Věda a perspektivy № 11(42) 2024
ISSN 2695-1584 (Print)
ISSN 2695-1592 (Online)
https://doi.org/10.52058/2695-1592-2024-11(42)-422-431

Maryana Tomenchuk

PhD in Philology, Associate Professor of the
Department of Applied Linguistics,
Uzhhorod National University, Uzhhorod, Ukraine
https://orcid.org/0000-0002-2036-4616

Kseniia Popovych

Master Graduate Student of Applied Linguistics,
Uzhhorod National University,
Uzhhorod, Ukraine

LARGE LANGUAGE MODELS AND MACHINE TRANSLATION

Abstract. Our article deals with linguistic peculiarities of machine translation provided by large language models, focusing on its lexical, semantic and syntactic aspects. Due to the fact that large language models are a rapidly developing notion of 21st century, the ability to understand their architecture and relation to linguistics and translation in particular is of a great importance.

The study is based on the investigation of ChatGPT – a very efficient large language model, which is capable of producing human-like outputs when provided with a certain task. Through the analysis of ChatGPT-responses, we further investigate its ability to convey linguistic meaning in different aspects of a language.

The main purpose of our investigation is to analyze linguistic notions from different subfields, such as syntax, lexis and semantics by reviewing the translations of GPT from target to source language. In our case, these are represented by English and Ukrainian languages.

The corpus of texts which were analyzed in our research has been taken from English scientific papers published in online scientific journals [11], [13], [15] from various academic fields, such as artificial intelligence, computer science, mathematics, chemistry, biology etc. The motivation to choose scientific discourse lies within the fact that academic language is rich in terminology, formality and special structures that allow to investigate the performance of GPT in different areas. After the analysis of a model was conducted, the outputs of ChatGPT were compared to the human translation and further investigated.

Our investigation has shown that such a linguistic model is a powerful tool than can be used in the area of machine translation. Its structure and architecture contribute a lot to the identification of various patterns in text, which imitates the

way human percept and analyze information given. However, although being very precise and natural in the formation of responses, ChatGPT proved to be not yet a perfect machine that would substitute a professional's work at the whole. The results of our investigation show that it tends to make mistakes when it comes to different complex structures of a language, in our case – English and Ukrainian.

In our work, we used descriptive analysis of the content of large language models and linguistics to provide the relevant data on the investigated topic and to highlight the most common features of academic language translation and exploratory analysis to find most common issues occurring when translating such a corpus from English into the Ukrainian language. The inferential analysis was used in order to reach some conclusions in investigated data, including the highlight of linguistic aspect of machine translation. We used the comparison method to discover the differences of lexis and grammar in the English and Ukrainian languages.

Keywords: machine translation, large language model, artificial intelligence, lexis, syntax, semantics.

The problem statement: Due to the growing interest to artificial intelligence in the field of linguistics and machine translation, it is essential to investigate the performance of such models for this purpose. The wide range of linguistic peculiarities cause various issues for a machine, which may result into improper translation and other problems. Our aim is to investigate the most common issues that occur in such an activity and provide possible solutions to them.

Review of recent publication: As the area of artificial intelligence is continuously evolving, the amount of scientific literature for these purposes is rapidly expanding. This article draws upon the works of prominent scholars such as Begus [2], Curry [4], and Gubelmann [6].

The **aim** of the study is analyze the most common lexico-semantic and syntactic features that occur in the process of translation from English to Ukrainian by a large language model, in our case – ChatGPT and propose the potential solutions to problems that persist in such a phenomenon.

Results and Discussions: Large language models play a crucial role in the area of modern linguistics, serving as a helping tool in translation of corpora from various fields of language.

Modern approaches in linguistics enable the application of various computer technologies in order to make the process of translation more clear, easy and less time-consuming. Machine translation, a subfield of traditional translation practices, incorporates different techniques from other branches like artificial intelligence, machine learning and deep neural networks in particular.

