Neural Network Algorithm for Detection of New Word Meanings Denoting Named Entities

Vladimir V. Bochkarev¹, Stanislav V. Khristroforov¹, Anna V. Shevlyakova¹, and Valery D. Solovyev¹
¹Kazan Federal University, Kazan, 420008, Russia

Corresponding author: Vladimir V. Bochkarev (Vladimir.Bochkarev@kpfu.ru).

This research was financially supported by Russian Science Foundation, grant № 20-18-00206.

ABSTRACT Lexical semantic change detection has been a rapidly developing field of science in recent years. Existed algorithms of lexical semantic change detection face difficulties when they are used to work with words denoting named entities. This paper proposes a method that allows one to reveal a word in a large corpus that started being used as a named entity, as well as to date the first usage of this word as a proper name. To solve this problem, firstly, we offer an algorithm that allows for detecting words in a large corpus denoting named entities. The recognizer is based on an analysis of co-occurrences with the most frequent words and was trained on data from the English subcorpus of the Google Books Ngram corpus. The achieved recognition accuracy of named entities is 98.44% on the test sample. Secondly, we test the possibility of applying the trained recognizer to diachronic data. The analysed cases show that the recognizer initially trained using the total bigram frequencies for a long time interval, at least for any frequent word, provides stable results for the annual frequency values. This can make the recognizer a good tool for language evolution studies, especially for detecting new meanings of words. The analysed cases show that the proposed method allows revealing new word meanings associated with named entities, as well as detecting genericized meaning of words that were earlier used as proper names.

INDEX TERMS computational linguistics, distributional semantics, Google Books Ngram, multi-layer neural network, lexical semantic change, named entities, natural language processing

I. INTRODUCTION

Diachronic semantics has been a traditional focus of interest in linguistic theory [1,2]. Now we observe a breakthrough in semantic studies since creation of large text corpora (such as Google Books Ngram [3,4], COHA [5]) and innovative methods of machine text processing has triggered a reversion of the long-standing issue of semantics.

Distributional approaches to semantic change started being used around the 2010s [6,7,8]. They are based on the hypothesis that there is a correlation between a change in meaning and a change in the context of use [9,10,11]. It means that a word distribution can be used to estimate the word meaning.

Various vector representations of words are used in works on semantic change detection. Good reviews of the applied approaches can be found in [12,13]. Currently, the most widely used methods apply vector models based on neural networks. However, simpler representations based on explicit word vectors are also employed to solve various problems concerning natural language processing.

Words associated with named entities is a real challenge for existing methods of lexical semantic change detection. Many works on this problem give a recommendation to pre-filter and not consider words associated with named entities (see, for example, [14,15]). In our paper, on the contrary, we will consider a specialized method for detecting new meanings of words associated with named entities.

Most of the words associated with names originally had a nominal basic meaning and later began to be used as names as well. The algorithm proposed in this paper allows one to detect and date the appearance of a new word meaning denoting named entity. To solve this problem, we use statistical data on word co-occurrence extracted from a large diachronic corpus.

It should be noted that we didn’t aim at clusterization of new word meanings. The main task was to detect the first appearance of a word as a named entity in the corpus.
However, variety of the detected word meanings can be further studied using the methods developed in [16,17,18].

Our study is based on the Google Books Ngram (GBN) corpus data. It appeared in 2009 and includes data on frequency of words and phrases in 8 languages over the past five centuries [3,4]. The English (common) subcorpus of GBN (Version 3) contains data on frequencies of individual words and n-grams from texts of 16.6 million books (published in 1470-2019) of the total size of 2 trillion words. The size of this corpus significantly exceeds the size of any other text corpus, which makes GBN a unique tool for language evolution studies. In addition to the Common English Corpus, GBN includes subcorpora of British and American English, as well as English Fiction corpus.

The GBN corpus does not provide free access to the source texts, only to frequencies of words and word sequences (2-, 3-, 4- and 5-grams) in a given year. This makes it impossible to use named entity recognition algorithms based on the word context of the named entities. A neural network recognizer that makes a decision based on the analysis of a word collocability is considered in [19]. This work uses Explicit Word Vectors, composed of relative frequencies of a word used in various contexts (as part of various bigrams). In [19], the described recognizer is tested on data from the Russian subcorpus of Google Books Ngram.

Firstly, the present study repeats the results presented in [19] for the English language. Secondly, we apply the trained recognizer to diachronic annual data and check the possibility of using it to detect new word meanings associated with named entities. When testing the recognizer on synchronous data, we use data on the frequencies of both ordinary and syntactic bigrams (as in [19]). However, when analyzing diachronic data, we only use data on the frequencies of ordinary bigrams. Thus, we do not use the syntactic markup available in the corpus in any way and only analyse changes in the distribution of words.

The results of the present work can be useful for improving conventional named entity recognizers. For example, the proposed recognizer can be applied to classify a significant part of the words contained in Google Books Ngram and create vast dictionaries of named entities. Besides, the recognizer can serve a good tool for language evolution studies because it allows detecting new meanings of words and date the first appearance of a word as a named entity in a text corpus.

The paper has the following structure. Section 2 provides a short review of works on lexical semantic change detection and the problem of named entity recognition. Section 3 describes the dataset used and the way of constructing a neural network recognizer of named entities. Section 4 describes the results of testing the trained recognizer on synchronous data from the English corpus of Google Books Ngram. Section 5 describes the application of the recognizer to diachronic data, as well as detailed analyses of the words that acquired new meanings associated with named entities. Section 6 discusses accuracy of the proposed method.

II. RELATED WORK

In this section, we provide a short overview of works on lexical semantic changes detection and named entity recognition.

A. LEXICAL SEMANTIC CHANGES DETECTION

Various papers discuss the problem of lexical semantic change in terms of distributional semantics that considers the interaction between use and meaning.

There are different algorithms for distributional semantic analysis. Early works mainly used representations based on combinability vectors [20,21,22,7,23]. It was proposed in [22] to use vectors based on Point Mutual Information (PMI). Cavallin [24] represented semantics through a set of words ranked according to the strength of their association with the word under study; and Mitra [25] represented word meanings by thesaurus-based graphs using a graph clusterization technique. Various ways for reducing the dimensions of vector representations were also considered, for example, those involving the use of SVD [26,27].

An improved word embedding analysis technique appeared in 2013 [28] and provided a new impetus for research. Various applications of word embeddings to the study of semantic change have been proposed in [8,29,30]. Despite a relatively short history, there are good reviews devoted to this area [12,13]. The most modern model for the semantic representation of words, BERT, was used in [31].

Currently, the most widely used methods are based on vector models of neural networks. However, simpler representations based on explicit word vectors are also used to solve various problems, such as studying the evolution of a language, discovering new meanings of words, and other problems in computational linguistics.

Hamilton compared different methods and concluded that SVD and Skip-gram models provide better results than PPMI vectors [14]. On the other hand, it is noted in [32] that applying methods that use one or another variant of dimension reduction (SVD, Skip-gram model, etc.) can lead to artifacts in the process of analysing semantic changes.

In [33], the authors perform an analysis based on the UK Web Archive corpus of terabyte size. It is noted that the method based on the direct use of combinability vectors requires less computational resources than word embeddings and application of methods based on word embeddings to extra-large corpora can cause difficulties. Moreover, direct use of combinability vectors has some advantages – they are easily interpreted and can be simply matched for different time intervals. The disadvantage of the method of direct use of combinability vectors is the high dimension of vectors.

An objective comparison of various methods for detecting semantic changes is hampered by the lack of a gold standard for testing the methods [12]. A test corpus consisting of 100
English words is provided in [7]. The method of comparison with dictionary entries is applied in [34].

As far as we know, all published works have studied the change in the semantics of only a small number of words, i.e., this research is at the ‘Case Study’ stage.

The work [25] uses the GBN subcorpus and considers changes in word meanings between 1909–1953 and 2002–2005. The algorithm detected 48 cases of appearance of new meanings; 29 of the cases were proved by the expert’s assessment. The experts also considered 21 cases of “splitting and joining of senses” and confirmed 12 cases. The words continuum, diagonal and intonation are given as examples of the meaning split.

