Title: ChatGPT "Anglicizing" Kazakh across different discourse domains: specifically in terms of calques and directly borrowed terms.

Contributor: Dilnaz Tursynbekova

Introduction

Artificial Intelligence is expanding and developing day by day. One of the most popular and successful examples is ChatGPT. It has millions of users from all over the world. ChatGPT's popularity since its introduction in 2022 has been increasing rapidly, as it acquired over 1 million users in just 5 days (Lee, 2023). However, such LLMs mainly get trained in English and do not work well in low-resource languages (LRLs) (Kshetri, 2024). Kazakh is an example of such LRLs (Asmaganbetova et al., 2024). So, there is a clear inequality between the performance of ChatGPT in languages with a high number of digital resources and those with data scarcity. From my personal user experience, when having conversations with ChatGPT in Kazakh, it frequently happens that it generates grammatically correct, but unnatural-sounding sentences. Doing a bit of analysis, it was noticeable that the structures were Anglicized, causing that unnaturalness. This research aims to address this issue by answering the research question "How does ChatGPT 'Anglicize' the Kazakh language across different discourse domains, specifically in terms of calques and directly borrowed terms?" The main hypothesis is "ChatGPT's Kazakh outputs will have an uneven distribution of anglicisms across different discourse domains".

The scope of this study was limited to evaluating ChatGPT's performance in Kazakh with occasional reference to Kyrgyz, due to the limited literature focused only on the Kazakh language. Also, due to time and data constraints, the focus was on ChatGPT powered by GPT-5.

The analysis was based on text generation tasks, with prompts in Kazakh as input data, and outputs were evaluated for different signs of “Anglicisation”, namely several types of calques, or overuse of borrowed terms. In addition, speech and multimodal outputs were also not covered in this research due to time and data constraints.

Literature Review

ChatGPT and any other AI-driven tools have many advantages. However, they are not perfect with various limitations and caveats. One of the major problems is that almost all of the modern AI tools are being trained in English. Kshetri (2024) points to the over-reliance of large language models’ (LLMs) training process on the English language, and that this dominance escalates the issues with Generative AI that people and companies from developing countries, mainly non-English, face (p. 95). Various advanced machine learning algorithms used for Natural Language Processing are primarily suitable for English, although they are not widely utilized in developing nations (Kshetri, 2024, p. 95). Modern GAI tools, which are rapidly improving, are possibly leaving out approximately six billion first-language speakers of over 7000 languages from different job opportunities, education, art, and other fields (Kshetri, 2024, p. 96). Such disadvantaged languages are called low-resource languages (LRLs). According to Du et al. (2023), there are 7151 languages in the world that are known to humanity and are still alive (p. 1). However, the major part of these 7000+ languages consists of low-resource ones (Du et al., 2023, p. 1). LRLs correspond to several or all of the points such as the absence of a distinctive writing system or a consistent orthographic structure, a minimal presence on the Internet, limited knowledge in linguistics, and a limited range of digital resources for natural language processing including language corpora, various dictionaries, audio transcriptions and many more (Du et al., 2023, p. 1). Du et al. (2023), in their study, describe the Kazakh language as one of the examples of low-resource languages along with Uyghur and Kyrgyz (p. 1). Generative Artificial Intelligence instruments like ChatGPT, which are trained on high-resource languages, often produce superficial and sometimes even incomprehensible results in LRLs (Kshetri, 2024, p. 97). To illustrate, ChatGPT generates responses with errors in even simple structures when asked to generate output in the Kazakh language. Asmaganbetova et al. (2024) exemplify it through prompting ChatGPT by asking it to define what Kazakh grammar is, and it resulted in its poor skills in Kazakh with various errors (p. 3).