In the field of natural language processing, a large language model is a computational model which is designed to perform a wide range of NLP tasks, including general-purpose language generation and classification. These models

achieve their capabilities by learning statistical relationships from extensive text corpora through a computationally intensive training process that involves both self-supervised and semi-supervised techniques [2, p.63]. Large language models are particularly adept at text generation, a special type of generative AI, where they process input text and predict subsequent tokens or words iteratively.

Large language models have become very successful recently in many areas. Most of these models use a transformer-based design, which is built to handle sequential data like language. ChatGPT, the focus of our study, is based on this transformer model. OpenAI's GPT [3] (which is generative pre-trained transformer) is especially important in natural language processing because it works in an autoregressive way. This means that it predicts the next word in a sequence based on previous words, using large amounts of text data to understand language patterns.

From a language perspective, GPT has several features that make it good for tasks like machine translation. One key feature is its autoregressive modeling, where it generates text one word at a time by considering the words that came before. This makes it strong at creating coherent and meaningful responses, useful in conversations and creative writing. Another important feature is its two-step training process: pre-training and fine-tuning. During pre-training, GPT learns general language rules and patterns from a large amount of text. After that, it is fine-tuned for specific tasks like translation, summarization, or answering questions, making it more specialized and effective.

Machine translation techniques have traditionally been categorized based on their translation strategies, with three main systems being direct systems, transfer systems, and interlingua systems [5, p. 120]. Direct systems focus on mapping phrases or structures from the source language directly to the target language through detailed pattern matching, while making adjustments for elements like word order. This approach was common in early translation systems and remains in use in some modern translation tools for personal computers. Large language models like GPT enhance this process by using their extensive pre-trained data to predict and generate translations more flexibly and accurately. By adjusting word order and other syntactic elements in real-time, these models surpass traditional direct systems that depend solely on fixed rules and databases.

Transfer systems, on the other hand, use a more abstract translation process. They convert the source language input into a generalized structure, removing specific grammatical features of the source language before transforming this abstract structure into a corresponding one in the target language. The level of abstraction can vary, with greater abstraction making it easier to create a compatible transfer module. LLMs excel at this form of translation due to their ability to create and manipulate abstract representations of language [10, p. 13]. By analyzing and generalizing input into intermediate forms, these models can more accurately transform text into the target language structure, preserving meaning and context

more effectively than traditional transfer systems. This capability allows for nuanced and contextually aware translations that capture the original text's structure and intent.

Another linguistic notion that is widely used in the structure of LLM is the notion of corpus [4, p. 15]. These corpora serve as the foundation for building language models used in AI systems, including chatbots, in our case – ChatGPT. Corpus linguistics and AI-driven chatbots intersect at the level of language data processing and interaction design. This convergence involves using large linguistic corpora to train and refine chatbots' language models, enabling them to understand, generate, and respond to human language more naturally and effectively [9, p. 120].

Lexical aspects of machine translation represented by ChatGPT comprise various notions, such as idiomatic expression, neologisms, register and formality, internationalisms and pseudo internationalisms.

One of the major concerns we faced during examination of these features is the accuracy of the definitions and examples generated by ChatGPT. While the model can produce coherent text, it sometimes generates content that is factually incorrect or lacks the nuance required for precise dictionary entries. This phenomenon, often referred to as hallucination, where the model invents information, poses a significant challenge for lexicographers [12, p. 12].

Beside the mention fact, ChatGPT appeared to be not yet consistent in translating from English to Ukrainian language. English is an analytical language that relies heavily on word order and auxiliary words to convey meaning, while Ukrainian is a synthetic language, characterized by its extensive use of inflections and a flexible word order to express grammatical relationships. This fundamental difference poses challenges for language models, which may struggle to capture the full complexity of Ukrainian morphology, leading to inaccuracies or awkward translations [3, p. 7].

