UDC 8.81᾿3
DOI https://doi.org/10.24919/2308-4863/75-1-22

Iryna BASARABA,
orcid.org/0000-0002-3209-9119
PhD in Philology,
Lecturer at the Foreign Languages Department
National Academy of the State Border Guard Service of Ukraine
(Khmelnytskyi, Ukraine) irynaborovyk2017@ukr.net

CHALLENGES ENCOUNTERED IN AUTOMATICALLY CLASSIFYING PHRASEOLOGICAL UNITS

The author argues that an urgent task related to automatic text processing is the automated classification of phraseological units in English texts, which is explained by the large amount of textual data, which, in the author's opinion, makes it impossible to manually analyze and classify all phraseological units present in the English texts under study. The article determines that there are currently different approaches to the classification of phraseological units in English texts and the author identifies the most common ones, namely: corpus approach; cognitive linguistic approach; grammatical approach; stylistic approach; interlinguistic approach. One of the possible ways to improve the accuracy and efficiency of software tools for automatic classification of phraseological units is to integrate them with machine learning algorithms. In the author's opinion, automatic classification of phraseological units in English texts is an important task in natural language processing (NLP), involving the identification and categorization of groups of words or phrases based on their semantic and syntactic properties. One of the most effective approaches to automatic phraseology classification is the integration of machine learning algorithms. The article argues that the integration of machine learning algorithms into the automatic classification of phraseological units involves training algorithms on a labeled dataset of phraseological units and their respective categories, which allows algorithms to learn the characteristics and properties of different types of phraseological units and accurately classify new instances. Automatic phrase classification consists of two main stages: feature extraction and classification. The choice of feature extraction method depends on the characteristics of the data set and available computing resources. Another approach to integrating artificial intelligence algorithms with English text is to use neural networks.

Key words: *phraseological units, artificial intelligence, machine learning, recognition, classification of phraseological units.*

Ірина БАСАРАБА,
orcid.org/0000-0002-3209-9119
доктор філософії в галузі філології,
викладач кафедри іноземних мов
Національної академії Державної прикордонної служби України
(Хмельницький, Україна) irynaborovyk2017@ukr.net

ПРОБЛЕМАТИКА АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ

Автор стверджує, що актуальним завданням, яке стосується автоматичної обробки текстів, є автоматизована класифікація фразеологічних одиниць англійських текстів, що пояснюється великим обсягом текстових даних, що на думку автора, унеможливлює «вручну» аналізувати та класифікувати всі фразеологічні одиниці, які присутні в англійських досліджуваних текстах. В статті визначається, що наразі існують різні підходи до класифікації фразеологічних одиниць англомовних текстів та автор виокремлює найбільш поширені, а саме: корпусний підхід; когнітивний лінгвістичний підхід; граматичний підхід; стилістичний підхід; міжлінгвістичний підхід. Одним із можливих шляхів підвищення точності та ефективності програмних засобів автоматичної класифікації фразеологічних одиниць є їх інтеграція з алгоритмами машинного навчання. На думку автора, автоматична класифікація фразеологізмів в англійських текстах є важливим завданням в обробці природної мови (NLP), передбачаючи ідентифікацію та категоризацію груп слів або словосполучень на основі їхніх семантичних та синтаксичних властивостей. Одним з ефективних підходів до автоматичної класифікації фразеологізмів є інтеграція алгоритмів машинного навчання. В статті стверджується, що інтеграція алгоритмів машинного навчання в автоматичну класифікацію фразеологічних одиниць передбачає навчання алгоритмів на маркованому наборі даних фразеологічних одиниць та їхніх відповідних категорій, що дозволяє алгоритмам вивчати характеристики та властивості різних типів фразеологічних одиниць і точно класифікувати нові екземпляри. Автоматична класифікація фразеологізмів складається з двох основних етапів: вилучення ознак і класифікація.

Вибір методу вилучення ознак залежить від характеристик набору даних і наявних обчислювальних ресурсів. Інший підхід до інтеграції алгоритмів штучного інтелекту з текстом англійською мовою полягає у використанні нейронних мереж.