A recent study on evaluating ChatGPT's performance in a multilingual context also confirms the chatbot's bias towards English. Lai et al. (2023) tested ChatGPT on 37 different languages with high to extremely low resources on 7 NLP tasks (in particular, "Natural Language Inference (NLI), Question Answering, and Common Sense Reasoning, [...] Part-of-Speech (POS) Tagging, Named Entity Recognition (NER), Relation Extraction, and Summarization") (p. 13172). One of the extremely low-resource languages was Kyrgyz (Lai et al., 2023, p. 13174), which is linguistically very close to Kazakh, and there is a high possibility of encountering the same issues in both of them. According to Lai et al. (2023), ChatGPT performed better in English than in low-resource languages in the above-mentioned 7 NLP tasks, and it showed better performance with prompts in English even with tasks and input texts in a different target language, confirming ChatGPT's bias towards the English language (p. 13173). In another study, Adilmetova et al. (2025) evaluated the ability of ChatGPT to produce personalized, consistent, and practical dietary advice in English, Kazakh, and Russian on 50 mock patient cases (p. 729). Even though the prompts were the same, the model resulted in "largely inapplicable for evaluation" advice in Kazakh (Adilmetova et al., 2025, p. 733). The scores in all three categories were around 1 out of 5 for Kazakh, while English and Russian scored around 3 out of 5 (Adilmetova et al., 2025, p. 732). So, it again shows a language bias in ChatGPT rooted in the scarcity of training data.

Methodology

For data acquisition, I had conversations with the chatbot in Kazakh in 4 different chats with different domains: everyday, academic, cultural, and administrative topics. It was decided to create 4 prompts to begin the conversations with ChatGPT-5 for each domain, and then follow the conversations as they unfold until 5,000 words per domain were collected to have the same amount of raw data. This way, I could control the word count of 5,000 words, and had more natural conversations, while with the initial method of creating 25 prompts per domain in advance and sending them to ChatGPT, I would have ended with different word counts for each domain as ChatGPT's response lengths differ for each prompt, and I would have had just 25 different not really related conversations with each other, as a result of creating prompts in advance. The first prompts for academic, administrative, cultural, and daily were as follows, respectively:

1) Ғылыми зерттеу жоспарын қалай құруға болады? - How to create a research plan?

2) Қазақстан Конституциясының бөлімдері жазып бере аласың ба? Жақсы мәлімет керек - Can you write down the part sof the Constitution of Kazakhstan? I need good information

3) Қазақ мәдениетінің ерекшеліктері неде? - What are the features of the Kazakh culture?

4) Уақытты тиімді жоспарлау үшін не істеуге болады? - What can be done to plan time efficiently?

The collected data was copied and pasted into four separate Google Docs, removing the prompts and leaving only ChatGPT's responses, because the purpose of the research is not about analyzing prompts and how ChatGPT's responses differ depending on their quality. So, it was not necessary to keep my prompts there.

The identification of direct English loanwords was straightforward. However, a calque is generally a broad term, so it was important to define which types of calques I focus on in this research, other than borrowed English loan terms. In this study, I used Betz's (1949) typology of calques and 'syntactic' calques (González & Knospe, 2019):

1) "loan translations, which are based on a literal translation of every component of a complex",

2) "loan renderings (also called 'loan renditions'), which deviate from the meaning and/or morphology of the English [or Russian] source by translating it more freely",

3) "loan creations [...] are free equivalents formally independent from the patterns of the source language",

4) “semantic calques”: “in this case, speakers impose an additional (metaphorical) meaning on an existing sign in the receiving language, orienting themselves towards an English [or Russian] model sign”,

5) “syntactic calques”: “inclusion of more abstract (grammatical) transfer features” (Gonzàles and Knospe, 2019, p. 238)

