3.29 |
| Representative of the Kremlin | 11 | 3.17 |
| President of the United States | 13 | 3.03 |
| Normandy Format | 9 | 2.99 |
| Sergey Naryshkin | 9 | 2.95 |
| Minister of Defense | 9 | 2.93 |
| Law enforcement bodies | 8 | 2.81 |
| Alexey Pushkov | 8 | 2.79 |
| Council of Europe | 8 | 2.69 |

**Table 5. The first 16 significant collocations in the descending order of the value of the Dice measure, extracted from news flows for the period of June 2-16, 2016**

| *Information object* | *Frequency* | *Dice* |
|---|---|---|
| Normandy Format | 9 | 1.00 |
| François Hollande | 3 | 1.00 |
| Harry Truman | 3 | 1.00 |
| Angela Merkel | 3 | 1.00 |
| Recep Tayyip Erdoðan | 3 | 1.00 |
| Kerch Strait | 3 | 1.00 |
| Anastasia Volochkova | 4 | 0.88 |
| Ottoman Empire | 7 | 0.87 |
| RIA News | 24 | 0.87 |
| Black Sea Fleet | 3 | 0.85 |
| Innovation Fund Skolkovo | 3 | 0.85 |
| Ella Pamfilova | 5 | 0.83 |
| Petro Poroshenko | 5 | 0.76 |
| British newspaper the Times | 3 | 0.75 |
| Barack Obama | 6 | 0.66 |
| Donald Trump | 5 | 0.66 |

**Table 6. The first 16 significant collocations in the descending order of the value of the MI measure, extracted from news flows for the period of June 2-16,2016**

| *Information object* | *Frequency* | *MI* |
|---|---|---|
| François Hollande | 3 | 10.74 |
| Harry Truman | 3 | 10.74 |
| Angela Merkel | 3 | 10.74 |
| Recep Tayyip Erdoðan | 3 | 10.74 |
| Kerch Strait | 3 | 10.74 |
| Black Sea Fleet | 3 | 10.33 |
| Innovation Fund Skolkovo | 3 | 10.33 |
| Anastasia Volochkova | 4 | 10.01 |
| British newspaper The Times | 3 | 10.01 |
| Livonian Order | 3 | 9.74 |
| Council of the Parliamentary Assembly | 3 | 9.74 |
| Catherine the Great | 3 | 9.59 |
| Pranker Vovan | 3 | 9.59 |
| Ella Pamfilova | 5 | 9.52 |
| Akhmad Kadyrov | 3 | 9.33 |
| Petro Poroshenko | 5 | 9.26 |

## 4. DISCUSSION

The main objective of the paper was to propose methods and algorithms forextractinginformation objects of the types "person", "organization" and "geographical Location", containing more than one word, from unstructured flows oftext data. The proposed method is based on the identifying statistically significant word combinations through the use of various statistical measures. The paper doesn'tdescribe in detail the method for determining the belonging of the identified collocations to a certaintype of information objects, since there are ready mechanisms and solutions allowing to achieve this,which were described in this paper. The qualitative assessment of the efficiency of the method was demonstrated through the example of extracting information objects from news messages for the periods from October to November of 2015and June 2-16, 2016, and relating these objects to the events that occurred during these periods.

For the quantitative evaluation of the method, the authors used manual verification and the search for multiwordinformation objects that were not included in the list of identified objects. As a result of the verification on this sample, no new multiwordinformation objects, occurring in the text corpus more than two times, were detected. However, the method can identify one and the same object several times. For example, for the information object "Free Syrian Army", two fixed collocations were identified:"Syrian army" and "Free Syrian Army", which can be wrongin the general case. The paper doesn't deal with the methods of mergingintersecting collocations and includingcollocations in each other.

## 5. CONCLUSION

The proposed method makes it possible to extract information objects of the types "Person", "Organization" and "Geographical Location", consisting of more than one word, from the flows of unstructured textual information. The testing of the method provided acceptable results on news flows for a specific period of time.

To optimize the performance of the proposed method, it is possible to add algorithms for determining the syntactic coherence of potential collocations, i.e. the words within the considered collocation should be coherent in terms of gender, number and case. This improvement will make it possible to reduce the total amount of considered potential collocations, whichshould have a positive impact on the speed of the work of the method, but this requires further research, as the overhead costs related to the determination of the word coherence may exceed the benefits due to the reduction in the amount of collocations. To determine the word coherence, it is necessary to use the syntactic analysis of thesentences containing these words, which can be implemented using GLR grammars [15],link grammars [16-17], building decision trees and other methods. At this stage, the most appropriate in terms of performance may be link grammars, which will allow, without full parsing of the sentence, to determine the coherence oftwo words according to the rules of the language.

The proposed method doesn't deal withthe issue of including some of the identified collocations in other collocations. The algorithm of such inclusion is not very difficult to describe and implement, but this should be done in the future.

In the proposed method,onlythree statistical measures are considered: Dice, T-score and MI. In the future, other measures should also be considered and evaluated: Log-likelihood [18-19], MI3-Score, Log-log, logDice and others [20].

One of the drawbacks of this method is ignoring the synonyms and abbreviations in the process of identifying information objects. For example, the collocations "MFA of the Russian Federation", "Ministry of Foreign Affairs of Russia" and "Russian Foreign Ministry" will be defined as three distinct information objects, whichis actually incorrect. This problem can be partially solved by buildingsemantic relationships between these objects or by using the method of ontologies.

