Automated Analysis of Micro-contexts of Word for Construction of Its Lexicographic Description

Nadezhda Lagutina, Yuliya Tsofina, Ilya Paramonov,
Ksenia Lagutina, Natalia Kasatkina

P.G. Demidov Yaroslavl State University,
Sovetskaya Str. 14, 150003, Yaroslavl, Russia
lagutinans@gmail.com, yutch@bk.ru, ilya.paramonov@fruct.org,
lagutinakv@mail.ru, ninet75@mail.ru

Abstract. The authors present an approach to automate analysis of micro-contexts of a word. The approach unites in a single software tool search of micro-contexts, calculation of their statistical characteristics, detection, visualization, and comparison of micro-contexts. The tool was applied to solve the task of constructing a lexicographic description of an interjection at the stage of analysis of its meanings in context. This experiment proved the usefulness of the tool and showed 12 times increase of total performance.

1 Introduction

The problem discussed in this paper is automation of analysis of micro-contexts of a word. Context is treated as a characteristic of word use that allows to determine its meaning. In linguistics context can be considered as an arbitrary factor (e.g., linguistic, physical, social, etc.) that affects interpretation of linguistic signs [1]. It can include language environment, a situation of verbal communication, or object surroundings. In turn, micro-context is the nearest lexical environment of a word [2].

Search and examination of a large number of micro-contexts containing a particular word can be used to establish its meanings, either represented in vocabularies or missing from there but existing in written texts. Frequencies of micro-contexts comprise an essential factor that allows to determine which meanings are dominant in modern linguistic consciousness and which ones are peripheral [3].

In order to find out micro-contexts and calculate their characteristics, a linguist determines words and phrases that occur near a given word and searches them in a text corpus. For this purpose, national corpora, such as COCA, BNC, or Russian National Corpus, are often used. Some corpora also allow to calculate frequency of occurrence automatically. However, detection and comparison of micro-contexts has to be done manually.

In order to use a custom text corpus, a researcher has to do all work herself. This process is routine, laborious, and error-prone when done by a human,

whereas a computer would perform this task quickly and correctly. That is why even partial automation of search, detection, overview, and comparison of micro-contexts of a word seems topical.

In this paper the authors examined state-of-the-art and proposed an approach and a tool for automation of analysis of micro-contexts of a word. The paper is structured as follows. Section 2 states the problem from a linguistic point of view. In Section 3 existing tools and related work are discussed. It is also shown that these tools cannot automate analysis of micro-contexts well enough for comfortable use. Section 4 is devoted to the authors' approach including the description of an algorithm proposed to solve the problem under consideration. Section 5 describes a tool developed by the authors on the basis of this approach. Section 6 reveals the results of experiments conducted by a linguist applying the developed tool and assessment of its usefulness. In Section 7 these results are discussed in a broader context including applicability of the tool for knowledge extraction and other linguistic tasks and its further development directions. Conclusion summarizes main results of the paper.

2 Problem statement

From a linguistic point of view the problem of analysis of micro-contexts can be stated as follows. A linguist chooses a word and finds all its occurrences in a text corpus. For each occurrence, she detects a micro-context (i.e., words nearest to the given word, also denoted as contextual words), determines semantic features (i.e. components of word meanings) of the word in each micro-context, and calculates statistical characteristics of all micro-contexts [4].

In order to determine semantic features, a linguist analyzes use of the given word in extended context, i.e., considers a sentence or several sentences in a word neighborhood. For each micro-context the frequency of occurrence is calculated, then all micro-contexts are sorted by descending frequency. This allows to demonstrate dominance of particular micro-contexts as well as to discover appropriate examples to illustrate separate semantic features.

The main challenge of micro-context processing consists in the following. A linguist finds 10–100 contextual words and juxtaposes each of them with a particular semantic feature. During this process calculation of statistical characteristics is necessity but not the main difficulty. The most important issue is to find a particular semantic feature in a particular micro-context, as it requires to remember or to find again and again a lot of information: extended context for each case, previous cases of contextual word use, word co-occurrences. In this paper we name this process micro-context detection. Its automation would significantly facilitate linguist's work and increase its performance and quality.