ChatGPT can translate text between many languages with a good level of accuracy. For example, when translating idioms that are frequently used in both languages, there is a very low probability of him to make some mistakes [14, p. 340]. This simple sentence demonstrates the model's ability to handle basic multilingual translation tasks effectively:

English idiom	ChatGPT's translation
*"**Bear in mind** the significance of cultural context when interpreting historical texts, as it can profoundly influence the intended meaning and reception of specific terms and expressions."* [15, p. 130]	*"Майте на увазі важливість культурного контексту під час інтерпретації історичних текстів, оскільки він може суттєво впливати на задумане значення та сприйняття окремих термінів і виразів."*

Table 1. Demonstrating ChatGPT's ability to convey lexico-semantic meaning

However, the quality of translation can be inconsistent when it comes to complex or idiomatic expressions. The example below demonsrates that translating the English idiom literally into Ukrainian makes no sense, while the correct translation, provided by a human translator, "*заповнити прогалину*", accurately conveys the meaning of the given linguistic notion. [14, p. 200].

English idiom	ChatGPT's translation
"Then, in April of that year, George P. Smoot and his colleagues at Berkeley released evidence that might ***fill this gap*** *in the theory."* [13, p. 56]	*"Потім, у квітні того року, Джордж П. Смут і його колеги з Берклі оприлюднили докази, які могли б* ***заповнити цю яму*** *в теорії."*

Table 2. ChatGPT's inability to translate more complex idioms

This example shows that ChatGPT can struggle with idioms and phrases that require a deeper understanding of cultural context and language-specific nuances beyond simple word-for-word translation.

In addition to the mentioned issues, LLMs like ChatGPT are trained on large datasets sourced from the internet, which may contain biases. Such a training process results in the fact that the model's output can reflect these biases, leading to dictionary entries that may inadvertently perpetuate stereotypes or underrepresent certain groups. This is particularly problematic in a lexicographic context, where the goal is to provide balanced and accurate language resources.

For example, when generating dictionary entries for terms associated with gender, race, or socioeconomic status, ChatGPT might reflect societal stereotypes. This is problematic because dictionary entries are expected to be neutral and objective. The model might also underrepresent minority dialects or non-standard forms of English, as it is predominantly trained on mainstream, standardized language data [6, p. 78]. This limitation can lead to dictionary entries that fail to fully capture the diversity and richness of language as it is used across different communities.

Although machine translation has become a very powerful tool in the area of Applied Linguistics, syntactic aspects of every language remain a challenging task for automated translation systems. Syntax, which refers to the arrangement of words and phrases in a sentence, is a very ambiguous notion for different language groups. In this subchapter, we will examine basic rules applied to generating well-structured corpora and highlight the common issues which occur in the field of machine translation by large language models, in our case – ChatGPT.

As to any other lexico-semantic features analyzed in our work, these comprise context preservations in specialized terminology, cultural and contextual adaptation, polysemy, metaphors and jargon. Traditional machine translation systems, like rule-based or statistical ones, used specific programs to ensure correct grammatical structure during translation. While this helped maintain accuracy, it required a lot of manual effort and language-specific resources [5, p. 25]. Neural machine translation systems work differently—they learn from large amounts of text data and remember how words relate to each other, without explicitly learning syntax rules.

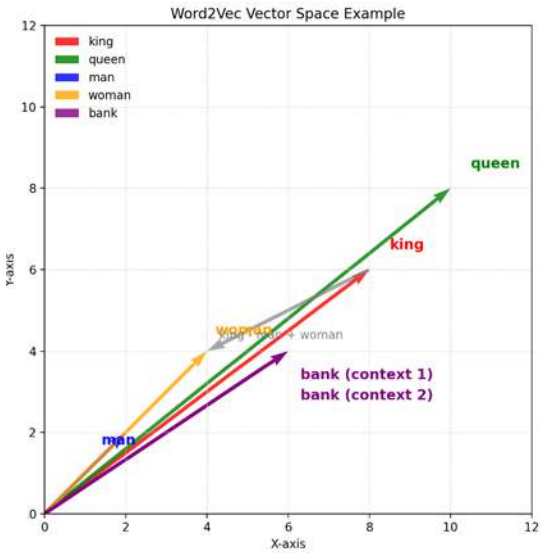
This approach allows neural models to handle complex language patterns, but they sometimes make mistakes if they haven't seen enough examples to understand all the syntactic nuances [7, p. 98]. While these models are good at creating fluent text, they can struggle with complex syntax, especially when translating between languages with very different grammatical rules. Improving their ability to better understand syntax can help make translations more accurate and natural.