Kulkarni discusses three algorithms for detecting changes in semantics [35]. The first one (Frequency Method) assumes that change in semantics leads to an abrupt change in word frequency. The second one (Syntactic Method) identifies semantic change based on information concerning part-of-speech change of words. The third algorithm (Distributional Method), proposed in [35], detects changes in word meanings through changes in their distribution. Comparative testing of the algorithms was carried out using the Google Books Ngram Corpus, the Amazon Movie Reviews Dataset and Twitter data. In total, the study considers 32 examples of words for which the emergence of new meanings was found. Three words are of particular interest to us. These words are Apple, Bush and Windows. New meanings of these words associated with named entities were revealed using the Syntactic Method. In order to apply this, the authors used the syntactic n-gram frequency dataset presented in [36] based on the second version of the English subcorpus of Google Books Ngram. Kulkarni used the corpus that had already been marked up and the words denoting named entities had also been marked by the corresponding tag (proper noun). The authors of [35] note that the proposed Distributional Method did not manage to reveal a new meaning for the word apple. It was revealed only by the Syntactic Method, through the presence of the corresponding markup in the corpus. Thus, the methods proposed in [35] can reveal new meanings associated with named entities only through the use of a large corpus where named entities have been already tagged. This significantly limits the possibilities of the method proposed in [35]. For example, only the English subcorpus belonging to the 2nd version of the Google Books Ngram corpus has such mark-up. The other subcorpora of the GBN corpus do not have such mark-up including the 3rd version of the English subcorpus.

B. NAMED ENTITY RECOGNITION

Named entity recognition (NER) is a fundamental problem in automated text processing. It is necessary to be able to solve it in many applied problems such as automatic analysis of news [37], analysis of users’ feedback on goods and services [38], etc. Recently, remarkable progress has been achieved in this area. Without pretending to comprehensively cover this topic, we will highlight only aspects that are important for our work.

Traditionally, most of the works was done for the English language, however, there are interesting studies for other languages such as Chinese [39] and Russian [40], etc. Good reviews of this area are presented in [41,42]. Usually, named entities of the following types are extracted from texts: people (names, surnames), organizations and locations. Names of drugs, moments of time, goods, proteins, etc. are recognized in various applied works.

An important infrastructural element of this area of research is carefully manually marked corpora of texts, which can be used to check the quality of newly developed programs. The most famous of such corpora is CoNLL 2003 [43]. The commonly used F1-measure is used to assess the quality of the NER program. Presently, the best result of 93.5% was obtained on this corpus in [44] by using neural networks.

It should be noted that training neural networks requires a large amount of marked-up data. Obtaining such data is an expensive procedure. Therefore, in recent years, considerable attention has been paid to reduction of the size of the training sample. For example, in order to cut costs of training external resources (for example, dictionaries) [45] and active learning [46] have been used for assistance in targeted selection of the most useful training examples. Using these approaches also implies some difficulties. The creation of large dictionaries of named entities is very labour-consuming. Another approach is proposed in [19], where statistics on the use of uppercase and lowercase words in a large corpus are used to create a training sample.

Little attention was paid to the problem of homonymy in NER studies so far. In this context, let us mention the work [47], in which the problem of homonymy is considered for the case of female and male names. This paper shows that female names are more likely to be incorrectly recognized. For example, names like Charlotte and Sofia are often mis-tagged as city names, or not recognized as names at all, even though they appear in contexts that explicitly indicate that they are people’s names. It is difficult to disambiguate words denoting locations and organizations. For example, Liverpool is both a city and a football club. A common situation is when words that were not originally named entities begin to be used as the names of newly created organizations or significant objects, an example being the word apple used as the name of a company.

To correctly process such words and identify cases of their use as names, it is necessary to resolve homonymy. In this article, we propose a technique that allows one to find the moment when a word acquired a new meaning (turned into a named entity).

III. DATA AND METHOD

Making a marked-up training set is the most time-consuming part in the process of creation of the named entity recognition
system. In [19], statistics on the use of words starting with lowercase and uppercase letters in the Russian subcorpus of Google Books Ngram was used to create a large training sample. It is a well-known fact that named entities start with uppercase letters and words that do not denote named entities are written with lowercase letters in many languages. There can be exceptions from this rule; however, this observation can be helpful when working with a large text corpus.

The training set included only words that satisfy the following requirements:

1. The word always starts with an uppercase letter, or the word always starts with a lowercase letter;
2. The word consists of the letters of the English alphabet (possibly with one apostrophe);
3. The word is marked up as a noun not less than in 90% cases.

The 50,000 most frequent words that satisfy these requirements were selected. They included 27,200 words that always start with an uppercase letter (54.4%) and 22,800 words that always start with a lowercase letter. 80% of the obtained words were used to train the recognizer and 20% were used to test it. The 20% test set was randomly selected from the whole sample.

In order to recognize named entities, the method of co-occurrence with the most frequent words (CFW) is used. The method is described in detail in [23,48,19]. In accordance with this approach, a word is represented by a vector of frequencies of 2-grams that include this word. Let the N most frequent words in the corpus (henceforth reference words) be chosen. To characterise the target word \( W \), \( N \) frequencies of 2-grams of the type \( Wx \) (where ‘\( x \)’ is a reference word) and the same number of 2-grams of the \( xW \) type are extracted from the corpus. Some of these word combinations can neither be found in the corpus nor in the language, so their frequency will be zero. It is not an easy task to select an optimal number of reference words. Thus, a list of \( N=5000 \) reference words, the most frequent in 1890 and 1999, was chosen in [23] to construct a vector representation. The present work, following [19], uses the 20,000 most frequent words as reference words. We conducted a series of preliminary experiments training the neural network with different \( N \). Using \( N=20,000 \) allows one to significantly improve the accuracy compared to the case when 5,000 and 10,000 reference words were used. A further increase in \( N \) up to 50,000 does not provide a further increase in recognition accuracy. At the same time, if a dimension of the vector representation is too large, it causes difficulties in using a large data set. Another consideration in favor of using 20,000 reference words is comparison of our results with those obtained in [19]. A word is thus described by a vector of 40,000 dimension. If the obtained vectors are normalized to 1, they can be interpreted as distributions of the probability of using a word in various contexts.

Besides frequencies of ordinary 2-grams (pairs of words that are direct neighbours in the text), the Google Books Ngram corpus contains data on frequencies of syntactic bigrams [4]. Syntactic bigram are units of syntactic structures denoting a binary relation between a pair of words in the sentence. In each syntactic bigram, one word is called the head, and the other is its dependent [49]. Recently, approaches based on the extraction of syntactic bigrams and analysis of their frequency have found application in various studies devoted to natural language processing [49]. Representation of words by vectors of ordinary and syntactic bigram frequencies (analogous to the one described above) was used in this paper. As it was noted above, the corpus contains information on frequencies not only of ordinary and syntactic bigrams, but also of 3-, 4- and 5-grams. This data can potentially be used to reveal named entities. For example, one can build a vector by calculating co-frequencies of a target and reference words in a n-gram. However, this is beyond the scope of our work.

The architecture of the classical feedforward network was chosen as in [19] (see the scheme in Figure 1). The network was a four-layer perceptron with 40,000 inputs, 128, 128, and 64 neurons in the three hidden layers, respectively.

![Scheme of the neural network](image)

**FIGURE 1.** Scheme of the neural network

As in [19], the rectifier activation function (RELU) was used as the activation functions of all hidden layers, and the neuron biases were equal to zero. In this case, the outputs of the last hidden layer are a homogeneous function of the inputs:

\[
y = Ax \rightarrow Ay = f(Ax)
\]

(1)

Here, \( x \) and \( y \) are the input and output vectors, respectively; and \( A \) is some positive number. Thus, the ratio of the two outputs does not depend on the module of the
input vector. It means that the result will not depend on the absolute word frequency and on the corpus size.

The dimension of the last four layers is equal to 2, according to the number of classes in the training sample (the first class includes words that are most often used as named entities, the second class contains words that are mainly used as common nouns). Activation of the output layer is performed by the softmax function [50]. This ensures non-negativity of the neural network outputs, as well as normalization of their sum to one that makes it possible to interpret the output data as distribution probabilities referring to the target classes.