Ключові слова: *фразеологічні одиниці, штучний інтелект, машинне навчання, розпізнавання, класифікація фразеологічних одиниць.*

Statement of the problem. Due to the rapid development of computer technology in recent years, more and more researchers are paying attention to the problems of automatic text processing. One of the most urgent tasks related to automatic text processing is the automated classification of phraseological units in English texts.

This is due to the following reasons. The huge amount of textual data available today makes it impossible to manually analyze and classify all the phraseological units present in English texts. Phraseological units often play a crucial role in the interpretation and creation of the language as a whole. Automatic classification of phraseological units has numerous applications in various fields, including language learning, translation, natural language processing, and computational linguistics. Phraseological units pose an important problem for machine translation systems, which may find it difficult to translate idiomatic expressions and well-organized phrases**.**

Traditionally, there are different approaches to the classification of phraseological units in English texts. Among the most common ones are the following: *corpus-based approach; cognitive linguistic approach; grammatical approach; stylistic approach; interlinguistic approach.*

The complexity of the task of classifying phraseological units determines the use of various software tools to solve it. Although the use of these tools for automatic classification of phraseological units in English texts has many advantages, there are also some disadvantages that should be taken into account. These include limited accuracy, difficulties with context, limited coverage, the need for customization and cost.

One of the possible ways to improve the accuracy and efficiency of automatic phraseological unit classification software is to integrate it with machine learning algorithms. This may involve training algorithms on large datasets of annotated text with human verification to confirm the accuracy of the classifications. By continuously improving the accuracy of algorithms through machine learning, software tools can become more effective in identifying and classifying a wider range of phraseological units, including those that are complex or ambiguous.

Analysis of recent research and publications. Today, linguistics is faced with the tasks of focusing on the general theoretical basis of phraseology, mainly on the basis of specific languages and texts (L. Nazarenko), clarification of the concept of phraseological units (O. Matviienkiv, I. Bekhta), study of their specificity in comparison with other linguistic units (N. Onishchenko, O. Matviienkiv, A. Markovska), including aspects of the emergence and integration of new phraseological expressions into the language system (M. Hamziuk), peculiarities of their use, study of synonymy, antonymy, polysemy and homonymy in phraseology (L. Skrypnyk, M. Sydorenko, V. Shkoliarenko), as well as the development of methods for studying phraseological units (S. Sukhorolska, L. Derevianko).

Statement of the task. Thus, the purpose of the article is to generalize methods in order to increase the efficiency of automatic classification of phraseological units in English-language texts and reduce the number of erroneous interpretations.

Presentation of the main research material. Phraseological units are multiword lexical units characterized by a certain degree of fixation or idiomatism of their components. In other words, phraseological units are a combination of words whose meaning is not necessarily derived from the meaning of its components, i.e. the words together can mean more than their sum of parts.

These linguistic structures are also known in the literature as phrases, stable expressions, and polysyllabic expressions. Although native speakers can easily learn such expressions, their interpretation poses a serious challenge for computing systems due to their flexible and heterogeneous nature. In addition, phraseological units are not as frequent in lexical resources as they are in real text, and this coverage problem can affect the performance of many natural language processing tasks.

Phraseological units are widely used by humans. The number of phraseological units expressed in polysyllabic expressions has the same order as the number of simple or isolated words.

We believe it is necessary to note that the appropriate definition of phraseological units (PhUs) is expressions consisting of two or more words that function as a single lexical unit. They have a fixed structure and their meaning almost cannot be derived from the meanings of their individual components. Examples of phraseological units in the English language include *"to kick the bucket", "to hold one's horses", "card up one's sleeve"*, etc. Phraseological units are an integral component of the language, which is widely used in everyday communication, literature and other forms of discourse (Ayto J., 2020). They provide a concise and often vivid way to express complex ideas, emotions, and actions. For example, the idiom *"to hold your horses"* means to be patient and wait, while the literal meaning of the words loses this idea. Phrases can be divided into several categories depending on their structure and meaning. The most common classification is based on their lexical composition, which includes idioms, phrases, phrasal verbs, and proverbs.