3 Existing tools and related work

The problem of analysis of micro-contexts is often solved manually by linguists, although there are many automated linguistic analysis tools for natural language

search, statistical analysis of texts, text structure definition, etc. [5] These tools can be divided into two groups: multi-function systems that implement many algorithms of text retrieval and statistical analysis and specialized applications, each of which solves a particular research task.

An example of a system from the first group is the Corpus of Contemporary American English (COCA, `https://corpus.byu.edu/coca`). It provides such functions as search of particular words, phrases, grammatical forms, synonymic series; visualization of context with graphs and diagrams; calculation of word usage frequency, and so on.

The similar open access tool is the Russian National Corpus (RNC, `http://www.ruscorpora.ru/en/`). It is an information system based on a collection of texts in Russian primarily intended to support research in vocabulary and grammar. This tool allows searching by particular words and phrases, searching by grammatical criteria, browsing context of search results, and calculating word usage frequency. It can also provide metadata of texts in the corpus that include size, genre, date of creation, information about authors, etc.

However, such systems cannot be used to analyze a custom corpora created by a linguist for a particular research. Also they do not provide the possibility to detect micro-contexts. In most cases a researcher cannot do sentiment analysis based on the context provided by the system. Finally, using the whole functionality provided by national corpora usually requires payment.

The Computational Social Science Laboratory at the University of Southern California developed another multi-functional linguistic system called TACIT [6]. It contains three main components: plugins for automatic collection of text from online sources, a module for corpus management, and plugins for analysis including algorithms for word count, sentiment analysis, clustering, classification, etc. However, using this toolset requires skills of software compiling and even programming to integrate custom plugins to the system. Besides, application of TACIT to non-English texts requires additional research.

The Sketch Engine system (`https://www.sketchengine.co.uk/`) proposes wider range of features than national corpora and TACIT. It allows to load and use custom texts and finds word sketches. The word sketch means that for a particular word the system counts and displays frequency of occurrence and collocations with others. Also Sketch Engine is able to calculate statistical characteristic based on collocations like mutual information score or logDice. Nevertheless, this system does not provide features for detection and investigation of micro-contexts.

The second group of automated linguistic tools consists of specialized applications developed to solve particular text analysis problems. LIWC (`http://liwc.wpengine.com/`) is a proprietary tool for classification of texts into psychological categories. It is used in many research devoted to sentiment analysis [7,8]. LIWC provides searching words in texts and determines sentiment polarity. However, this tool does not allow to detect micro-contexts.

Several linguistic tools visualize text structure, e.g., SinTagRus automatically builds and visualizes structure of sentences in Russian [9]. Another similar tool

[10] helps to determine similarity or difference of texts in natural language. These tools can potentially be used by linguists to investigate micro-contexts and cases of their usage. Nevertheless, they do not automate the most laborious tasks of the micro-context analysis: search and detection of words and calculation of the frequency of occurrence.

Summarily, there are no currently available tools of linguistic analysis that allow to automate analysis of micro-contexts as a whole. They can be useful for solving some subtasks, but detection and visualization of all found contextual words and most of calculations are left beyond automation.

4 Proposed approach to automation of analysis of micro-contexts

Our approach to automation of the task under consideration unites in a single tool search of micro-contexts, calculation of their statistical characteristics, detection, visualization, and comparison of micro-contexts. Such a tool would quickly and accurately perform a significant part of the routine processing of a large amount of information. Besides, it would provide functionality that allow linguists to quickly determine semantic characteristics of any particular micro-context.