In this case, vector-based semantics appears to be an essential tool. These models represent linguistic elements – such as words, phrases, and sentences – as continuous vectors in a high-dimensional space. The distances and relationships between these vectors capture semantic similarities and distinctions, allowing the models to comprehend context and meaning beyond the literal forms of words.

One of the key applications of such vector spaces is word embeddings, which convey the semantic meaning of words or phrases during translation from the source language to the target language.

Word embeddings are essentially mathematical representations of words as dense vectors. They have played a pivotal role in modern natural language processing and translation systems. By encoding words as vectors, embeddings help capture both syntactic and semantic relationships, facilitating the model's ability to understand and generate contextually accurate translations. This approach enables nuanced and context-aware interpretations, enhancing the overall quality and fluidity of translated content.

For instance, in the vector space generated by Word2Vec, words with similar meanings, like *"king"* and *"queen,"* are placed close to each other. Relationships between words, such as *"king - man + woman = queen,"* can be represented as linear transformations within this space [11, p. 3-5]. However, these embeddings lack context sensitivity, meaning a word like "bank" would have the same representation whether it refers to a riverbank or a financial institution, which limits the model's ability to distinguish between different meanings of the same word based on context.



Pic. 3 Word2Vec example usage

Large language models, like ChatGPT with its transformer-based architecture, are designed to handle translation between multiple languages. However, this can sometimes lead to challenges due to the limited availability of data for certain languages. To address these issues, such models use innovative techniques like up-sampling and back-translation to enhance translation quality. These methods help the model better learn from limited data, allowing it to infer meaning and produce semantically coherent translations even in less-resourced languages. Consequently, these techniques contribute to more accurate and contextually meaningful translations across a diverse range of languages.

Traditional machine translation methods often relied on explicit rules and syntax models, such as using syntactic parsers to ensure proper grammar. However, modern large language models like ChatGPT learn syntax indirectly by training on huge amounts of text data [8, p. 7-9]. This allows them to produce natural translations, but sometimes they can make mistakes with word order, misplaced words, or grammar when translating complex sentences.

Challenges arise when translating between languages with different structures. For example, languages like Japanese and Korean, which use a subject-object-verb (structure, are harder for models trained mostly on English's subject-verb-object structure. Morphologically complex languages, like Finnish or Turkish, also pose difficulties because words change based on their role in a sentence. Ambiguous syntax, where sentences can have more than one meaning, adds another layer of complexity [8, p. 90].

Improving LLMs for translation often involves multilingual fine-tuning. By training on a variety of texts with complex sentence structures in different languages, models become better at handling syntax. Cross-lingual learning is another strategy where knowledge from high-resource language pairs, like English-French, helps improve translations for less-resourced languages like Ukrainian.

Hybrid approaches that combine LLMs with explicit syntactic models can further enhance accuracy. By using syntactic rules alongside machine learning, these models ensure better grammar and word order. Post-translation corrections also help by reviewing and refining translations to match grammatical standards. In the future, better training data and improved correction systems will make machine translations even more accurate and human-like.

One of the possible solutions to the mentioned problems in lexis, semantic and syntax is represented by the notion of self-attention mechanisms. It allows models to focus on different parts of an input sentence to understand the relationships between words, regardless of their position. This capability is especially useful for capturing syntactic and contextual nuances across languages.