Since the dimension of the input vector is high, the number of weights between the input and the first hidden layer is also high, which can lead to overfitting of the model. To prevent the overfitting effect, a dropout layer [51] with parameter 0.4 was created in front of the first hidden layer.

The model was trained using the backpropagation method based on the Nadam algorithm, which is a type of stochastic gradient descent [52,53].

To ensure high performance in the process of the model training, the entire training set is divided into a set of batches of a fixed size. Moreover, the network weights are updated, and the error gradient vectors are aggregated after all the examples from the batch are provided. The batch size was chosen to be 256. Thus, approximately 150 updates of the network weights occur during one epoch. The neural network was trained according to the minimum binary cross-entropy criterion [50]. The library of neural network calculations PyTorch was used.

IV. NAMED ENTITY RECOGNIZER

The trained recognizer was tested on a sample of 10,000 words that were not presented to the network at the training stage. The decision is made depending on the value of the ratio of the two outputs of the neural network. Depending on the chosen threshold, one obtains certain values of the error probabilities of the first and second type. Figure 2A shows receiver operating characteristics (ROC) of the obtained recognizer. The solid line shows the recognition results obtained using 2-gram frequencies; the dotted line shows the recognition results obtained using syntactic bigrams.

Since the error probabilities are quite small, Figure 2B shows the dependence of the type 2 error probability (β) on the error probability of type 1 (α) on a log-log scale. We choose as the threshold the value at which the probabilities of errors of the 1st and 2nd type are equal, as shown in the figure. In this case, the error probability is $\alpha = \beta = 1.563\%$ (the corresponding value of F1-score is 0.987) for the recognizer using frequencies of syntactic 2-grams, and 1.782% (the corresponding value of F1-score is 0.983) for the recognizer using frequencies of ordinary 2-grams. Error probability values of 2.71% and 3.27% were obtained in [19] under the same conditions for Russian data. The high recognition accuracy for English can be explained by the fact that the English corpus is significantly larger than the Russian one (more than 22 times as large).

Thus, we managed to obtain high recognition accuracy. However, these results cannot be directly compared to the recognition accuracy obtained in the well-known works on the recognition of named entities [41,42,44], since a substantially different problem is solved in our work.

- As a rule, the initial data used in the works on NER is text. The initial data for the recognizer used in our work is statistics on the distribution (co-occurrence) of the word extracted from the large corpus.
- Output of a traditional recognizer may be the decision that some words or phrases in a text belong to named entities; and required disambiguation can be performed by analyzing context of use of a word (phrase) in the text. Output of our algorithm is classification of words that occur in a large corpus.

The created recognizer can be applied not only to the vectors of the summed bigram frequencies but also to the frequency vectors for a given year; further, we will use the logarithm of the ratio of the recognizer outputs $L(t)$:

$$L = \log \frac{y_1}{y_2}$$

Here $y_1$ and $y_2$ are outputs of the recognizer (see the scheme in Figure 1). This characteristic can be interpreted as the logarithm of the likelihood ratio when testing the null hypothesis that the word denotes a named entity. Values above zero indicate that a word distribution is typical for a named entity, values below zero indicate that the word distribution is not characteristic of a named entity.

In this paper, we are primarily interested in changes in $L(t)$ over time. Significant jumps in $L(t)$, especially associated with a change in the sign of this value, may indicate that a word that was previously used mainly as a common noun has begun to be used as a proper name or vice versa. Examples below will show that sometimes several jumps can be observed in the $L(t)$ graph. As a rule, their presence indicates emergence of a new meaning of a word or increase in the relevance of one of the previously existed meanings.

![FIGURE 2. A - Receiver operating characteristics of the named entity recognizer for the English language. B - the receiver operating characteristic on a log-log scale for the same case. The solid line relates to the use of syntactic bigrams, the dotted line relates to the case of ordinary bigrams.](image-url)
The question arises whether it is possible to perform a fully automatic (without participation of an expert) analysis of the \( L(t) \) curve to highlight jumps corresponding to appearance of a new value. There are many works devoted to change point detection in time series [54, 55]. To solve the problem of change point detection, various methods were proposed that use nonparametric methods, state space models, Kalman filters, etc. (see, for example, [56, 57]). Special attention deserves the algorithm described in [58] and used in [35] to determine appearance of a new word meaning. There are also some recent relevant works [59, 60]. The question of which of the numerous algorithms developed to date is best suited for the analysis of the \( L(t) \) series will apparently require additional research. Note, however, that in the examples under consideration, the length of the analyzed time series is rather small (from several tens to 200–250 samples), which makes it preferable to use the simplest algorithms.

In this paper, we use the method proposed in [60]. We apply this algorithm to the \( L(t) \) curve for the selected word (within the considered interval 1800–2019) choosing 3 candidate points in each case. It is possible to test significance of changes in the point selected by the algorithm using the method proposed in [61]. The testing is performed in the following sequence:

- Selecting two time intervals before and after the proposed change point. In this work, we (unless otherwise stated) took two 5-year intervals before and after the change point.
- Generating random vectors of bigram frequencies for each of the selected intervals using the algorithm described in [61]. In this work, we generated 100 vectors for each interval.
- Submitting the generated vectors to the input of the neural network. Having done this, we obtained two samples of estimates \( L(t) \) (for the first and second time intervals);
- Testing significance of differences between the two samples, for example using the Wilcoxon rank sum test.

Let us analyse some examples.

**V. CASE STUDY**

**A. APPLE**

The word *apple* is one of the most vivid examples of how a common noun becomes a named entity. The central meaning of the word *apple* is “the fleshy, usually rounded red, yellow, or green edible pome fruit of a usually cultivated tree of the rose family” [62]. Figure 3A shows the percentage of the uppercase use of the word *apple* of the total usage of this word. The figure is built on the data of the English (common) subcorpus of Google Books Ngram. As one can see, this percentage does not rise above 20–25% until the mid-70s of the 20th century. The increase in the percentage has been observed since the second half of the 1970s; and a sharp peak was observed in 1979–1983.

Figure 3B shows the change in the value of \( L(t) \) for the word *apple* over time (here and hereafter, data on the frequencies of ordinary bigrams are used for plotting). Until the middle of 1970s of the 20th century, the curve goes significantly below zero. However, in 1978–1985, the \( L(t) \) graph shows a sharp jump; and in 1984–1985, the curve crosses the zero mark. This behavior of the curve indicates that the word acquires the meaning of a named entity at that time. Note that the algorithm [60] automatically selects the year 1981 as a change point. At that, the p-value is \( 3 \times 10^{-11} \). In fact, the brand ‘Apple’ has come to be widely used since that time. The company name Apple Computer, Inc. was officially registered on April 1, 1976 [63]. In 1977, the company introduced the Apple II computer, which became one of the first personal computers and was released in a large series. In 1984, Apple launched an innovative 32-bit computer Macintosh that brought the company a great success [63].

**FIGURE 3.** A - the percentage of the uppercase use of the word *Apple* of the total usage of this word; B - change in the \( L(t) \) value for the word *Apple* over time.

It is of great interest to identify the word combinations that cause such a great change in the \( L(t) \) value. The approach applied is similar to the method of detecting new word meanings described in [64]. First, we choose two time intervals, that are the interval of the peak \( L(t) \) values (1982–1994) and the reference interval before the jump (1932–1974). We then build vectors of bigram frequencies \( (v^o, v^c) \) for these intervals representing the word studied. The task is to reveal individual contribution of each of the vector components to the \( L(t) \) increase. To do this, we build vectors \( v^{(k)} \) where the \( k \)-th component is taken from \( v^o \) and the rest are taken form the vector \( v^c \):

\[
v^{(k)}_i = (1 - \delta_k) v^c_i + \delta_k v^o_i \quad (3)
\]

Then we form the partial increments of \( L(t) \):

\[
\delta L_k = L(v^{(k)}) - L(v^c) \quad (4)
\]

We then select the components that meet the following requirements:

1) the frequency of the corresponding bigram increases,
2) $\delta L$ has the same sign as the $L\left( v^P \right) - L\left( v^r \right)$.

Finally, we sort the selected components in descending order, according to the $\delta L$.