For this approach the authors defined a formal model of micro-context as follows. Consider a text in natural language as a sequence of words $T = (w_1, w_2, \ldots, w_n)$. Let d be a word for which micro-contexts should be searched. The set of surrounding words for d is

$$W_d = \{w_i : w_j = d,\ 0 < |i - j| \leq 2,\ j = 1, 2, \ldots, n\}.$$

From this set we remove words of general use W_g and words W_n to be ignored in the linguist's particular research (for example, evidently neutral in the word context). Besides, some words W_m can be marked by a linguist as mandatory to investigate regardless of their location. Then we define micro-contexts of the word d as a set

$$M_d = W_d \cup W_m \setminus W_g \setminus W_n.$$

Such a choice of the context model is based on the idea that the inclusion of two neighboring words from each side (the search radius equals 2) gives a good initial approximation to detect a micro-context. An increase of the searching radius for neighboring words leads to a large number of extraneous words appearing in the set M_d that would apparently be removed manually by a linguist. Besides, the set W_m is introduced in order to take into account already known micro-contexts (e.g., from vocabulary) that can occur outside the specified radius. This approach allows to achieve a good balance between the speed and the quality of text processing.

For the purpose of research a linguist specifies a word d, sets W_g, W_n, W_m, and a text corpus to analyze micro-contexts.

5 A tool for automated analysis of micro-contexts

The developed tool for automated analysis of micro-contexts of a word is based on the approach described in the previous section. It is a cross-platform application written in Java with the use of JavaFX UI library. It is available under the Open Source MIT license at `https://github.com/ivparamonov/word-micro-contexts-searcher`.

The program takes four text files specified by the user. The first one contains a word for which micro-contexts are analyzed. The second one consists of words of general use. The third file includes words that should be ignored in text processing due to a conducted research. The fourth file contains words that should be taken into account regardless of their location relative to the specified word. Also the user should provide a path to the text corpus (a directory with files in the docx format) chosen for processing.

The tool provides the following functionality:

1. Search over the corpus for the given word and its micro-contexts.
2. Forming a list of contextual words. A linguist can browse the found micro-contexts and their features. All contextual words are displayed in the table on the left side of the main window (Figure 1). For each micro-context the frequency of occurrence is shown. The last column is filled with marks for words found automatically (false) and added by the linguist manually (true).
3. Editing the list of contextual words. A linguist can add a word to the list or remove it depending on the conducted linguistic experiment. For example, she can delete a contextual word found by the tool, but unrelated to the meaning of the given word under particular research.
4. Search of a contextual word in the list. This function is necessary, because the list of contextual words can be too large for manual browsing. In addition, the list can be sorted alphabetically.
5. Calculation of the frequency of a contextual word and sorting the word list by the frequency. For clarity, the frequencies of selected contextual word can be visualized as a histogram in the right part of the main window.
6. Plotting a histogram of frequencies for several words from the generated list. A linguist can select several context words to compare the frequencies of their occurrence. The selected words are displayed in a separate list. An example of such a list containing four selected context words is shown in Figure 2 in the right part of the application window. The histogram of corresponding frequencies is displayed below.
7. Browsing an extended semantic context for a selected contextual word. We define the extended context as a sentence or a significant part of it, where the word under consideration occurred. For each word from the list of micro-contexts a linguist can browse all the cases where this word was found. The extended semantic context is displayed in a separate window. Figure 3 on the left shows an example of the found extended context for the contextual word “говорит” (“says”). The selected word is highlighted in the text in capital letters.

