# Creation of a Parallel Corpus of Contexts of Idioms in Russian and English Languages for Evaluating the Quality of Machine Translation

# Abstract

Idiomatic expressions translation is attracting considerable research interest due to the challenge this task presents in the field of machine translation. Previous studies have addressed the problem of automatic translation of idioms (Baziotis et al., 2023; Fazly et al., 2009; Takeuchi et al., 2007). In addition, a certain amount of literature has been published on the task of compiling corpora of potentially idiomatic expressions (PIEs) (Adewumi et al., 2022; Fadaee et al., 2018; Haagsma et al., 2020). However, the task of designing large parallel corpora of contexts of idiomatic expressions, especially in Russian — English translation task has been overlooked. The aim of the present study is to create a parallel corpus of contexts of Russian and English potentially idiomatic expressions and to use it to fine-tune and test a neural translation model. The current study combines continuous sampling method and corpus compilation methods in order to create a parallel corpus of idiomatic expressions. In addition, machine learning techniques are used for the evaluation of the compiled corpus. The project proposal might result in the elaboration of an effective tool for future research in the machine translation field, which may increase the accuracy of automatic translation of idioms by 15%.

*Keywords*: machine translation, parallel corpus, idiomatic expression, machine learning, Russian, English

# Introduction

At present, the translation of idioms is of significant research interest. Although the recent introduction of deep learning models into the field of machine translation has significantly enhanced the quality of translated texts, the translation of idiomatic expressions still poses a challenging task even for large language models. For this reason, researchers are actively seeking ways to address this problem through the selection of specific training datasets, the development of specialized test suites, and the creation of specialized text corpora.

One of the possible solutions to the problem of translating idioms is the creation of specialized text corpora containing a range of potentially idiomatic expressions in their respective contexts. Such corpora provide an opportunity to evaluate machine and neural translation systems with respect to their ability to perform a translation of specific phenomena. They also provide new, more focused material for training translation models, which typically results in improved translation accuracy. One of the most notable and widely recognized corpora in this area is the MAGPIE corpus, developed by Haagsma et al. This corpus contains over 50,000 instances of potentially idiomatic English expressions and was compiled and annotated using crowdsourcing. Another notable corpus of English-language idiom contexts is that created by Adewumi et al. The expressions in this corpus are divided into classes reflecting their idiomatic nature. These include metaphors, euphemisms, irony, paradoxes and other idiom types. In addition, there are similar parallel corpora, such as the English-German corpus of potentially idiomatic expressions compiled by Fadaee et al. Despite the existence of several such corpora, there has been an oversight in regard to the need for the creation of large parallel corpora containing idiomatic expressions in contexts, particularly with respect to Russian — English translation.

The aim of the present study is to create a parallel corpus of contexts of Russian and English potentially idiomatic expressions and to use it to fine-tune and test a neural translation model. In order to achieve this goal, it is necessary to identify and select relevant contexts of possibly idiomatic expressions in both the Russian and English languages. This task also requires conducting a statistical analysis of the compiled corpus and fine-tuning the neural network using the collected data. The final step involves evaluation of the translation accuracy of the fine-tuned model compared to its original state.

The process of creating a parallel corpus requires the use of a combination of methods, including continuous sampling and corpus compilation techniques. These techniques involve text preprocessing and markup, analysis of the corpus, and methods for providing access to it. In addition, deep learning techniques are used to verify the compiled corpus. In particular, fine-tuning and inferencing of neural network models are conducted.

# Literature Review

Translating idiomatic expressions often poses challenges for human translators and automatic systems alike. Among other problems, the most significant, according to researchers, are the tasks of identifying idioms and determining their meaning, which can be literal or figurative (Anastasiou, 2010). This is especially relevant for the field of machine translation, since automatic systems tend to translate such expressions literally, which is often incorrect (Anastasiou, 2010; Baziotis et al., 2023; Rodina & Lakiza, 2023). This also applies to modern neural translation systems, which continue to struggle with idiomatic expressions, despite a generally higher level of translation quality.