Identifying English loanwords was easy, as it was very clear when a word/phrase was a loanword or not. But the calques were more challenging. The first phase of spotting calques was through intuition, when the phrases did not sound very natural in Kazakh. However, it was not sufficient to make reliable claims. The evidence that I used for identifying whether a word/phrase is a calque was checking whether there was a native alternative. As a result of tight language contact with Russian, Kazakh has many calques already absorbed into it and used normally not only in spoken Kazakh but also in academic Kazakh. The strategy that helped me identify how “normal” a calque is in Kazakh was checking the frequency of use through the corpora of the Kazakh language. I searched them on the National Corpus of the Kazakh language, which is free and public (Aitova, 2025). Then, searching for similar constructions and how frequent they are in the British national corpus (Davies, 2004) and the Russian national corpus (Savchuk et al., 2024) was used as supporting evidence to prove that something is a calque and to identify the origins of a calque. Depending on the frequency of use of British or Russian alternatives of identified calques, it was possible to make initial decisions regarding the source language. Another method was searching for calques on Wikipedia in Kazakh, because it is mostly translated from English, so I used it to have an additional confirmation whether the calques on ChatGPT were translated from English. However, there was also a qualitative analysis. There is a possible underrepresentation of newer terms in the English corpus, which made the quantitative criteria alone insufficient. In some cases, highlighted with yellow in the Excel sheet, the calques' source language was identified as English despite a low or zero frequency in the English corpus because there was a clear contextual link to English as a source language (e.g., Аралықпен қайталау (Spaced repetition) and Mind map (ақыл картасы)). In addition, such factors as structural similarity, in what context it was used, and phrase patterns were also considered when making final judgments about the source of calques. During the manual annotation of the collected data, I used color-coding (green for loan translations, blue for loan renderings, pink for loan creations, yellow for semantic calques, and orange for syntactic calques). Elaborating on the choice of written sources over spoken, the conversations with ChatGPT were in a written format, not speech. Therefore, the preference was given to written corpora and Kazakh Wikipedia.

After that, every identified loanword and calque was listed in an Excel sheet as a data point in rows. For calques, there was also information about what type of calque each of them was, absence in Qazcorpora and Kazakh Wikipedia, natural construction in Kazakh, and the source constructions and their frequency (English and Russian). The frequencies across different domains were counted and listed in frequency tables for each feature. Afterwards, chi-square tests of independence were done in Python using the SciPy library (see Appendix A for code) to evaluate whether the anglicism types (loanword or calque) are related to a discourse domain. The null hypothesis was that there is no correlation between a borrowing type and a domain. The alternative hypothesis stated that there is an association between a borrowing type and a domain.

Results

Domains	English loanwords	English calques	Russian calques	Ambigious	Total Anglicisms (n, %)
Everyday	19	14	16	0	33 (10.1)
Culture	17	7	5	0	24 (7.4)
Academic	211	22	5	1	233 (71.5)
Administrative	21	15	2	1	36 (11.0)
Total	268	58	28	2	326

Table 1. Full frequency data for transparency.

Domains	Loanwords (n, %)	Calques (n, %)	Total Anglicisms
Everyday	19 (57.6)	14 (42.4)	33
Culture	17 (70.8)	7 (29.2)	24
Academic	211 (90.6)	22 (9.4)	233

Administrative	21 (58.3)	15 (41.7)	36
Total	268 (82.2)	58 (17.8)	326

Table 2. Number of anglicisms for Chi-square testing.

Domains	**Loan translations (n, %)**	**Loan renderings (n, %)**	**Loan creations (n, %)**	**Semantic calques (n, %)**	**Syntactic calques (n, %)**	**Total**
Everyday	4 (28.6)	0 (0.0)	0 (0.0)	7 (50.0)	3 (21.4)	14
Culture	4 (57.1)	1 (14.3)	0 (0)	1 (14.3)	1 (14.3)	7
Academic	18 (81.8)	0 (0.0)	0 (0.0)	0 (0.0)	4 (18.2)	22
Administrative	4 (26.7)	0 (0.0)	1 (6.7)	5 (33.3)	5 (33.3)	15
Total	30 (51.7)	1 (1.7)	1 (1.7)	13 (22.4)	13 (22.4)	58

Table 3. Types of English calques.

```
Chi-square statistic: 40.94945702906753
Degrees of freedom: 3
P-value: 6.702617178823735e-09

Expected frequencies:
[[  5.87116564  27.12883436]
 [  4.26993865  19.73006135]
 [ 41.45398773 191.54601227]
 [  6.40490798  29.59509202]]
```

Figure 1. Chi-Square test results.

A chi-square test of independence was run to test the relationship between the type of anglicism (loanword or calque) and a discourse domain. The results ($\chi^2 = 40.95$, df = 3, $p < 0.001$) show a statistically significant association. All of the expected frequencies were within the acceptable range, making the chi-square test valid. The results reject the null hypothesis.