The proposedmethod is not a universal mechanism of detecting information objects; it is merely a method of identifying potential information objects, consisting of more than one word. The proposed mechanism could be used as a part of a decision support system or as a part of a system of semi-automatic formation of ontological databases [21] in conjunction with syntactic and semantic analysis of texts in natural languages.

## 6. ACKNOWLEDGEMENTS

The work was financially supported by the Ministry of Education and Science of the Russian Federation (Grant Agreement 14.579.21.0088, the unique identifier for applied scientific research RFMEFI57914X0088).

## 7. REFERENCES

1. Kuznetsov, I. P. (2012).The Methods of Discovery of Objects and Their Links Presented Implicitly in Texts.In *Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "Dialogue 2012"*.
2. Khokhlova, M.V. (2010). *Issledovanie leksiko-sintaksicheskoy sochetaemosti v russkom yazyke s pomoshch'yu statisticheskikh metodov (Dissertatsiya na soiskanie uchenoy stepeni kandidata filologicheskikh nauk)* [A Study of Lexical-Syntactic Compatibility in the Russian Language Using Statistical Methods (Thesis for the Degree of Candidate of Philological Sciences)]. Saint Petersburg: St. Petersburg State University.
3. Vlavatskaya, M.V. (2015). Kombinatornaya leksikologiya: funktsional'no-semanticheskaya klassifikatsiya kollokatsiy [Combinatorial Lexicology: Functional-Semantic Classification of Collocations]. *Filologicheskie nauki.Voprosy teorii i praktiki, 11(53), Part 1*.
4. Kilgarriff, A. (2006). Collocationality (And How to Measure It).In*Proceedings of the Euralex International Congress*. Torino.
5. *Morphological Analyzer Pymorphy2* (2013, February 14). Retrieved September 14, 2015, from https://pymorphy2.readthedocs.io/en/latest/index.html.
6. Zakharov, V.P., & Khokhlova, M.V. (2010). Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts.In *Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "Dialogue 2010"*.
7. Magerman, D.M., & Marcus, M.P. (1990). Parsing a Natural Language Using Mutual Information Statistics. In *Proceedings of the Eighth National conference on Artificial Intelligence*.
8. Manning, D.C., & Schutze, H. (1999). Collocations. In D.C. Manning & H. Schutze, *Foundations of Statistical Natural Language Processing* (pp. 151-191). Massachusetts: Massachusetts Institute of Technology.
9. Dice, L.R. (1945). *Measures of the Amount of Ecological Association between Species*. Retrieved July 6, 2016, from http://www.jstor.org/stable/1932409.
10. Khokhlova, M.V. (2014). Avtomaticheskoe vydelenie terminov i terminologicheskikh slovosochetaniy iz spetsial'nykh tekstov [Automatic Extraction of Terms and Terminological Phrases from Specialized Texts]. In *Materialy XLIII Mezhdunarodnoy filologicheskoy nauchnoy konferentsii, Rossiya, Sankt-Peterburg, 11-16 marta 2014* [Proceedings of the 43th International Philological Conference, Russia, St. Petersburg, March 11-16,2014].
11. Porter, M.F. (1997). *An Algorithm for Suffix Stripping*.San Francisco: Morgan Kaufmann Publishers Inc.
12. Korobov, M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages. In *Proceedings of the Fourth International Conference on Analysis of Images, Social Networks and Texts, AIST 2015* (pp. 320-332).
13. Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics:Investigating Language Structure and Use*. New York.
14. Bird, S., Klein, E., & Loper, E. (2009).*Natural Language Processing with Python*. Gravenstein Highway North, Sebastopol:O'Reilly Media, Inc.
15. Meng, H., Luk, P., Xu, K., & Weng, F. (2002). GLR Parsing with Multiple Grammars for Natural Language Queries. *ACM Transactions on Asian Language Information Processing (TALIP), 1(2),* 123-144.
16. *Link Grammar*. (2004). Retrieved June 2, 2016, from http://www.link.cs.cmu.edu/link.
17. Protasov, S.V. (2006). Obuchenie s nulya grammatike svyazey russkogo yazyka [Teaching Russian Link Grammar from the Start].In *Materialy desyatoy natsional'noy konferentsii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2006, Obninsk, Rossiya, 25-28 sentyabrya 2006 goda* [Proceedings of the 10th International Conference on Artificial Intelligence with International Participation KII-2006, Obninsk, Russia, September 25-28, 2006].
18. Dunning, T.E. (1993). Accurate Methods for the Statistics of Surprise and Coincidence.*Computational Linguistics, 19(1),* 61-74.
19. Dunning, T.E. (1998). *Finding Structure in Text, Genome and Other Symbolic Sequences* (Ph.D. Thesis). Sheffield: University of Sheffield, Department of Computer Science.
20. *Kollokatsii* [Collocations]. (2011,September 26). Retrieved June 14, 2016, from http://www.opencorpora.org/wiki/%D0%9A%D0%BE%D0%BB%D0%BB%D0%BE%D0%BA%D0%B0%D1%86%D0%B8%D0%B8.
21. Vakulenko, A.A., &Sytnik, D.A. (2015). Interactive Visualization of Multi-Dimensional Data of Heterogeneous Information Objects in the Decision Support Systems for Solving Multicriteria Choice Problem.*Journal of Theoretical and Applied Information Technology, 81(3),* 453-465.