Approaches to the problem of translating idiomatic expressions in the field of machine translation have evolved along with the development of the field and the advent of new, more advanced translation tools. The key issue in the early studies was detecting idioms in the source text. According to Émile Delavenay (1962), it could be done by implementing a word-by-word analysis of the source text and a search for idiomatic combinations in it. The translation of these expressions involved the use of a supplementary dictionary of idiomatic expressions. Subsequently, this approach was implemented in a number of more complex solutions, including, for example, a method that calculates the probability that the selected sentence fragment is an idiom, an approach using so-called "local grammar rules", as well as various combinations of parsers with dictionaries or databases (Fischer & Keil, 1996; Pedrazzini, 1994; Stock, 1989). However, these approaches were largely ineffective, which, on the one hand, may be explained by their simplicity, and on the other by the limitations of the resources available at that time for the development of machine translation systems.

Despite the indicated low efficiency of the abovementioned solutions, developments in this direction continued, which was, to a greater extent, due to the limitations of classical translation methods based on rules and dictionaries. Nevertheless, the development of technology has allowed researchers to develop more advanced systems for detecting idiomatic expressions. Thus, systems that take into account morphological and syntactic information, as well as phrasal context, began to be developed. Although some developments showed promising results, these systems were still not effective enough, in particular, in distinguishing idioms from similar literal expressions (Anastasiou, 2010; Fazly et al., 2009; Takeuchi et al., 2007).

At the same time, approaches using machine learning methods in order to detect idiomatic expressions in the text have made headway. While the unsupervised learning method showed results comparable to the aforementioned approaches, supervised learning algorithms, by way of contrast, have proven to be more effective. Take the case of Feldman’s study, in which supervised machine learning algorithms outperformed unsupervised algorithms in the accuracy of detecting idioms by 20% with a final accuracy of about 80% (Feldman & Peng, 2013). However, this study was carried out on an imbalanced dataset, which could affect the results. Nonetheless, other studies also confirm the greater effectiveness of this approach (Feldman, 2021; Muzny & Zettlemoyer, 2013).

Despite the fact that the above approaches show potential, neural networks are widely used in the field of machine translation today. Large datasets in several languages, called parallel text corpora, are used to train neural networks. Despite the significantly better translation quality, neural networks trained in this way on the corpus of texts are not trained to solve such specific tasks as the detection of idioms and the resolution of semantic ambiguity (Baziotis et al., 2023; Feldman, 2021). Thus, neural networks still have some difficulties in translation.

Research on improving the quality of translation is currently underway in several directions. One of them is the use of hybrid systems, including the combination of various neural network architectures. Researchers use different ways of representing idiomatic expressions and different architectures of neural networks in a number of studies (Baziotis et al., 2023; Feldman, 2021; Senaldi et al., 2019; Serdyuk & Vlasova, 2021). Significant results were obtained by Kurfali and Östling, who managed to achieve 94% accuracy in determining idiomatic expressions in monolingual representations of texts (Kurfalı & Östling, 2020). However, the results were slightly worse for multilingual representations and were limited to German-English data. A notable contribution to this area has also recently been made by Christos et al., who developed a metric for evaluating the quality of translation of idioms, whose efficiency has yet to be proved (Baziotis et al., 2023). In brief, this is a burgeoning research area, and researchers present promising results.

Another direction in the development of this field is the compilation of special corpora containing contexts of potentially idiomatic expressions (PIEs). Such corpora can be used both to test the ability of translation systems to detect idioms and solve the problem of semantic ambiguity, and to train neural networks in order to improve the quality of translation of idioms. This area has emerged recently, and there are not many works related to the compilation of PIEs corpora. One of the most notable works in this area is the study of Haagsma et al. (Haagsma et al., 2020). They managed to compile “a high-quality corpus containing a total of 56 622 PIE instances.” However, there are several limitations in this work. Firstly, the created corpus is monolingual and contains texts only in English. Secondly, to create such a large set of texts, the authors resorted to using crowdsourcing, which required funding and caused a number of additional problems such as unreliable annotations that had to be addressed. Other existing corpora of PIEs are either also limited to one language, primarily English, or contain a parallel translation of contexts into one of the European languages, for instance, German (Adewumi et al., 2022). Thus, there is clearly a scarcity of research on the creation of a parallel corpus of PIEs for the Russian language.