Idioms are stable expressions whose meaning cannot be derived from the individual words that make up the expression. Phrases are word combinations that occur frequently and have a strong lexical association. Phrasal verbs are verbal phrases that consist of a verb and one or more particles that together convey a specific meaning. Proverbs are expressions that convey a general truth or advice. A verbal phraseological unit is a phraseological unit that contains a single verb as its grammatical center. Verbal phraseological units perfectly illustrate general richness. Given this feature, as well as the fact that verb phrases have a paradigmatic gap, make us focus on this type of phraseology, which implies a very complex research line in terms of semantic identification and classification of phraseology (Ellis, 2018: 9).

Thus, it is important to study the nature of such linguistic structures so that we can design automatic methods for working with these units.

Automatic classification of phraseological units in English texts is an important task in natural language processing (NLP), which involves identifying and categorizing groups of words or phrases based on their semantic and syntactic properties. One of the most effective approaches to automatic phraseology classification is the integration of machine learning algorithms (Davis, Barret, 2012: 7).

Machine learning is a subfield of artificial intelligence (AI) that involves training algorithms that learn from data and make predictions or decisions without explicit programming. Integration of machine learning algorithms into automatic phraseological unit classification involves training algorithms on a labeled dataset of phraseological units and their respective categories, which allows the algorithms to learn the characteristics and properties of different types of phraseological units and accurately classify new instances. Automatic phraseological classification consists of two main stages: feature extraction and classification. Feature extraction involves identifying relevant features of phraseological units that can be used to create a vector representation. The choice of feature extraction method depends on the characteristics of the data set and available computing resources. Another approach to integrating AI algorithms with English text is to use neural networks (Levchenko, Romanyshyn, 2019: 289).

Neural networks are a type of machine learning algorithm modeled after the structure of the human brain. They consist of interconnected nodes that process information and learn from experience. Neural networks can be trained to recognize patterns in English text, such as the structure and meaning of phraseological units, and use this information to classify new expressions (Pawar, Mago, 2012: 12).

The subject area of automatic phraseological unit classification lies at the intersection of computational linguistics and natural language processing. It involves the development of algorithms and methods that can automatically identify and classify phraseological units, which are fixed or semi-fixed expressions in a language that have figurative or idiomatic meanings that cannot be easily derived from the meanings of their individual words. Philologists have conducted a large number of studies and asked themselves how to identify phraseological units in a text and have identified the main hypotheses that help identify these linguistic units in a text. Among the main ones are the following:

1) *The fixation hypothesis*. The clearer the verb phrase is, the higher its probability of being a verbal phraseological unit. Each component of the target verbal phrase can be replaced by their close synonyms in order to check whether the new verbal phrase loses its meaning. To verify the meaning of a new verbal phrase, it may be considered to use a reference corpus where evidence of such a phrase can be searched.

2) *The translation hypothesis*. The more literal the translation of a verbal phrase, the lower the possibility of it being a verbal phraseology. A word phrase can be translated from one language into another. Evidence of such translation is then sought in the reference corpus created in the target language.

3) *The hypothesis of intrinsic appeal and contextual correlation*. The higher the intrinsic attraction and the lower the contextual correlation in a word combination, the higher the probability that the word combination is a verbal phraseological unit. Statistical methods are used to determine the level of intrinsic attractiveness and contextual correlation between the terms of a verbal phrase and the terms of their context.

4) *The terminological domain hypothesis*. The greater the number of vocabulary terms outside the current domain for a verbal phrase, the higher the probability that it is a verbal phraseological unit. The use of out-of-domain terms in real-world VPhUs is quite common, so the terminology can be identified out-of-domain to determine whether a verbal phrase is a real-world VPhUs. Automatic classification of phraseological units can be useful in a variety of applications such as machine translation, text analysis, and natural language generation. It involves techniques such as pattern recognition, machine learning, and statistical analysis to identify and classify these expressions based on their syntactic and semantic properties. Some concepts are expressed in a language through a set of words or phrases that are intuitively used by native speakers, thus characterizing different cultural communities. Phraseology, which is considered the cultural heritage of a language community, is aimed at studying these blocks of words, which are commonly referred to as phraseological units. The study of phraseological units has become increasingly important in recent years, partly because the linguistic and computational linguistic community has realized that this phenomenon encompasses all components of a sentence, which includes various aspects of natural language (Thomas, 2016: 228).