Discussion

According to the Chi-Square test results ($\chi^2 = 40.95$, df = 3, $p < 0.001$), there is a statistically significant correlation between a discourse domain (everyday, culture, academic, and administrative) and an anglicism type (calque or loanword). It means that English loanwords and calques are not randomly distributed across these domains on ChatGPT.

The academic domain has the highest percentage of loanwords (90.6%) and comparatively low calque percentage (9.4%). So, this domain shows a clear preference for English loanwords rather than calques, and it could be explained by this domain's high exposure to English terminology. There are global academic norms, and research is dominated by English.

It could suggest that the training data related to this domain is mostly in English. Thus, it could mean that ChatGPT prefers English terminology over equivalents in Kazakh.

On the other hand, in everyday and administrative domains, English loanwords and calques produced by ChatGPT are distributed in a more balanced way. In the everyday domain, there are 57.6% of loanwords and 42.4% of calques, while in the administrative domain, it is very similar (58.3% and 41.7% respectively). It suggests that, unlike academic discourse, in these less specialized domains, ChatGPT relies on more semantic and structural anglicisms, not just direct English borrowings.

The last cultural domain shows 70.8% loanwords and 29.2% calques. It is between academic and everyday+administrative domains. It could be because this discourse domain is heavily globalized, and the media enhances the use of English in other languages, through either direct borrowings or calques. ChatGPT follows this pattern as well when generating Kazakh responses.

The analysis of the types of English calques in the responses by the chatbot shows in what ways exactly ChatGPT anglicizes Kazakh. Loan translations are the major part of calques (51.7%), then it is followed by semantic and syntactic calques (both 22.4%). It means that when the chatbot does not borrow direct loanwords, it tends to translate and transfer them into Kazakh. But it is not limited to translations only. The next two calques suggest that ChatGPT not only anglicizes the vocabulary, but also the semantics and grammar of the generated text in Kazakh. Loan renderings and loan creations are only 1.7% each, meaning ChatGPT almost never integrates borrowings in Kazakh through partial rendering or full creation.

Regarding the distribution of calque types among domains, academic mostly has loan translations (81.8%), showing that ChatGPT tends to translate literally English terminology, when it is not borrowing them directly into Kazakh. On the contrary, ChatGPT's responses on administrative topics have a more balanced distribution of three calque types: loan translations, 26.7%, and semantic and syntactic calques, both 33.3%. It suggests that in this domain, English influence goes beyond mere translations to the semantics and syntax of the Kazakh language. Everyday domain has 50% semantic calques, 21.4% syntactic calques, and 28.6% loan translations. It means that in such more informal discourses, ChatGPT mostly integrates English meanings into already existing Kazakh words and phrases. In the cultural domain, loan translations are the most frequent calque type (57.1%), although the total number of calques is small (only 7). This means that when generating responses related to cultural topics, ChatGPT relies more on direct borrowings, rather than deeper semantic or structural transfer.

Overall, anglicisms are distributed very unevenly across four domains. Out of the total 326 identified anglicisms in the responses of ChatGPT, 233 (71.5%) correspond to academic, 36 (11.0%) administrative, 33 (10.1%) everyday, and 24 (7.4%) cultural. So, more than two-thirds of the total anglicisms occur in the academic domain. This huge difference from other domains suggests that when generating academic Kazakh, ChatGPT's responses are heavily influenced by English, possibly due to the English dominance in the academic context and English being the model's primary training language.

Another important point to add is the inclusion of Russian. It was just added as a reference language in findings, unusual calques created by ChatGPT, in order to avoid over-attributing every unusual calque to English in a context where Russian has a huge influence on Kazakh. Russian calques were obviously not considered anglicisms. They were helpful in identifying the source of unusual calques more accurately. The fact that there were, in total, only 2 ambiguous cases shows that the source identification of calques was relatively stable.

Conclusion

This project examined the way ChatGPT anglicizes the Kazakh language in four different discourse domains through the deep analysis of English calques and loanwords. According to the results, their distribution is uneven. ChatGPT uses the largest number of anglicisms in the academic domain (71.5%), whereas its responses in everyday, administrative, and cultural domains have considerably fewer cases. In addition, there is a significant association with a type of anglicization and a domain. ChatGPT borrows substantially more direct English words in academic Kazakh responses, followed by a cultural domain, while in everyday and administrative domains, the responses have a higher frequency of calques. The findings of this study show in what ways large language models rely on the language patterns of English in multilingual contexts.