In translation tasks, self-attention helps models weigh the importance of each word relative to others in the sentence, ensuring that key dependencies, such as word order and grammatical structures, are preserved. When combined with self-

correction mechanisms, self-attention can further refine translations by identifying areas where context, tense, or meaning may have been misinterpreted [7, p. 35]. This focused re-evaluation helps the model better handle complex syntactic dependencies, idiomatic expressions, and variations in word meaning, ultimately leading to more accurate and contextually appropriate translations.

Another effective approach involves using self-correction mechanisms to improve translation quality. In this method, the LLM first generates an initial translation, which is then refined through multiple iterations based on syntactic feedback. The process can include explicit error detection using metrics like Translation Edit Rate to spot and correct syntactic issues in the output. Essentially, this self-refinement framework allows the model to make adjustments based on any syntactic discrepancies identified during the initial pass [12, p. 45]. This approach is especially useful for languages with complex syntactic rules, as it helps correct issues with word order or tense usage. By refining translations of sentences with complex structures, such as nested clauses or subordinate phrases, self-correction mechanisms can significantly enhance overall translation quality.

Moreover, cross-lingual few-shot learning is another valuable approach for improving syntactic translation in large language models. It leverages syntactic examples from one language pair to enhance translation performance for another pair, focusing on shared linguistic similarities. For instance, an LLM might use examples from a high-resource language pair, such as English-French, to improve translations for a lower-resource pair like English-Ukrainian. This method helps align syntactic structures in languages where data availability is limited.

What makes cross-lingual few-shot learning so effective is its ability to help LLMs model syntactic structures in underrepresented languages by drawing on similarities in grammatical rules across languages. By doing so, it reduces reliance on large amounts of parallel data, as the model generalizes syntactic knowledge from widely-studied languages to less common ones. This allows for improved accuracy and contextual understanding even in low-resource translation scenarios.

All these features of English and Ukrainian languages contributed to the diversity of issues a model is facing when trying to convey the meaning of a sentence, paragraph or even the entire article from source into the target language.

Conclusions: Large language models like ChatGPT have proven highly valuable for modern linguistics and machine translation due to their extensive training on diverse data. In our study, we evaluated ChatGPT 4.0's ability to translate scientific texts between English and Ukrainian. We focused on its handling of lexico-syntactic, and semantic features. While ChatGPT is effective as a support tool for human translators, providing functions like error-checking and context searches, it still has limitations with complex terms and context-specific nuances.

One significant challenge is the limited availability of Ukrainian-language corpora, which affects translation accuracy. Ukrainian is less represented in

international scientific literature compared to English, which often serves as the primary language for academic work. We analyzed 70 scientific passages across fields such as chemistry, artificial intelligence, and biology to assess the model's performance. We found that it struggles with precise terminology and domain-specific language, highlighting the need for richer Ukrainian resources.

Lexical challenges were common, with ChatGPT often misinterpreting idiomatic expressions, new words, and technical terms. This led to inconsistent or inaccurate translations. The model's performance could be improved by training it on more varied data, including regional, colloquial, and specialized language used in Ukrainian. This would help it better understand and translate terms with more contextual accuracy.

On the semantic level, the differences between English and Ukrainian pose unique challenges. English often relies on word choice and sentence structure for nuanced meanings, while Ukrainian, with its inflectional system, conveys meaning through changes in word endings and forms. ChatGPT sometimes struggled to capture these nuances, particularly when translating complex terms, metaphors, and context-dependent language. Expanding the model's training on scientific texts and refining its handling of implicit meanings could enhance its semantic translation capabilities.

Syntactically, ChatGPT had difficulty managing differences in word order, subject-predicate agreement, and sentence flexibility between English and Ukrainian. English uses a fixed word order, while Ukrainian's case system allows for varied word positions to emphasize meaning. To improve translations, the model needs to better adapt word order and respect complex grammatical agreements in Ukrainian.