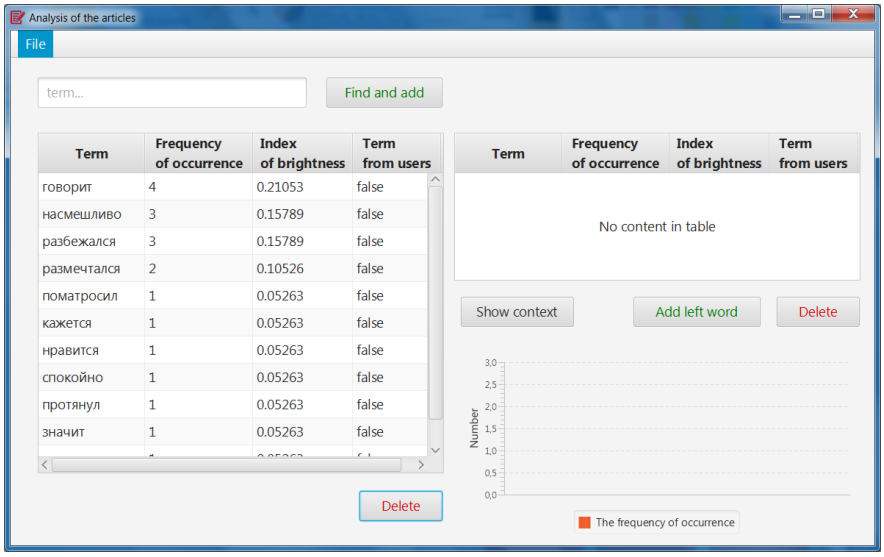

Fig. 1. Main application window

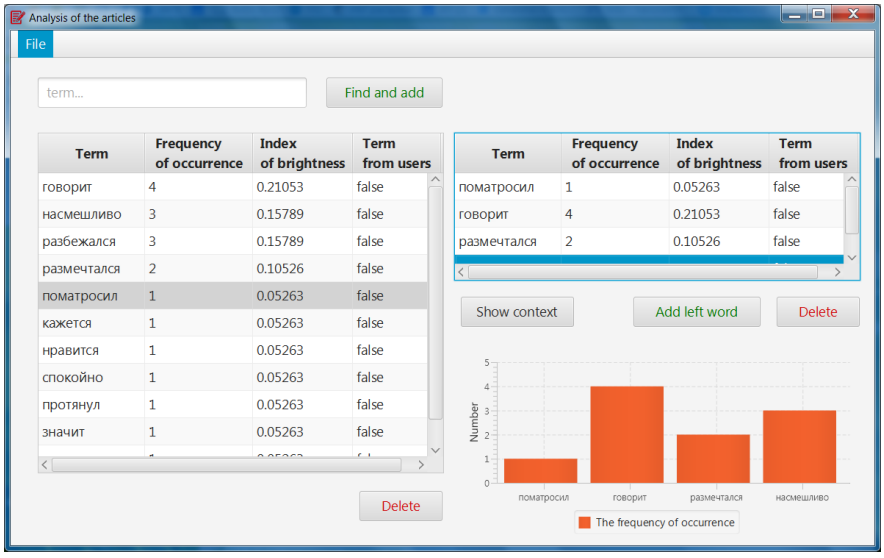

Fig. 2. Histogram of occurrence of words

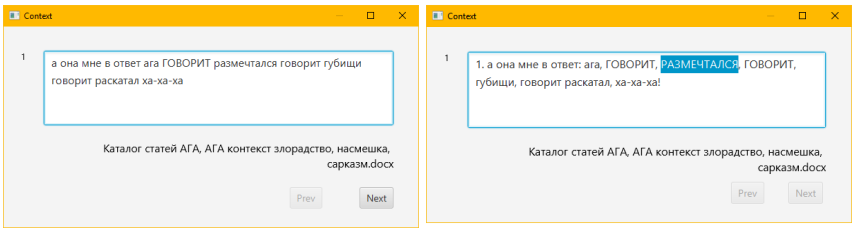

Fig. 3. Browsing extended context for one word (left) and two words (right)

8. Search of the extended semantic context of two selected words from the list. The most interesting context for studying the semantic features of a word consists of several contextual words, most often two. A linguist can select two contextual words and browse all the cases where these words appear in one sentence. In Figure 3 on the right the extended semantic context of the words "говорит" ("says") and "размечтался" ("are you kidding me?" in this context) was found.

6 Experiments

To prove the effectiveness and efficiency of the developed tool, the authors applied it to construct lexicographical descriptions of interjections. This task is especially complicated due to meaning diffuseness typical for the chosen part of speech. One of the ways to solve the stated problem is to use a method of psycholinguistic description of word semantics [3]. It consists of three stages.

During the first stage, meanings of interjections are generalized on the basis of existing dictionary entries. This process reveals dominant and peripheral word meanings.