The scientific and methodological apparatus for automatic classification of phraseological units has become a subject of research in the field of computer linguistics and natural language processing. Several approaches have been proposed for automatic classification of phraseological units, which can be broadly classified into three categories: *rule-based approach; statistics-based approach; and machine learning approach.*

Rule-based approaches rely on manually creating rules for identifying and classifying phraseological units. While these approaches can be effective in certain cases, they are limited by the complexity of the rules required and their inability to handle new or unseen expressions.

The advantages include: flexibility – the rule-based approach allows you to easily customize the classification of phraseological units to meet the specific needs of the program. Clarity – a rule-based approach is easy to interpret because the rules are clearly defined and can be easily understood by humans. High accuracy – when the rules are carefully designed and tested, the rule-based approach can achieve high accuracy in the classification of phraseological units.

We consider the disadvantages to be: limited scalability, i.e., the rule-based approach is limited by the ability to develop rules manually. Error-prone – a rule-based approach can be error-prone, especially when it comes to complex or ambiguous rules. Maintenance overhead – a rule-based approach requires manual maintenance, which can be time-consuming and expensive. Lack of adaptability – a rule-based approach may not be suitable for dynamic or changing environments where classification rules need to be updated frequently. Statistical approaches rely on the analysis of large corpora of text to identify and classify phraseological units. These approaches involve identifying words and phrases that occur simultaneously in a language and using statistical measures such as frequency and mutual information to identify and classify expressions. For example, statistical approaches can identify the phraseological unit *"a piece of cake"* by analyzing the frequency of its occurrence in a large corpus of text. These approaches can be effective for identifying common phraseological units, but may be limited in their ability to identify rare expressions or handle variations in expressions.

Advantages: Statistics-based approach is an approach based on statistical data analysis that provides an objective way to identify patterns in language use. The scalable approach can handle large amounts of data, making it scalable for analyzing text from different sources and domains. At the same time, statistical methods can be applied quickly and efficiently, which makes it possible to process large amounts of data in a short period of time. The statistical approach can be generalized to different languages and domains, making it applicable in different contexts.

The disadvantages include: limited accuracy – statistical models rely on patterns in data that may not always capture the complexity of language use; lack of contextual understanding – statistical models do not understand the contextual meaning of language and may not capture the nuances of meaning in certain situations; data bias – the accuracy of statistical models can be affected by data bias, where training data may not reflect real-world language use.

Machine learning approaches rely on learning algorithms to identify and classify phraseological units based on examples from a corpus of text. These approaches use techniques such as deep learning and neural networks to identify patterns and structures in the language and classify expressions based on their syntactic and semantic properties. For example, machine learning-based approaches can identify the phrase *"to bark up the wrong tree"* by training a neural network on a large corpus of text and identifying patterns of words and phrases that occur simultaneously. These approaches can be very effective in identifying and classifying phraseological units, but require a large amount of training data and computational resources.

Advantages: high accuracy – machine learning models can learn from data and improve their accuracy over time, making them highly accurate in identifying phraseological units in text; contextual understanding – machine learning models can capture the contextual nuances of language use, making them more accurate in determining the correct meaning of a phraseological unit in a particular context; reliability – machine learning models can cope with noise and data variations, making them more reliable in real-world settings; scalability – machine learning models can be trained on large amounts of data, which makes them scalable for analyzing text from different sources and domains.

The disadvantages include: large amounts of training data – machine learning models require large amounts of labeled training data to train, which can be time-consuming and expensive to obtain; adaptability – machine learning models can adapt to training data, which means they may not generalize well to new data; interpretability – there are tendencies for machine learning models to be difficult to interpret, making it difficult to understand how they make decisions; lack of transparency – models of machine learning can be a "black box" approach, which makes it difficult to understand how the algorithm arrived at its decisions (Twitto-Shmuel, Ordan, Wintner, 2015: 67).