In our work, we also provide potential solutions to the mentioned and analyzed problems. For example, increasing the volume and diversity of Ukrainian-language texts used for training, especially those taken from the field of science, would provide a more robust foundation for translation. This can be achieved by the availability of larger datasets that cover scientific texts, technical documents, colloquial expressions, and regional variations. Collaborative efforts among researchers, linguistic experts, and institutions can accelerate the creation and availability of such resources, leading to better contextual understanding and nuanced translations.

Moreover, an effective strategy for improving translation accuracy is the continuous refinement of LLMs through feedback and iterative learning cycles. Collaborations with human translators, linguists, and domain experts can help identify translation errors, suggest corrections, and feed this feedback into the model's training pipeline. The iterative process allows for dynamic adaptation, ensuring that the model evolves in response to evolving language patterns and user needs.

References:

1. A Deep Decomposable Model for Disentangling Syntax and Semantics in Sentence Representation / D. Li et al. *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic. Stroudsburg, PA, USA, 2021. URL: https://doi.org/10.18653/v1/2021.findings-emnlp.364

2. Begus G., Dabkowski M., Rhodes R. Large linguistic models: analyzing theoretical linguistic ablities of LLMs. *Rutgers University Manuscript*. 2023. August 21. P. 1–28. URL: https://arxiv.org/abs/2305.00948

3. ChatGPT by OpenAI 4.0. URL: https://chatgpt.com

4. Curry N., Baker P., Brookes G. Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*. 2023. P. 100082. URL: https://doi.org/10.1016/j.acorp.2023.100082

5. Estimating and Comparing Translation Skills: A Comparative Study of ChatGPT and Human Translation. *Journal of Development and Social Sciences*. 2024. Vol. 5, no. III. URL: https://doi.org/10.47205/jdss.2024(5-iii)08

6. Gubelmann R. A Loosely Wittgensteinian Conception of the Linguistic Understanding of Large Language Models like BERT, GPT-3, and ChatGPT. *Brill*. URL: https://brill.com/view/journals/gps/99/4/article-p485_2.xml

7. Jiang L. P. Unifying Syntax and Semantics in Cognitive Concept Generation for Natural Language Expression. *International Journal of Humanities, Social Sciences and Education*. 2023. Vol. 10, no. 10. P. 10–16. URL: https://doi.org/10.20431/2349-0381.1010002

8. Karaban, V. (2004). Pereklad anhliiskoi naukovoi i tekhnichnoi literatury *[Translation of English scientific and technical literature]*. Vinnytsia: Nova Knyha [in Ukrainian].

9. Kiselev D. An AI using Construction Grammar: Automatic Acquisition of Knowledge about Words. *12th International Conference on Agents and Artificial Intelligence*, Valletta, Malta, 22–24 February 2020. 2020. URL: https://doi.org/10.5220/0008865902890296

10. Koutsoudas A., Humecky A. Linguistics and Machine Translation. *WORD*. 1959. Vol. 15, no. 3. P. 489–491. URL: https://doi.org/10.1080/00437956.1959.11659712

11. New England Journal of Medicine. URL: https://www.nejm.org/

12. Patil A. Top 5 Pre-trained Word Embeddings. *Medium*. URL: https://patil-aakanksha.medium.com/top-5-pre-trained-word-embeddings-20de114bc26

13. Taylor J. R. Semantic structure in Cognitive Grammar. *Cognitive Grammar*. 2002. P. 96–120. URL: https://doi.org/10.1093/oso/9780198700333.003.0006

14. The Lancet. URL: https://www.thelancet.com/.

15. Trujillo A. Computational Linguistics Techniques. *Translation Engines: Techniques for Machine Translation*. London, 1999. P. 85–119. URL: https://doi.org/10.1007/978-1-4471-0587-9_5

16. Wiley Online Library. URL: https://onlinelibrary.wiley.com/.