Bigrams that contribute most to the $L(t)$ jump will be at the top of the resulting list. The proposed approach allows to obtain a ‘combinability portrait’ of the word used in its new meaning.

The words that combine with the word Apple and contribute most to the $L(t)$ value increase in 1978–1985 are IBM, services, system, computer, is, family, computers, at, from, logo, software, 's. The last one deserves some comment. In the process of tokenizing texts during the creation of the Google Books Ngram corpus, the apostrophe was considered a separator in most cases. Thus, the possessive form of Apple’s in the corpus is considered a bigram. The possessive form (in most cases) is characteristic of a word denoting a named entity. The bigrams also contain two Roman numerals, II and III, as well as the word form plus, all of which is a designation of different computer models. It would be natural to expect the word Macintosh to appear in this list as well. However, this word was not included in the list of 20,000 reference words due to its lower frequency. As a result, it is not detected using the above-described method.

It should also be noted that Figure 3B shows a smooth curve with a relatively small number of fluctuations of the $L(t)$ values from year to year. For example, throughout the time interval 1925–1974, the standard deviation of $L(t)$ is $1.54 \cdot 10^4$, with an average value of $-5.62 \cdot 10^4$. This suggests that there is enough statistics on the use of this word during a given year for a reliable estimate of $L(t)$.

B. VOYAGER

The word voyager is one more example of how a common noun becomes a named entity. Originally this word means someone who travels on a long journey, especially by boat [65] or a person who goes on a long and sometimes dangerous journey [66].

However, the word obtained one more meaning and started denoting a named entity. Figure 4 shows the percentage of the uppercase use of the word voyager of the total use of this word in the English (common) subcorpus of Google Books Ngram (Figure 4A) and changes in the $L(t)$ value for the word voyager with time (Figure 4B). The graph shows outlier $L(t)$ values occurring in 1977–1981. These are due to the Voyager mission, which was certainly an important milestone in space exploration and was widely discussed in the literature. The twin spacecrafts Voyager 1 and Voyager 2 were launched in 1977 and flew past Jupiter and Saturn in 1979–1981 [67]. The algorithm [60] selected the year 1978 as a change point; and the p-value equaled $2.9 \cdot 10^{-11}$.

The years 1980–1995 were chosen as the interval of the peak values of $L(t)$; and the interval 1900–1940 was chosen as the reference interval. Words that combine with Voyager and contribute most to the $L(t)$ increase are spacecraft, 's, I, II, mission, missions, observations, project, images, data, encounter. All of them (in one way or another) relate to the Voyager spacecraft project.

C. TWAIN

The word twain is an interesting example of a word changing from being a numeral to a word denoting a named entity. Twain is an archaic word for ‘two’. This word was also used in river navigation – there was a term Mark Twain, which meant the minimum depth suitable for passing river vessels (2 fathoms).

A sharp increase of the $L(t)$ value occurs in 1869–1873 (see Figure 5), a period corresponding to the beginning of the active literary activity of Samuel Clemens, a famous American writer who wrote under the pseudonym Mark Twain [68]. The year 1872 is selected by the algorithm as a change point; and the p-value equals $3.5 \cdot 10^{-10}$. The words Mark, 's, says, by, said, book, work, works, tell, himself, library, has, had once, tells, was, went, called, describes combine with the name Twain and contribute most to the sharp increase of the $L(t)$ value (The years 1875–1910 were selected as an interval of peak values and the interval 1800–1865 was selected as a reference interval). These words are directly connected with Twain’s person and his works. Until the present time, the word Twain is preferably used as a named entity, as shown by the graph.

---

FIGURE 5. A - the percentage of the uppercase use of the word Twain of the total usage of this word; B - change in the $L(t)$ value for the word Twain over time.

A sharp increase of the $L(t)$ value occurs in 1869–1873 (see Figure 5), a period corresponding to the beginning of the active literary activity of Samuel Clemens, a famous American writer who wrote under the pseudonym Mark Twain [68]. The year 1872 is selected by the algorithm as a change point; and the p-value equals $3.5 \cdot 10^{-10}$. The words Mark, 's, says, by, said, book, work, works, tell, himself, library, has, had once, tells, was, went, called, describes combine with the name Twain and contribute most to the sharp increase of the $L(t)$ value (The years 1875–1910 were selected as an interval of peak values and the interval 1800–1865 was selected as a reference interval). These words are directly connected with Twain’s person and his works. Until the present time, the word Twain is preferably used as a named entity, as shown by the graph.
D. TITANIC

The following example is the adjective *titanic* which means ‘extremely large, powerful, or important’ [65]. It also means ‘made of titan’. However, since 1910, this word has been widely used in the printed press as a named entity denoting a famous ship. ‘Titanic’ was laid down in 1909, it was launched in May 1911, and crashed in 1912. At that time, it was the largest passenger liner in the world; B *Age of the uppercase use of the word Titanic in 1997*, the centenary of the ship Titanic and the tragedy caused by its crashing. According to the data obtained using the American English subcorpus data. Figure 6 shows a sharp increase of the $L(t)$ value in 1910–1912. It should be noted that the $L(t)$ value is positive only in 1912.

![Figure 6](image-url)  
**FIGURE 6.** A - the percentage of the uppercase use of the word Titanic of the total usage of this word; B - change in the $L(t)$ value for the word Titanic over time.

The words *the, of, disaster, to, is, from, steamship, SS* (abbreviation of the word combination steam ship), *liner, reached, was, had, went, were, steamer, memorial, sank, called, tragedy* combine with *Titanic* an contribute most to the sharp increase of $L(t)$ (the years 1912 and 1890–1909 were used for the analysis). Some of the words are function words. However, most of the words are associated with the ship Titanic and the tragedy caused by its crashing.

Note that, the algorithm [60] does not find the point of appearance of a new meaning. Apparently, this happens since in this case, in contrast to the previous examples, we observe a short-term burst of $L(t)$. Nevertheless, having manually selected two intervals 1890-1909 and 1912 (as above), we will check the null hypothesis about absence of changes in $L(t)$. In accordance with the method [61], we obtain a $p$-value equal to $6.3 \cdot 10^{-12}$. Thus, the change in $L(t)$ is statistically significant.

It should be noted that the $L(t)$ value increases in the end of the 20th and at the beginning of the 21st century. At that, there are three peaks in the graph that can be associated with such events as Robert Ballard’s expedition and discovery of the wreckage of the liner in 1985, the release of the feature film ‘Titanic’ by James Cameron in 1997, the centenary of the catastrophe in 2012, as well as the crash of the Costa Concordia cruise ship in January 2012. This catastrophe was compared with the Titanic ship crash.

E. BUSH

The previous examples show that the graphs of $L(t)$ and the percentage of the upper-case use of the word are largely similar in shape. Let us consider an example of a word for which these graphs are significantly different.

The word *Bush*, in addition to its basic meaning, denotes a common surname of English or German origin. Among other famous American families, the Bushes have been described as “the most successful political dynasty in American history” [70]. Members of the family have held various national and state offices for four generations.

![Figure 7](image-url)  
**FIGURE 7.** A - the percentage of the uppercase use of the word *Bush* of the total usage of this word; B - change in the $L(t)$ value for the word *Bush* over time. The data on bigram frequencies come from the American and British subcorpora.

Figure 7A shows the percentage of the uppercase use of the word *Bush* of the total usage of this word in the American and British English subcorpora of Google Books Ngram. Figure 7B shows the change in the $L(t)$ value of the word *Bush* according to the data of these subcorpora. Consider first the results obtained using the American English subcorpus data. Attention should be drawn to the large jumps on the graphs of both values under consideration, starting from the 1980s of the 20th century. Obviously, these jumps are associated with the activity of the two most famous representatives of the Bush family: George H. W. Bush, vice president in 1981–1989 and president of the United States in 1989–1993, and George W. Bush, president of the United States in 2001–2009.

It is seen that the percentage of the uppercase use of the word *Bush* increases in 1988–1989 and in 2000–2001 and decreases in each case after the end of the presidential term. The $L(t)$ graph differs significantly in shape. First, a sharp increase has been noted since 1980. Moreover, there is no decline in $L(t)$ up to the present time. Note that the year 1980 is automatically selected as a change point; at that, the $p$-value is $3.1 \cdot 10^{-11}$. The years 1980–2019 is chosen as the interval of the $L(t)$ peak values, and the interval 1966–1988 is chosen as the reference one. Bigrams that contain the word *Bush* contribute much to the observed sharp increase and include the following words: president, George, administration, ‘s, said, had, Barbara, campaign, took, announced. These bigrams are associated with G. Bush’s political activity and his life.