The second stage includes analysis of interjection meanings in context. It is often difficult to determine and describe meanings of interjections as in oral speech their main semantic features are communicated through intonation, mimics, and gestures of a speaker. In written speech, intonation can be defined only with the help of lexical markers distinguished from micro-contexts and examination of the corresponding extended context. Such lexical markers are important for the lexicographical description, as they allow readers to identify the semantics of interjections and phrases in writing, which is specifically helpful for non-native speakers. Thereby, the second stage is devoted to determination of lexical markers and discovery of word meanings that were not found during the first stage.

The third stage allows to verify discovered meanings by a psycholinguistic experiment with native speakers to decrease subjectivity. After this process, semantic structure of an interjection can be described in terms of nucleus and periphery.

The developed tool is used at the second stage of this method. A linguist forms a corpus of texts to study a particular semantic feature of a word. The

Table 1. Example of a search of the interjection “ага” in semantic feature “mockery, sarcasm, malevolence” (extraction)

Contextual word	Frequency of occurrence	Example of extended semantic context
говорит	4	а она мне в ответ: ага, **говорит**, размечтался, **говорит**, губищи, **говорит** раскатал
насмешливо	3	я уже заплатил кому надо, ага, — **насмешливо** протянул он
разбежался	3	ага, сейчас, **разбежался** Романец
размечтался	2	а она мне в ответ: ага, **говорит**, **размечтался**, **говорит**, губищи, **говорит** раскатал
поматросил	1	слушай, Верка, это вообще не ваше дело, — ага, **поматросил** и бросил
значит	1	Они пришли, увидели детей, говорят: Ага, **значит**, у вас детский приют
кажется	1	начал было Грым, но остановился. — Ага. **Кажется**, понимаю.

tool organizes a list of micro-contexts and calculates word frequency. It should be mentioned that interjection micro-context essentially depends on a particular word form, which is important for its meaning, that is why stemming is not used when forming a list of micro-contexts. The list is sorted by descending contextual word frequency. For each micro-context a linguist studies the extended semantic context to define additional lexical markers. Additionally, the extended context allows to select illustrative material (i.e., word usage examples) for the lexicographical description.

In the described experiment the interjection “ага” was investigated. A linguist prepared three text corpora in Russian, each of average size of 1 100 words. The first corpus consists of 43 texts, containing the interjection “ага”, which has the semantic feature “mockery, sarcasm, malevolence”. The second corpus consists of 33 texts, containing “ага” with the semantic feature “guess, sudden insight, remembrance”. The third corpus consists of 44 texts, containing “ага” with the semantic feature “reflection, deliberation”.

In Table 1, there are some results retrieved by the tool for the first test corpus. It is shown that in this corpus the most frequent collocation is “ага” + “говорит”. Additional analysis of the extended semantic context performed by a linguist allowed to determine the following micro-contexts that communicate the semantic meaning of the interjection: “ага” + “размечтался”, “ага” + “разбежался”, “ага” + “поматросил”, “ага” + “насмешливо”. These contextual words perform as lexical markers that communicate the semantic feature under analysis.

Moreover, a new lexical marker of the semantic feature “mockery, sarcasm, malevolence” was found: “губищи раскатал, ха-ха-ха”. Examination of the extended semantic context of two collocations (“ага” + “говорит” + “размечтался”)

Table 2. Example of a search of the interjection “ага” in semantic feature “guess, sudden insight, memory” (extraction)

Contextual word	Frequency of occurrence	Example of extended semantic context
вот	7	Ага, так **вот** зачем все это было, — чтоб я писал.
понял	3	Ага, **понял** я, вот откуда взялись эти 333.33 %.
обрадовался	2	Ага, — **обрадовался** я, — значит, на предыдущую мне должны дать скидку
вспомнил	2	Ага, вот, — **вспомнил** он.
нашла	1	Ага, **нашла**. Вот оно.