To summarize all the above, it can be said that the field of machine translation has undergone significant transformations. Despite the noticeable increase in the overall quality of translation with the advent of neural networks in this area, the translation of idiomatic expressions is still problematic. One of the relevant and actively developing areas aiming at resolving this issue is the creation of corpora of potentially idiomatic expressions. Although there is a number of such corpora, some of them are monolingual. Meanwhile, when building multilingual corpora, the Russian language is overlooked. This work aims to fill in this gap in the literature.

# Methods

The texts included in the OPUS parallel corpora collection and the Russian National Corpus act as the basis for compiling the corpus. Using these reliable sources ensures the provision of high-quality text pairs without the need for translation. The search for idioms in texts will be conducted using the Oxford Dictionary of Current Idiomatic English and the Database of Russian Idioms.

The Russian-English Parallel Corpus, part of the National Corpus of Russian Language, contains 1 322 texts, which are divided into various domains, including the day-to-day life, academic, fiction, theological and other fields. A significant advantage of this corpus is the presence of automatic linguistic tagging, which is performed using Mystem and similar programs. This feature is useful for detecting idiomatic expressions in the texts (Sichinava, 2019).

The collection of parallel corpora in the OPUS system includes 41 corpora consisting of Russian and English texts. Among the corpora available in the system, the Book Corpus, which contains over 17 000 pairs of sentences written in Russian and English, as well as the Open Subtitles Corpus, with more than 25 million pairs of Russian-English sentences, may be of particular relevance in completing my task. All the corpora within the OPUS system utilize a standardized encoding and linguistic markup format, making it easier to work with these corpora (Tiedemann & Nygaard, 2004).

Generative neural networks with the encoder-decoder architecture will be used to validate the compiled corpus. As the corpus of idiomatic expressions is a collection of specific texts, pre-trained models are utilized, for which the performance metrics of translation before and after fine-tuning are compared. Architectures that utilize graph-based techniques are also of significant interest, as they enable the incorporation of semantic information, thus enhancing the accuracy and efficiency of translations (Yang et al., 2019).

Automatic Comet and chrF metrics will be used to assess the quality of the translated output. In addition to these metrics, a novel LitTER metric developed with the specific aim of evaluating an automatic translation system's ability to handle idiomatic expressions will also be employed (Baziotis et al., 2023). Despite BLEU being a more commonly used metric for evaluating the quality of machine translation, it has recently been discovered that this metric can sometimes produce inaccurate results (Freitag et al., 2020). Therefore, the use of these metrics would allow for more accurate and reliable results.

# Results

The most significant outcome of the research is the development of a parallel corpus, which could become a valuable resource for the development of models for automatic translation between English and Russian languages. This implies that the final corpus should contain more than 5,000 contexts of PIEs in both languages. The corpus will be made available on GitHub, which would ensure free access to it.

Conducting experiments will determine the potential utility of the corpus for future applications in neural translation model training. Fine-tuning of neural translation models through the use of corpus data is expected to increase the overall accuracy of idiom translation by 15%, based on the metrics employed. The code for the conducted experiments will be made available in the Jupiter Notebook, enabling the possibility of repeating them if required.

# Conclusion

To summarize, the expected outcome of this project is a high-quality parallel corpus of contexts of potentially idiomatic expressions in English and Russian. The corpus will be compiled taking into account the techniques used in similar studies, as well as considering the problems faced by other researchers when creating other corpora. When conducting experiments on neural networks, previous work in the field will also be considered. We anticipate that our results will be similar to those reported by Fadaee et al. as the dataset used in their study allowed to produce high-quality neural translations of idiomatic expressions with a good level of accuracy (Fadaee et al., 2018). Thus, the quality of the compiled case will be confirmed.

**Word Count:** 2139 words

**References**

Adewumi, T., Vadoodi, R., Tripathy, A., Nikolaido, K., Liwicki, F., & Liwicki, M. (2022). Potential Idiomatic Expression (PIE)-English: Corpus for Classes of Idioms. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 689–696). European Language Resources Association. https://aclanthology.org/2022.lrec-1.72

Anastasiou, D. (2010). *Idiom treatment experiments in machine translation*.