The increase of the $L(t)$ value is observed on the graph in 1952–1963. After selecting the target intervals 1952–1963
and 1940-1952, we find the bigram “senator Bush”, which makes the greatest contribution to this increase. This allows one to associate the observed sharp increase of $L(t)$ with the Prescott S. Bush’s activity. He was the Senator from the State of Connecticut in 1952-1963. The first 10 bigrams that contribute most to the described change include Prescott Bush and Bush said, the high frequency of which is due to Prescott S. Bush’s political activity. Unlike the previous examples, in this case, the contribution of a single bigram, calculated in accordance with expression (4), is 71% of the observed increase in $L(t)$. The contribution of the second most important bigram is 28 times less! Figure 8 shows the change in the frequency of the word bush (case-insensitive) and the bigram senator Bush in 1930–1980. It can be seen that the sharp increase in the use of the bigram occurs synchronously with a slight increase in the frequency of use of the word bush and has the similar shape. Even though the change in the percentage of the uppercase use of the word bush in these years does not exceed the level of random fluctuations (see Figure 7A), the neural network recognizer detects a change in the nature of the use of the word.

**FIGURE 8.** **Frequency of the word bush (case-insensitive) and the bigram senator Bush in 1930-1980.**

Let us consider the change in $L(t)$, calculated from the British subcorpus (see Figure 7B). In contrast to the graph based on the American subcorpus, one can see two jumps, in 1988 and in 1999-2001, and observe the decrease in $L(t)$ after the expiration of the presidential terms (George H. W. Bush and George W. Bush). There is no jump detected in 1953–1962 for the British subcorpus. As we can see, the British subcorpus reflects only those events in American history that are of global, worldwide significance.

**F. DREADNOUGHT**

The previous examples show that the word can start denoting a named entity in certain times. However, reverse processes can also be observed in the language when a word used as a named entity acquires a new widely used meaning and turns into a common noun. Let us consider the following example.

The adjective Dreadnought has traditionally been used as the name for the warships of the British Navy. As one can see in Figure 9, $L(t)$ fluctuates near zero up to 1904 (unlike previous examples, this figure is based on data from the British subcorpus of Google Books Ngram). In 1904, a new generation of battleship, which was later named ‘Dreadnought’ began to be built in Britain. The ship was laid down in 1905 and entered service in October 1906. The name of the innovative ship quickly became a common name and denoted an entire class of battleships [71]. In this sense, the word Dreadnought was borrowed by many languages. In Figure 9B one can see a significant decrease in the $L(t)$ values in 1906–1909, when mass construction of dreadnoughts began in Great Britain and other countries.

**FIGURE 9.** **A - the percentage of the uppercase use of the word Dreadnought of the total usage of this word; B - change in the $L(t)$ value for the word Dreadnought over time.**

To express this effect quantitatively, let us calculate the total frequencies of bigrams including the word Dreadnought in the period 1909–1920. The value of the parameter $L$ obtained based on these frequencies is equal to $-2.65 \times 10^4$. For comparison, the value of the parameter $L$ obtained using total frequencies for the interval 1800-1904 equals $-0.34 \times 10^4$. The same value calculated for the interval 1898–1904 turns out to be positive and equals $0.11 \times 10^4$. The algorithm automatically selects the change point in 1909 (the $p$-value equals $3 \times 10^{-11}$).

The words that combine with the word dreadnought and contribute most to the jump in 1906–1909 include the following ones: class, one, first, our, single, ships, build, strength, every, two, construction, design, vessels, their, Brazilian, large. The word Brazilian appears in the list because Brazil was one of the first countries to order battleships of the new class. (The intervals 1908–1920 and 1890–1903 were chosen for the analysis). Most of them are in one way or another associated with the design and construction of battleships of the dreadnought class.

**G. POTTER**

One more example of a word acting both as a common noun and as a named entity is the word potter. Originally, potter denotes someone who makes dishes or other objects out of clay. However, this word also denotes a common surname, that is, a named entity. Over the past two centuries, the $L(t)$ values have been consistently above zero, which is associated with the prevalence of the use of this word as a named entity (see Figure 10). Note that, as in the previous case, this figure is based on data from the British subcorpus of Google Books Ngram.

As one can see from Figure 10B, the $L(t)$ values are quite stable in 1960-2000. Since 2001, $L(t)$ has jumped to significantly larger values, and then there are changes in a new, higher range. It was in 2001 when the first film about
Harry Potter (based on J. K. Rowling’s books) was released. The year 2001 was selected by the algorithm as a change point (the p-value equaled $3 \cdot 10^{-11}$).

The words that form bigrams with the word Potter and contribute much to this jump include *Harry*, *and*, *at*, *said*, *prince*, *in*, *lord*, *see*, *Elisabeth*, *book*, *philosopher*, *school*, *and*, *mark*, *justice*, *looked*, *also* (The 2001–2019 and 1960–2000 intervals were chosen for the analysis). Some of them are directly related to the tales about the wizard boy. Thus, this is a good example of how a word that has already been used to denote named entities acquires a new meaning associated with the name of a popular literary character.

It should be noted that although we trained the recognizer on words free from homonymy, the above examples show that it provides quite reasonable results for homonyms.

### Table 1

| word     | Corpus   | Total frequency in 1800-2019, millions | Time interval | Average annual frequency, thousands per year |
|----------|----------|---------------------------------------|---------------|---------------------------------------------|
| Apple    | English  | 31.91                                 | 1968-1995     | 275.8                                       |
| Voyager  | English  | 1.338                                 | 1967-1991     | 16.84                                       |
| Twain    | English  | 5.478                                 | 1859-1883     | 7.069                                       |
| Titanic  | English  | 2.025                                 | 1902-1922     | 13.33                                       |
| Bush     | American | 21.59                                 | 1966-2019     | 293.6                                       |
| Bush     | American | 21.59                                 | 1940-1963     | 64.13                                       |
| Bush     | British  | 5.189                                 | 1966-2019     | 5.189                                       |
| Dreadnought | British  | 0.213                                 | 1896-1919     | 3.251                                       |
| Potter   | British  | 1.934                                 | 1991-2011     | 19.12                                       |

### H. Influence of Word Frequency

The pictures in the previous subsections show that different level of values of $L(t)$ fluctuations is observed for different words. The most obvious factor that determines the standard deviation of the values is word frequency. Table 1 summarizes information about the frequency of each of the items discussed above.

The table provides information on each of the analysed words, i.e., the corpus used for the analysis (English (common), American or British), total frequency for the period 1800–2019, and the average annual frequency in the period ±10 years from the time specified in the corresponding section of the event in question. For convenience, the time interval for which the average frequency was determined is also indicated in the table. The most frequent of the examples is the word *apple*, the rarest is the word *dreadnought*.

In all the cases considered above, the $L(t)$ was estimated using yearly data on frequencies of bigrams that include the target word. However, statistics on rare word co-occurrence for one year may not be enough to reliably estimate $L(t)$. In this section, we will briefly consider how choosing a time span influence the analysis. The calculation scheme requires only minimal changes. We extract bigram frequencies for each of the years within the selected time window and find their total frequencies over this time span. Next, we use the resulting frequency vectors to estimate the $L(t)$ as described above. Figure 11 shows the $L(t)$ estimates for the word *apple* (according to the common English corpus) obtained using different window widths $T$ ($T=1, 2, 4, 8, 16$). The figure illustrates that increase of the time window length causes decrease of the standard deviation of $L(t)$ fluctuations; however, the jumps in $L(t)$ associated with changes in the word meaning also smooth out (see the discussion above).

Thus, in practice, the window width should be selected as a compromise between the requirements for the accuracy of the $L(t)$ estimate and the required time resolution.