Table 3. Example of a search of the interjection “ага” in semantic feature “reflection, deliberation” (extraction)

Contextual word	Frequency of occurrence	Example of extended semantic context
значит	3	Это сколько же здесь на наши? Ага. **Значит**, Валька я верну сразу
понимаю	2	начал было Грым, но остановился. — Ага. Кажется, **понимаю**.
понятно	2	Шестой . . . ага . . . **понятно**
запоминаю	1	оглянулся и **запоминаю**: ага, слева кофейня, справа обувная лавка.
смекнул	1	Ага, — **смекнул** Панов и задумался.

enhanced lexicographical description with the following word usage example: “ага, говорит, размечтался, говорит, губищи, говорит раскатал, ха-ха-ха!”

Table 2 shows micro-contexts that appear most frequently in experiments with the second text corpus that was created to investigate semantic features of the interjection “ага”: guess, sudden insight, remembrance.

Table 3 describes the main results of the third text corpus processing.

To estimate time costs, the first text corpus was processed manually by a linguist. This work took 120 minutes. Most of the time was devoted to word search and frequency calculation. Automated text processing with the help of the developed tool lasted approximately 10 minutes, including 20 seconds for automatic search of micro-contexts and calculations, and the rest of time for analytical part of the research.

Automated detection of micro-context list with further manual editing became the most significant factor of linguist’s performance increase. Having this list allowed not only to automate calculation of statistical characteristics, but also to efficiently organize extraction of lexical markers, quick search of extended context for clarifying semantic features, and search of illustrative material.

7 Discussion

The previous section contains a description of an experiment devoted to analysis of micro-contexts of interjections conducted with the help of the developed tool. In this section we discuss other possible directions of its usage.

Search of micro-contexts of a word and ranking them by the word frequency allow to use this tool not only for the construction of lexicographical word descriptions, but also for updating existing dictionaries. It is required, because lexical surroundings of a word can change in time [11].

Automated search of micro-contexts of a word can be used to form associative dictionaries. Words in such dictionaries are grouped into fields where the central word unites the meanings of words surrounding it, which either have similar meanings or can be psychological associated with it [12]. Associative dictionaries are popular among linguists as they contain the most common words of modern literary language. Detection of micro-contexts and their ranking by the word frequency can be a basis for word entries in an associative dictionary.

Among various types of dictionaries there is a special place for dictionaries devoted to writers and their works. They contain and explain words used in writings of particular authors. These dictionaries are used to study language and literary style of a writer, history of a literary language. They give clues to understanding author's texts correctly. For instance, there is a dictionary of Shakespeare's medical terminology, which includes terms and corresponding text references [13]. This dictionary explains the meanings of a term related to the Shakespeare's epoch that can differ from the ones used nowadays. The tool described in this paper can help a linguist in construction of such dictionaries, as is allows to find all occurrences of a particular word and browse corresponding enhanced contexts.

Functionality of the developed tool can be enhanced to solve other similar tasks of text processing. For example, it can be adapted to construction of text corpora containing words characterized by a certain grammatical feature. This feature might require presence of particular words (e.g., verb forms) or word collocations [14]. To leverage the tool for such a task, it should be just supplemented with a feature to process each text of a corpus separately and subdivide texts into groups.

Adding a feature of comparing word frequency of micro-contexts by year would help to determine obsolescent words and neologisms. This function would not only allow to update dictionaries effectively, but to study language history.

Calculation of frequencies for pairs "a word" + "a contextual word" can be a basis for automatic formation of lexico-syntactic patterns that reflect semantic and syntactic features of a particular text fragment. Such patterns can be used for knowledge extraction [15], determining mood of a text and its emotional–expressive colouring [16], which, in turn, can be applied in systems of automatic text classifications.

To sum up, the described tool can be helpful in constructing and updating dictionaries of various types. Also, current functionality can be easily enhanced

for search of lexico-syntactical patterns, construction of corpora based on certain grammatical features, and determination of obsolescent words and neologisms.

8 Conclusion

In the paper the authors proposed an approach to automate analysis of micro-contexts of a word and implemented it in a tool that can be easily used by linguists in their research. The application takes the most of routine work on itself and allows a linguist to concentrate on the analytical part of research.