In general, the existing scientific and methodological apparatus for automatic classification of phraseological units includes a number of approaches, each of which has its advantages and disadvantages. Future research is likely to focus on combining these approaches and developing more sophisticated algorithms that can cope with the complexity and variability of language. The most basic algorithms for finding the right language structures still started out using paper dictionaries and physical search and classification. However, while alphabetical searching in paper dictionaries is convenient, reverse searching (i.e. searching for a phraseological unit by its definition) in a classic paper dictionary is literally tantamount to looking for a needle in a haystack. This realization has led to repeated attempts to create onomasiological dictionaries designed to search for words by their concepts. Despite the huge number of dictionaries that have proven their effectiveness in this field, there is still a significant gap in knowledge about their use and principles of compilation. On the other hand, using reverse dictionaries is not as easy as it may seem. For example, if a potential user were to search for all entries containing the lexical alphabet, their search in One Look Reverse Dictionary would yield more than a hundred results, including spelling, alphabet, language, etc., which is much easier to handle than the entire set of entries. While looking up basic words when writing or editing text can be made easier with reverse dictionaries, processing sequences of phraseological units can be a much more difficult task with unpredictable results, since many phraseological units do not appear in bilingual dictionaries. Moreover, even if they do, their translations often have a different basic semantics or connotation from the original (Stevenson, Fazly, North: 2014: 7). It is logical to assume that finding a suitable phrase can be a time-consuming process, and a translator, writer, or journalist sometimes has to spend hours trying to establish a connection between the meaning the user has in mind and the phrases that exist in the target query. Given these difficulties, compiling phraseological onomasiological dictionaries with convenient and understandable keywords that act as entry points for the user can optimize the search for a phrase with a given meaning. They should prove useful when a writer or translator does not know the desired phrase in the target language.

As noted earlier, bilingual or multilingual phrase dictionaries do not necessarily provide equivalent translations of phraseological units in terms of semantics, underlying motivating structure (or images), connotation, or meaning in the context in which the phrase is used at the time. If translated equivalents are sometimes far from desirable, can the same obstacles be attributed to phraseological units within the same language? And how can we define phraseological units given such a mismatch of criteria? Although interest in this concept has increased dramatically in the last decade, some infrequent definitions could be found before. Phraseological units are co-referential units of language belonging to the same grammatical class, either partially coinciding or completely independent of each other in their lexical structure, containing both common and differential components that coincide or differ in their styles. Although the concept of phraseological units has probably crystallized by now, its practical significance for translations has been clearly taken into account before, so the concept of interlingual phraseological unit is introduced as "a phraseological unit that coincides in the morphological composition of significant components, in the type of the whole phraseological unit, but lacks an interlingual lexical invariant" (Levchenko, Romanyshyn, 2019: 288).

Obviously, the practical need to use phraseological units can arise both in translation and in monolingual communication. However, it is in the field of translation that the problem becomes apparent, while communicators are less likely to realize the need for a phraseological unit if they are not experienced professionals. It may seem commonly accepted that phraseological units should belong to the same grammatical class. However, as soon as the practical needs of translation are considered and given that the phraseological level is highly susceptible to transformations in the target text, it becomes clear that transpositions, modulations and other procedures widely used in translation can often lead to distortion of the grammatical structure of the change in the target text. At the same time, it is often difficult to find a more appropriate translation of a given phrase (such as *all out of the blue*) in open sources accessible through a search system such as Google. However, this example (as well as many other possible similar illustrations) allows us to ignore grammatical structure as a requirement to exclude phraseological synonyms.