### I. Other Cases

Besides the words discussed above, new meanings of the words *Avenger*, *Blackberry*, *Diesel*, *Liberator*, *Monitor*, *Orange*, *Windows* and *Zeppelin* were also revealed. The new meanings are the following:
Avenger. A jump of L(t) values is detected in the American subcorpus in 1942–1945. It is associated with the name of the Avenger torpedo bombers type, as well as the name of the American Air Force training base Avenger Field, which was in use during the Second World War.

Blackberry. A dramatic change in L(t) values is detected in all English subcorpora around 2005, associated with the widespread adoption of BlackBerry smartphones during these years.

Diesel. This word originally represented a surname of German origin. A new meaning (nominal) of this word is found in all English corpora, including the fiction corpus. For example, the transition from initially positive values of L(t) to negative ones is observed in the common English corpus in 1919-1927, which is associated with the increasingly widespread use of diesel engines.

Liberator. A jump in the L(t) values in the American and British subcorpora is recorded in 1822–1829. It is associated with the use of this word as the personal name of Simon Bolivar (The Liberator, President Liberator, etc.). The English subcorpus also shows a peak in 1893–1895, associated with numerous mentions of the name of the Liberator Building Society, the bankruptcy of which was widely discussed in the British press in those years.

Monitor. A peak in L(t) values is detected both in the American and British subcorpora in 1862. It is associated with the construction of USS Monitor, an ironclad warship built for the Union Navy during the American Civil War.

Orange. A sharp increase in L(t) values is recorded in 1897–1900 in the British English subcorpus. It reaches its peak in 1902 and gradually declines by the beginning of the First World War. The analysis shows that this surge is associated with the events of the Anglo-Boer War of 1899-1902, and the use of the name Orange Free State, and, after the annexation of this state by the British Empire, the names of Orange River Colony and Orange Free State Province. Also, at the beginning of the 19th century, a number of less significant surges were recorded (for example, in 1816 and 1835–1836), which are associated with the political struggle surrounding the Loyal Orange Institution.

Windows. This word has shown a sharp increase in the L(t) value since 1992, which is associated with release of a family of operating systems (proprietary graphical operating system families) by Microsoft Corporation.

Zeppelin. Originally this word represented a German aristocratic surname. After the success of the rigid airships created by Count Ferdinand von Zeppelin, the word zeppelin came to be commonly used to refer to all rigid airships. The emergence of a common noun is determined by a sharp drop in L(t) in 1912-1914 (in the Common English subcorpus). After 1972–1973, L(t) increases again and remains positive until the present time. This is primarily due to the emergence of another new meaning of this word, which is included in the name of the famous musical band Led Zeppelin.

VI. DISCUSSION
To determine the accuracy of lexical semantic change detection algorithms, a number of test datasets have been created to date. Unfortunately, now there are no specialized datasets that contain enough words whose new meanings are associated with named entities.

For example, in 2020, a dataset [72] was presented, which is used in many works to test proposed semantic change detection algorithms. However, it is completely unsuitable for evaluating our algorithm because none of the 37 lemmas presented in it are related to any widely known named entities.

For lack of a better test dataset, we use the word list described in [35]. The proposed list consists of 32 words; and the authors indicate that 7 words obtained new meanings associated with named entities. These words are apple, bush, candy, mystery, sandy, twilight and windows. The words apple, bush and windows were discussed in previous sections of the present article. Let us briefly discuss the other four words.

It is stated in [35] that the word twilight gained new meaning in 2009 which can be explained by releasing a series of books by an American novelist Stephenie Meyer (2005) and a movie Twilight (2008) based on the same novel. The L(t) graph shows a peak connected with the appearance of the discussed meaning in 2010-2017. The peak is highly pronounced for the American English corpus data though it is also typical for the English corpus (common) data. This is due to the fact that the books and the film were published and shot in the USA.

The word sandy can be an adjective denoting colour or something “covered with or containing sand” [66]. Besides, Sandy is a widely used proper name. It is indicated in [35] that a new meaning of sandy appeared in 2012 due to the Hurricane Sandy (female names have been more often used as hurricane names). Since the 1960s, there had been an increase in L(t), which had become highly evident since the 1980s. At the same time, there is also an increase in the percentage of the uppercase use of Sandy. Since 2000, the number of the uppercase occurrences of Sandy has exceeded the number of the lowercase uses. Therefore, the proportion of Sandy used in the corpus texts as a proper name has been growing that is also indicated by the change in the L(t) value. Notably, the use of this word as a proper name in the expression Hurricane Sandy takes only a small percentage of
use of the word *sandy* as a named entity. The maximum occurrence frequency of the expression *Hurricane Sandy* in the corpus is revealed in 2014; its value was 2.45% of the total frequency of the use of *sandy* (considering both uppercases and lowercases uses). Thus, the use of the word *Sandy* as the name of the hurricane takes only a small percentage of its use as a proper name and is not revealed by the change in $L(t)$.

The word *candy* denotes a type of sweet confectionery. However, it can be used as a proper noun. For example, the new meaning of the word *candy* associated with the name *Candy Crush Saga* (a free-to-play match-three puzzle video game released by King.com Limited) is described in [35]. Besides, *Candy* often refers to real people and fictional characters denoting their names, nicknames, stage names or surnames. The English word *Candy* or *Candia* also refer to Crete, a Greek island. The $L(t)$ graph shows a number of peaks that can be caused by publishing of literary works (for example, in 1958 and 1998), musical compositions and albums (for example, in 1985, 1998–1999, 2006, etc.)

Moreover, the most significant jump at the beginning of the 19th century is associated with the Kandyan Wars, which were waged by the British Empire in 1796–1818 to conquer and annex the Kingdom of Kandy (the alternative spelling of *Candy*). Thus, according to the $L(t)$ graph, the word *Candy* is used in different years to denote various named entities. However, the meaning indicated in [35] is not found among them by our method. Indeed, the frequency of the expression *Candy Crush* in the corpus reaches its maximum in 2016 and equals 0.98% of the total frequency of the word candy in this year.

Therefore, the frequency use of the name *Candy Crush* in the corpus is small compared to the frequencies of other named entities denoted by the word *Candy*.

*Mystery* is described in [35] as a word that obtained a new meaning in 2012 caused by the release of *Mystery Manor*, a hidden object game developed by Game Insight. The $L(t)$ graph does not reveal the use of the word mystery to denote a named entity. It should be noted that the frequency use of the expression *Mystery Manor* reaches peak values in 1937 and 2016 being 0.044% and 0.023%, respectively (of the total frequency of the word *mystery*). That is, the relative frequency of use of the word *mystery* in the meaning indicated in [35] is very small.

To sum up, new meanings of 4 words form the list of 7 words considered in [35] were found by our algorithm. For the two more words, we revealed a number of meanings associated with named entities and dated their appearance in the corpus. However, for these 2 words the values indicated in [35] were not found. The main reason for this is relative rarity of words used in the meanings specified in [35] compared to the frequency of their use in the meanings associated with other named entities. Finally, we didn’t reveal the word *mystery* used as a named entity. It can also be explained by the fact that this word in the meaning described in [35] has a very low relative frequency.

VII. CONCLUSION

Recent work [19] has proposed a new algorithm that allows one to identify words denoting named entities in large corpus data. The algorithm was tested based on the Russian subcorpus of Google Books Ngram. In the present work, we repeated the same results for the English language.

Analogously to [19] the recognizer was based on the analysis of the combinability with the most frequent words (CFW) and was trained on the data of the English subcorpus of Google Books Ngram.

The recognition error probability gotten from the test sample of 10,000 words free from homonymy and used at least 750 times over the entire period was 1.563% ($F_1 = 0.987$). This result was obtained using the frequencies of syntactic bigrams. Just as for Russian [19], the use of frequencies of ordinary bigrams provides a slightly lower accuracy; in this case, the obtained error probability was 1.782%, and, accordingly, the value of the $F_1$-measure was 0.983. The obtained values are relatively high. However, as it was detailly explained above, they cannot be directly compared to the results of the well-known works on the recognition of named entities.

The main result of this work is a test of the possibility of applying the trained recognizer to diachronic data. The above examples show that the recognizer initially trained on data averaged for a long-time interval provides stable results for the data for each year (at least, for any frequent words). This allows for using the recognizer for language evolution studies, and above all for revealing new meanings of words.