It should be mentioned that the intention of the research was to automate analysis of micro-contexts as full as possible. That is why the developed tool is not as versatile as many existing text processing systems (RNC, Sketch Engine, etc.). However, for the task under consideration it reduces the amount of manual efforts significantly.

The experiment on construction of lexicographic description of a Russian interjection proved the usefulness of the tool and showed 12 times increase of total performance.

The developed tool is available under the Open Source MIT license at `https://github.com/ivparamonov/word-micro-contexts-searcher`.

Acknowledgements

The authors would like to thank professor I. A. Sternin from Department of General Linguistics and Stylistics of Voronezh State University for discussion of experimental results; D. V. Grushevskaya and E. E. Zolotova, students of P.G. Demidov Yaroslavl State University, for programming the presented in the paper tool.

References

1. I. Kecskes, "Dueling contexts: A dynamic model of meaning," *Journal of Pragmatics*, vol. 40, no. 3, pp. 385–406, 2008.
2. E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 873–882.
3. I. A. Sternin, "Psikholingvisticheskoe znachenie slova [Psycholinguistic word meaning]," *Vestnik Rossiiskogo universiteta druzhby narodov. Seriya: Russkii i inostrannye yazyki i metodika ikh prepodavaniya*, no. 1, pp. 5–13, 2011.
4. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
5. R. Iliev, M. Dehghani, and E. Sagi, "Automated text analysis in psychology: Methods, applications, and future developments," *Language and Cognition*, vol. 7, no. 2, pp. 265–290, 2015.

6. M. Dehghani, K. M. Johnson, J. Garten, R. Boghrati, J. Hoover, V. Balasubramanian, A. Singh, Y. Shankar, L. Pulickal, A. Rajkumar *et al.*, "TACIT: An open-source text analysis, crawling, and interpretation tool," *Behavior research methods*, vol. 49, no. 2, pp. 538–547, 2017.
7. R. Wynn, S. Oyeyemi, J.-A. Johnsen, and E. Gabarron, "Tweets are not always supportive of patients with mental disorders," *International Journal of Integrated Care*, vol. 17, no. 3, p. A149, 2017.
8. K. A. Stevens, K. Ronan, and G. Davies, "Treating conduct disorder: An effectiveness and natural language analysis study of a new family-centred intervention program," *Psychiatry research*, vol. 251, pp. 287–293, 2017.
9. I. Boguslavsky, "SynTagRus—a deeply annotated corpus of Russian," *Les émotions dans le discours-Emotions in Discourse*, pp. 367–380, 2014.
10. K. S. Choi, K. S. Jeong, and S. D. Kim, "A MVC framework for visualizing text data," *Journal of Intelligence and Information Systems*, vol. 20, no. 2, pp. 39–58, 2014.
11. M. Yang, D. Dai, L. Shen, and L. Van Gool, "Latent dictionary learning for sparse representation based classification," in *Proceedings CVPR 2014*, 2014, pp. 4138–4145.
12. N. V. Ufimtseva, "The associative dictionary as a model of the linguistic picture of the world," *Procedia-Social and Behavioral Sciences*, vol. 154, pp. 36–43, 2014.
13. B. H. Traister, "Shakespeare's medical language: A dictionary," *Shakespeare Studies*, vol. 42, pp. 309–314, 2014.
14. P. Kasparaitis and T. Anbinderis, "Building text corpus for unit selection synthesis," *Informatica*, vol. 25, no. 4, pp. 551–562, 2014.
15. J.-E. Kim, K. Park, J.-M. Chae, H.-J. Jang, B.-W. Kim, and S.-Y. Jung, "Automatic scoring system for short descriptive answer written in Korean using lexico-semantic pattern," *Soft Computing*, pp. 1–9, 2017.
16. K. Schouten, F. Baas, O. Bus, A. Osinga, N. van de Ven, S. van Loenhout, L. Vrolijk, and F. Frasincar, "Aspect-based sentiment analysis using lexico-semantic patterns," in *International Conference on Web Information Systems Engineering*. Springer, 2016, pp. 35–42.