Baziotis, C., Mathur, P., & Hasler, E. (2023). Automatic Evaluation and Analysis of Idioms in Neural Machine Translation. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3682–3700. https://doi.org/10.18653/v1/2023.eacl-main.267

Fadaee, M., Bisazza, A., & Monz, C. (2018). *Examining the Tip of the Iceberg: A Data Set for Idiom Translation* (arXiv:1802.04681). arXiv. http://arxiv.org/abs/1802.04681

Fazly, A., Cook, P., & Stevenson, S. (2009). Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, *35*, 61–103. https://doi.org/10.1162/coli.08-010-R1-07-048

Feldman, A. (2021). Hard Nut to Crack: Automatic Idiom Detection. *Interférences Littéraires/Literaire Interferenties*, *25*, 188–203.

Feldman, A., & Peng, J. (2013). *Automatic Detection of Idiomatic Clauses*. *7816*, 435–446. https://doi.org/10.1007/978-3-642-37247-6\_35

Fischer, I., & Keil, M. (1996). Parsing decomposable idioms. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. COLING 1996. https://aclanthology.org/C96-1066

Freitag, M., Grangier, D., & Caswell, I. (2020). BLEU might be Guilty but References are not Innocent. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 61–71. https://doi.org/10.18653/v1/2020.emnlp-main.5

Haagsma, H., Bos, J., & Nissim, M. (2020). MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 279–287). European Language Resources Association. https://aclanthology.org/2020.lrec-1.35

Halliday, M. A. K., & Delavenay, E. (1962). An Introduction to Machine Translation. *The Modern Language Review*, *57*(1), 73. https://doi.org/10.2307/3721978

Kurfalı, M., & Östling, R. (2020). Disambiguation of Potentially Idiomatic Expressions with Contextual Embeddings. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova, & A. Savary (Eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (pp. 85–94). Association for Computational Linguistics. https://aclanthology.org/2020.mwe-1.11

Muzny, G., & Zettlemoyer, L. (2013). Automatic Idiom Identification in Wiktionary. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1417–1421). Association for Computational Linguistics. https://aclanthology.org/D13-1145

Pedrazzini, S. (1994). *Phrase Manager: A System for Phrasal and Idiomatic Dictionaries* (Vol. 3). Olms Verlag, Hildesheim. https://repository.supsi.ch/1692/

Rodina, S., & Lakiza, E. (2023). LINGUISTIC PROBLEMS OF MACHINE TRANSLATION. *Universum:Philology and art history*, *108*(6). https://doi.org/10.32743/UniPhil.2023.108.6.15686

Senaldi, M. S. G., Bizzoni, Y., & Lenci, A. (2019). *What Do Neural Networks Actually Learn, When They Learn to Identify Idioms?*

Serdyuk, Y., & Vlasova, N. (2021). Idiomatic expression usage recognition by neural networks. *Program Systems: Theory and Applications*, *12*(3), 3–26. https://doi.org/10.25209/2079-3316-2021-12-3-3-26

Sichinava, D. (2019). Параллельные тексты в составе Национального корпуса русского языка: Новые языки и новые задачи. *Proceedings of the V.V. Vinogradov Russian Language Institute*, *21*, 41–61. https://doi.org/10.31912/pvrli-2019.21.2

Stock, O. (1989). Parsing with Flexibility, Dynamic Strategies, and Idioms in Mind. *Computational Linguistics*, *15*(1), 1–18.

Takeuchi, K., Kanehila, T., Hilao, K., Abekawa, T., & Kageura, K. (2007). Flexible automatic look-up of English idiom entries in dictionaries. In B. Maegaard (Ed.), *Proceedings of Machine Translation Summit XI: Papers*. https://aclanthology.org/2007.mtsummit-papers.60

Tiedemann, J., & Nygaard, L. (2004). *The OPUS corpus—Parallel and free http://logos.uio.no/opus/*.

Yang, H., Xie, G., Qin, Y., & Peng, S. (2019). Domain Specific NMT based on Knowledge Graph Embedding and Attention. *2019 21st International Conference on Advanced Communication Technology (ICACT)*, 516–521. https://doi.org/10.23919/ICACT.2019.8701980