The proposed recognizer managed to reveal new meanings (which relate to named entities) of the words *apple*, *voyager*, *twain*, *titanic* and *bush*, as well as to adequately date the moment of their appearance. Moreover, the method also makes it possible to single out a set of the most frequent bigrams that characterize the use of a word in a new meaning and, thus, to obtain a kind of ‘combinability profile’ of a new word. The method also allowed to detect a new meaning of the word *Potter* associated with a famous literature character.

The word *Dreadnought* is a good example of a word that, besides denoting a named entity, acquires a new common noun meaning. This event is also reliably detected and dated using the recognizer considered in this work, even though the word *Dreadnought* is the least frequent of all the words considered here.

Note that in 6 cases from the 7 considered ones, we managed to determine appearance of a new meaning fully automatically.

The performed analysis shows that the proposed method allows revealing new word meanings associated with named entities, as well as detecting genericized meaning of words that were earlier used as proper names. The question of the sensitivity of the method remains open. For example, what percentage of use of a word in a new meaning is enough to be detected? To answer this question, one needs marked-up datasets, indicating the percentage of use of words in
different meanings. This could be an objective of further research.

Besides, the results of the present work can be useful for improving conventional named entity recognizers. Using a trained CFW recognizer, one can classify a significant part of the words contained in Google Books Ngram and create vast dictionaries of named entities.

REFERENCES

[1] E. C. Traugott, R. B. Dasher, Regularity in semantic change, UK: Cambridge University Press, 2001.
[2] A. Deo, 2015. “Diachronic semantics,” Annu. Rev. Linguist., vol. 1, pp. 179–197, Jan. 2015.
[3] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, et al. “Quantitative analysis of culture using millions of digitized books,” Science, vol. 331(6014), pp. 176-182, Jan. 2011.
[4] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, S. Petrov, “Syntactic Annotations for the Google Books Ngram Corpus,” in 50th Annual Meeting of the Association for Computational Linguistics 2012, Proceedings of the Conference, vol. 2, Jeju Island, Korea, 2012, pp. 238-242.
[5] M. Davies, “Expanding horizons in historical linguistics with the 400 million-word Corpus of Historical American English,” Corpora, vol. 7, no. 2, pp. 121–157, 2012.
[6] E. Sagí, S. Kaufmann, B. Clark, “Semantic Density Analysis: Comparing word meaning across time and phonetic space,” in Proc. of the EACL 2009 GEMS Workshop, Athens, Greece, 2009, pp. 104-111.
[7] K. Gulordava, M. Baroni, “A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus,” in Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP, Edinburgh, UK, 2011, pp. 67–71.
[8] Y. Kim, Y. I. Chiu, K. Hanaki, D. Hegde, S. Petrov, “Temporal analysis of language through neural language models,” in Proceedings of the 32nd annual meeting of the association for computational linguistics, Baltimore, USA, 2014, pp. 61–65.
[9] Z. Harris, Papers in structural and transformational Linguistics, Dordrecht,Germany; Reidel, 1970.
[10] H. Rubenstein, J. Goodenough, “Contextual correlates of synonymy,” Communications of the ACM, vol. 8, no. 10, pp. 627-633, 1965.
[11] J.R. Firth. “A Synopsis of Linguistic Theory, 1930–1955,” Studies in Linguistic Analysis. Special volume of the Philological Society, Oxford, UK: Basil Blackwell, 1962.
[12] A. Kutuzov, L. Øvrelid, T. Szymanski, E. Vellidal, “Diachronic word embeddings and semantic shifts: A survey,” in Proceedings of the 27th international conference on computational linguistics, Santa Fe, New Mexico, USA, 2018, pp. 1384–1397.
[13] X. Tang, “A state-of-the-art of semantic change computation,” Natural Language Engineering, vol. 24, no. 5, pp. 649-676, 2018.
[14] W. L. Hamilton, J. Leskovec, D. Jurafsky, “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 2016, pp. 1489-1501.
[15] H. Gonen, G. Jawahar, D. Seddah and Y. Goldberg, “Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora,” in 38th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 538-555.
[16] T.R. Besold, K.-U. Kühnberger and E. Plaza, “Towards a computational- and algorithmic-level account of concept blending using analogies and amalgams,” Connection Science, vol. 29, no. 4, pp. 387-413, 2017, 10.1080/09540091.2017.1326463.
[17] H. Calvo, O. Méndez and M.A. Moreno-Armendáriz, “Integrated concept blending with vector space models,” Computer Speech & Language, vol. 40, no. C, pp. 79-96, 2016, 10.1016/j.csl.2016.01.004.
[18] Y. Nagai, T. Taura and F. Mukai, “Concept blending and dissimilarity: factors for creative concept generation process,” Design Studies, vol. 30, no. 6, pp. 648-675, 2009, 10.1016/j.desstud.2009.05.004.
[19] V. V. Bochkarev, S.V. Khrustilovor, A. V. Shevlyakova, “Recognition of Named Entities in the Russian Subcorpus Google Books Ngram,” Lecture Notes in Computer Science, vol 12469, pp. 17-28, Oct. 2020. 10.1007/978-3-030-60887-3_2.
[20] J. Weeds, D. Weir, D. McCarthy, “Characterising measures of lexical distributional similarity,” in Proceedings of the 26th International Conference on Computational Linguistics, Geneva, Switzerland, 2004, pp. 1015–1021.
[21] P. Pantel, “Inducing ontological co-occurrence vectors,” in Proceedings of the 43rd Conference of the Association for Computational Linguistics, Ann Arbor, USA, 2005, pp. 125–132.
[22] J. A. Bullinaria, J. P. Levy, “Extracting semantic representations from word co-occurrence statistics: A computational study,” Behavior Research Methods, vol. 39, pp. 510–526, 2007. 10.3758/BF03193020.
[23] Y. Xu, C. Kemp, “A Computational Evaluation of Two Laws of Semantic Change,” in Proceedings of the 37th Annual Meeting of the Cognitive Science Society,CogSci 2015, Pasadena, California, USA, 2015.
[24] K. Cavallin, “Automatic extraction of potential examples of semantic change using lexical sets,” in Proceedings of KONVENS 2012 (LThist 2012 workshop), Vienna, Austria, 2012, pp. 370-377.
[25] S. Mitra, R. Mitra, S. K. Maity, M. Riedl, C. Biemann, P. Goyal, A. Mukherjee, “An automatic approach to identify word sense changes in text media across timescales,” Natural Language Engineering, vol. 21, no. 5, pp. 773-798, 2015. 10.1017/S135132491500111X.
[26] D. Pantel, D. Turney, P. Panchanet, “From frequency to meaning: Vector space models of semantics,” J. Artif. Intell. Res., vol. 37, no. 1, pp. 141–188, 2010.
[27] J. Bullinaria, J. Levy, “Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-Lists, Stemming and SVD,” Behavior Research Methods, vol. 44, pp. 890-907, 2012.
[28] T. Mikolov, et al., “Distributed representations of words and phrases and their compositionality,” Advances in Neural Information Processing Systems, vol. 26, pp. 3111–3119, 2013.
[29] L. Ferrmani, M. Lapata, “A Bayesian model of diachronic meaning change,” TACL, no. 4, pp. 31–45, 2016.
[30] Z. Yao, Y. Sun, W. Ding, “Dynamic word embeddings for evolving semantic discovery,” in Proceedings of the eleventh ACM international conference on web search and data mining, Marina Del Rey, CA, USA, 2018, pp. 673–681.
[31] M. Giulianielli, M. Del Tredici, R. Fernandez, “Analysing Lexical Semantic Change with Contextualised Word Representations,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3960–3973.
[32] H. Dubossarsky, E. Grossman, D. Weinschall, “Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models,” in Proceedings of the 2017 Conference on Empirical Methods in Natural
Language Processing, Copenhagen, Denmark, 2017, pp. 1136–1145.

[33] P. Basile, B. McGillivray, “Exploiting the Web for Semantic Change Detection,” Lecture Notes in Computer Science, vol. 11198, pp. 194–208, 2018. 10.1007/978-3-030-01771-2_13.

[34] C. Rohrdanz, A. Haufl, T. Mayer, M. Butt, D. A. Keim, F. Plank, “Towards Tracking Semantic Change by Visual Analytics,” in The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, 2011, pp. 305-310.