Limitations

Despite the fact that the corpus for this project consisted of 20,000 words of raw data (ChatGPT's responses in Kazakh), the number of calques is only 58, which is comparatively smaller than loanwords (268). Some calque types have almost no or very few instances. Therefore, comparisons made among calque types in four domains cannot be interpreted with high certainty. Future research could expand the corpus even further to increase the number of anglicisms and statistical robustness.

Furthermore, it is methodologically challenging to classify all the calques' source languages in a heavy multilingual environment surrounding the Kazakh language. The influence of English and Russian on Kazakh could even be overlapping, considering a long period of language contact. So, even though Russian was added as a source language to minimize the possibility of misidentification, there is still some degree of subjective interpretation.

As this study focused only on ChatGPT powered by GPT-5, future studies could do comparisons of several large language models to examine whether anglicization trends are specific to models or whether there are broader patterns across them.

References

Adilmetova, G., Nassyrov, R., Meyerbekova, A., Karabay, A., Varol, H. A., & Chan, M.-Y. (2025). Evaluating ChatGPT's multilingual performance in clinical nutrition advice using synthetic medical text: Insights from Central Asia. *The Journal of Nutrition, 155*(3), 729–735. https://doi.org/10.1016/j.tjnut.2024.12.018.

Aitova, N. (Ed.). 2025. *Qazcorpora*. https://qazcorpora.kz/.

Asmaganbetova, K., Auyeskhan, U., Zarina, N., Serikbayev, A., & Yeskazina, M. (2024). Integration of AI techniques in low-resource language: The case of Kazakh language. In *2024 IEEE AITU: Digital Generation* (pp. 7–13). IEEE. https://doi.org/10.1109/IEEECONF61558.2024.10585350.

Betz, W. (1949). *Deutsch und Lateinisch: Die Lehnbildungen der althochdeutschen Benediktinerregel.* Bouvier.

Davies, M. (2004). *British National Corpus (from Oxford University Press)*. https://www.english-corpora.org/bnc/.

Du, W., Maimaitiyiming, Y., Nijat, M., Li, L., Hamdulla, A., & Wang, D. (2023). Automatic speech recognition for Uyghur, Kazakh, and Kyrgyz: An overview. *Applied Sciences, 13*(1), 326. https://doi.org/10.3390/app13010326.

González, F. R., & Knospe, S. (2019). *The variation of calques in European languages, with particular reference to Spanish and German: Main patterns and trends. Folia Linguistica, 53*(1), 233-276. https://doi.org/10.1515/flin-2019-2009.

Kshetri, N. (2024). Linguistic challenges in generative artificial intelligence: Implications for low-resource languages in the developing world. *Journal of Global Information Technology Management, 27*(2), 95–99. https://doi.org/10.1080/1097198X.2024.2341496.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 13171-13189).

Lee, H. (2023). The rise of ChatGPT: Exploring its potential in medical education. *Anatomical Sciences Education, 16*(5), 926-931. https://doi.org/10.1002/ase.2270.

Savchuk, S. O., Arkhangelsky, T. A., Bonch-Osmolovskaya, A. A., Donina, O. V., Kuznetsova, Y. N., Lyashevskaya, O. N., Orekhov, B. V., & Podryadchikova, M. V. (2024). Национальный корпус русского языка 2.0: новые возможности и перспективы развития [The Russian National Corpus 2.0: New possibilities and development prospects]. *Вопросы языкознания*, *2*, 7-34.

Appendix A. Chi-square Test Code in Python (SciPy library)

The following Python code was run to perform the chi-square test of independence:

```
!pip install numpy scipy

import numpy as np

from scipy.stats import chi2_contingency



table = np.array([

  [14, 19],   # Everyday

  [7, 17],    # Culture

  [22, 211],  # Academic

  [15, 21]    # Administrative

        ])



chi2, p, dof, expected = chi2_contingency(table)



print("Chi-square statistic:", chi2)
```

```
print("Degrees of freedom:", dof)

print("P-value:", p)

print("\nExpected frequencies:")

print(expected)
```