Additional issues of defining phraseological units were considered by R. Piniero, who raised the question of whether phraseological units can be variants of the same phraseological units and whether phraseological units with different distribution and semantic combination should be considered. On the other hand, there is a possibility that the mismatch of images in coreferent phrases raises the question of their synonymy; in addition, the question of quasi-synonymy should be defined in the field of phraseology. Thus, Piniero concludes that the difficulties faced by researchers are the lack of criteria for classifying some phraseological units as such (phrases or fixed phrases), their different syntagmatic combinatorics, belonging to different areas of use, and their polysemy (Pinero, 2012: 226).

Thus, it can be said that the scientific and methodological apparatus for automatic classification of phraseological units is quite difficult to study and at the time of our research there are no perfectly working methodologies that could identify and classify phraseological units in any English-language text without error and distortion.

Conclusions. The automatic recognition and ordering of phraseological units in English texts is a key task in the field of natural language processing (NLP), which includes the identification and grouping of word units based on their semantic and syntactic nature. One of the most effective methods for automated phraseology classification is to combine it with machine learning algorithms. Prospects for further research include the development of software for the automatic classification of phraseological units based on the implementation of a hybrid method algorithm that combines a rule-based method and a machine learning method, which would increase the efficiency of classifying the phraseological units of English-language texts, reduce the classification time and increase the number of features for classification.

BIBLIOGRAPHY

1. Ayto J. The Oxford Dictionary of Idioms. *Oxford University Press,* 2020. Pp. 46–84.
2. Davis R., Barrett L. Lexical semantic factors in the acceptability of english support-verb-nominalization constructions. *ACM Trans,* 2012. Pp. 5–15.
3. Ellis N. Phraseology the periphery and the heart of language. *Phraseology in foreign language learning and teaching,* 2018. Pp. 1–13.
4. Levchenko O., Romanyshyn N. Modern approaches to automated identification of metaphor. *Philological series,* 2019. Pp. 288–298.
5. Pawar A., Mago V. Calculating the similarity between words and sentences using a lexical database and corpus statistics. Cornwell University, 2010. Pp. 1–14.
6. Rodríguez-Piñero A. Variación y sinonimia en las locuciones. *Revista de Lingüística y Lenguas Aplicadas*, 2012. Pp. 225–238.
7. Stevenson S., Fazly A., North R. Statistical measures of the semi-productivity of light verb constructions. *Proceedings of the Workshop on Multiword Express,* 2014. Pp. 1–8.
8. Thomas J. Discovering English with Sketch Engine. *A Corpus-Based Approach to Language Exploration,* 2016. Volume 12. P. 228.
9. Twitto-Shmuel N., Ordan N.,Wintner S. Statistical machine translation with automatic identification of translationese. *Proceedings of WMT*, 2015. Pp. 67–68.

REFERENCES

1. Ayto J. (2020) The Oxford Dictionary of Idioms. *Oxford University Press.* Pp. 46–84.
2. Davis R., Barrett L. (2012) Lexical semantic factors in the acceptability of english support-verb-nominalization constructions. *ACM Trans*. Pp. 5–15.
3. Ellis N. (2018) Phraseology the periphery and the heart of language. *Phraseology in foreign language learning and teaching*. Pp. 1–13.

4. Levchenko O., Romanyshyn N. (2019) Modern approaches to automated identification of metaphor. *Philological series*. Pp. 288–298.

5. Pawar A., Mago V. (2010) Calculating the similarity between words and sentences using a lexical database and corpus statistics. Cornwell University. Pp. 1–14.

6. Rodríguez-Piñero A. (2012) Variación y sinonimia en las locuciones. [Variation and synonymy in phrases] *Revista de Lingüística y Lenguas Aplicadas*. Pp. 225–238. [in Spanish]

7. Stevenson S., Fazly A., North R. (2014) Statistical measures of the semi-productivity of light verb constructions. *Proceedings of the Workshop on Multiword Express*. Pp. 1–8.

8. Thomas J. (2016) Discovering English with Sketch Engine. *A Corpus-Based Approach to Language Exploration*. Volume 12. P. 228.

9. Twitto-Shmuel N., Ordan N.,Wintner S. (2015) Statistical machine translation with automatic identification of translationese. *Proceedings of WMT*. Pp. 67–68.