[35] V. Kulkarni, R. Al-Rou, B. Perozzi, S. Skiena, “Statistically significant detection of linguistic change,” in Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 2015, pp. 625–635.

[36] Y. Goldberg, J. Orwani, “A dataset of syntactic-ngrams over time from a very large corpus of English books,” in Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, Atlanta, Georgia, USA, 2013, pp. 241–247.

[37] A. Badjuar, S. Chen, P. Kambatta, T. Tran, A. Wang, K. Yu, P. Intrevedo, D. Brizan, “Quantum criticism: an analysis of political news reporting,” Machine Learning and Applications: An International Journal (MALIJ), vol. 7, no. 1/2, Jun. 2020.

[38] V. Solovyev, V. Ivanov, “Dictionary-Based Problem Phrase Extraction from User Reviews,” Lecture Notes in Computer Science, vol. 8655, pp. 225–232, 2014. 10.1007/978-3-319-10816-2_28.

[39] Q. Wang, Y. Zhou, T. Ruan, D. Gao, Y. Xia, P. He, “Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition,” J. Biomed. Inform., vol. 92, 103133, Apr. 2019. 10.1016/j.jbi.2019.103133.

[40] R. Gareev, M. Tkachenko, V. Solovyev, A. Simanovsky, V. Ivanov, “Introducing baselines for Russian named entity recognition,” Lecture Notes in Computer Science, vol. 7816, pp. 329-342, 2013.

[41] R. Shamagat, “Named entity recognition: a literature survey,” Cent. Indian Lang. Technol., Bombay, India, 11305R013, Jun. 2014.

[42] J. Li, A. Sun, J. Han, C. Li, “A survey on deep learning for named entity recognition,” IEEE Trans. Knowl. Data Eng., Mar. 2020. 10.1109/TKDE.2020.2981314.

[43] E. F. Tjong Kim Sang, F. De Meulder, “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition,” in Proceedings of the seventh conference on Natural language learning atHLT-NACL, 2003, vol. 2003, pp. 142–147.

[44] A. Baevski, S. Edunov, Y. Liu, L. Zettlemoyer, M. Auli, “Cloze-driven pretraining of self-attention networks,” CoRR, vol. abs/1903.07785, 2019.

[45] T. Liu, J.-G. Yao, Ch.-Y. Lin, “Towards improving neural named entity recognition with gazetteers,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 5301–5307.

[46] B. Y. Lin, D.-H. Lee, F. F. Xu, O. Lan, X. Ren, “AlpacaTag: An active learning-based crowd annotation framework for sequence tagging,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, Italy, 2019, pp. 58–63.

[47] N. Mchrlf, T. Gowda, F. Morstatter, N. Peng, A. Gallstyan, “Man is to Person as Woman is to Location: Measuring Gender Bias in Named Entity Recognition,” arXiv, abs/1910.10872v1, 2019.

[48] S. Khristoforov, V. Bochkarev, A. Shevlyakova, “Recognition of Parts of Speech Using the Vector of Bigram Frequencies,” Communications in Computer and Information Science, vol. 1086, pp. 132-142, Feb. 2020. 10.1007/978-3-030-39575-9_13.

[49] G. Sidorov, F. Velasquez, E. Stamatas, A. Gelbukh, L. Chanona-Hernandez, “Syntactic dependency-based N-grams as classification features,” Lecture Notes in Computer Science, vol. 7630, pp. 1-11, 2012.

[50] J. Goodfellow, Y. Bengio, A. Courville, Deep Learning, USA: MIT Press, 2016.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929-1958, 2014.

[52] D. Kingma, J. Ba, “Adam: A method for stochastic optimization,” arXiv, abs/1412.6980, 2014.

[53] A. Botev, G. Lever, D. Barber, “Nesterov’s accelerated gradient and momentum as approximations to regularised update descent,” CoRR, abs/1607.01981, 2016.

[54] M. Basseville and I. V. Nikiforov, Detection of Abrupt Changes: Theory and Application. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[55] H. Poor, O. Hadi, Quickest Detection, Cambridge: Cambridge University Press, 2008.

[56] M. Daumer, M. Fahl, “On-line Change-Point detection (for state space models) using multi-process Kalman filters,” Linear Algebra and its Applications, vol. 284, pp. 125-135, 1998.

[57] F. Akashi, H. Dette, Y. Liu, “Change point detection in autoregressive models with no moment assumptions,” arXiv, abs/1211.5201, 2016.

[58] W. A. Taylor, “Change-point analysis: A powerful new tool for detecting changes,” 2000. [Online]. Available: http://www.variation.com/cpa/tech/changepoint.

[59] M. Lavielle, “Using penalized contrasts for the change-point problem,” Signal Processing, vol. 85, pp. 1501–1510, Aug. 2005.

[60] R. Killick, P. Fearnhead and I.A. Eckley, “Optimal detection of change-points with a linear computational cost,” Journal of the American Statistical Association, vol. 107, no. 500, pp.1590-1598, 2012.

[61] V. V. Bochkarev and A. V. Shevlyakova, “Calculation of a confidence interval of semantic distance estimates obtained using a large diachronic corpus,” Journal of Physics: Conference Series, vol. 1730, 012031, 2021.

[62] Oxford Languages. [Online]. Available: https://www.oxforddictionaries.com/grammatical-sentence/.

[63] O. Linzmayer, Apple Confidential 2.0: The Definitive History of the World’s Most Colorful Company. 2nd ed. San Francisco, CA, USA: No Starch Press, 2004.

[64] V. Bochkarev, A. Shevlyakova, V. Solovyev, “A Method of Semantic Change Detection Using Diachronic Corpora Data,” Communications in Computer and Information Science, vol. 1086, pp. 94-106, Feb. 2020. 10.1007/978-3-030-39575-9_10.

[65] Macmillan Dictionary. [Online]. Available: https://www.macmillandictionary.com/.

[66] Cambridge Dictionary. [Online]. Available: https://dictionary.cambridge.org/.

[67] EXPLORE THE UNKNOWN. Selected Documents in the History of the U.S. Civil Space Program Volume V: Exploring the Cosmos, J. M. Logsdon, Ed. Washington, USA: NASA History Office, 2001.

[68] The Routledge Encyclopedia of Mark Twain. Edited By J. R. LeMaster, J. D. Wilson, eds., 1st ed. New York, USA: Routledge, 1993.

[69] Report into the Loss of the SS Titanic, S. Halpern, ed., Reprint ed. The History Press, 2016.

[70] P. Schweizer, R. Schweizer, The Bushes: Portrait of a Dynasty, 1st ed. USA: Doubleday, 2004.

[71] O. Parkes, British Battleships 1860–1950: A History of Design, Construction and Armament, Annandale, MS, USA: Naval Institute Press, 1990.
[72] D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, and N. Tahmasebi, “SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection,” in Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Spain, 2020, pp. 1–23.
VLADIMIR V. BOCHKAREV – graduated from the Faculty of Physics at Kazan Federal University (Russia) in 1991. Currently, he is a Scientific Researcher at Kazan Federal University. His scientific interests include mathematical modelling, cosmology, data analysis, corpus linguistics and quantitative linguistics.

STANISLAV V. KHRISTOFOROV – graduated from the Faculty of Physics at Kazan Federal University (Russia) in 2014. Currently, he is a Scientific Researcher at Kazan Federal University (Russia) and a Team Leader at “TGT Oilfield Services” company. His research interests include machine learning and artificial neural networks.

ANNA V. SHEVLYAKOVA – graduated from the Faculty of English and German at Pyatigorsk State Linguistic University (Russia) in 2008. She received her PhD degree in philology from the same university in 2011. Currently, she is a Senior Researcher at Kazan Federal University (Russia). Her scientific interests include cognitive, corpus and statistical linguistics.

VALERY D. SOLOVYEV – Professor, Principal Researcher at the Institute of Philology, Kazan Federal University (Russia). He defended his doctoral dissertation in the field of mathematical cybernetics at the Computing Center of the Russian Academy of Sciences in 1996. His scientific interests include mathematical and computational linguistics. He is an author of 4 monographs.