References

European Union (2024). *Translation Quality Evaluation Info Pack for External Contractors*. https://eur-lex.europa.eu/eli/dec/2011/833/oj. [15.08.2024].
Github. (2024). *COMET*. https://github.com/Unbabel/COMET [12.10.2024].
Google Patents. (2019). *DE 102011086742B4*.
https://patents.google.com/patent/US20240341407A1/en?q=(eletronic+german)&oq=el
etronic+german&sort=new&page=1&clustered=true#citedBy [20.08.2024].
Hou, X.Y. Google Drive. (2024). *Masterarbeit Dokumente*.
https://drive.google.com/drive/folders/1DbrHH8SdQM2d6e4om0yy_AhfeFQYviEq. [27.02.2025].
MIT Sloan. (2024). *Sam Altman believes AI will change the world*. https://mitsloan.mit.edu/ideas-made-to-matter/sam-altman-believes-ai-will-change-world-and-everything-else [12.07.2024].
Pérez-Ortiz, Juan Antonio, Mikel L. Forcada & Felipe Sánchez-Martínez (2022). *How neural machine translation works*. In Dorothy Kenny (ed*.). Machine translation for everyone: Empowering users in the age of artificial intelligence*, 141–164. Berlin: Language Science Press. DOI: 10.5281/zenodo.6760020.
Teo, S. (2023). *How I won Singapore's GPT-4 prompt engineering competition*. *Towards Data Science.* https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41 [12.08.2024].
Thakur, K., Barker, H.G., & Khan Pathan, A.-S. (2024). *Artificial Intelligence and Large Language Models: An Introduction to the Technological Future* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003474173.

Xinyu Hou
International University SDI Munich

Bridging Languages and LLMs in Translation: A Study on Prompt Engineering for Customized LLM Model Using the COSTAR Framework

This article focuses on the implementation of a specialized AI assistant based on the ChatGPT language model for translation tasks. By employing specific prompts and conducting translation experiments, the effectiveness of these methods will be empirically analyzed. The results can improve the efficiency and quality of translations in specific industrial sectors, thereby bridging the gap between different languages and industries.

***Key words**: large language models, machine translation, prompt engineering, ChatGPT, translation quality evaluation.*

Li, Xintian
International University SDI München
Supervisor – Mayer, Felix, PhD, Professor

https://doi.org/10.33989/pnpu.791.c3288

A Comparative Experiment of Traditional Tools and LLM Tools in Terminology Extraction: Sketch Engine and ChatGPT as Examples

Terminology plays a pivotal role in both our everyday and professional lives. Identifying widely accepted expressions marks the beginning of translation and interpreting work, while reviewing and refining terminology completes the process. With the rise of various Large Language Model (LLM) tools and the widespread speculation that LLMs may replace many translators and interpreters in the market, newcomers to the field may feel uncertain about how to navigate the constantly evolving landscape of the profession. Given the undeniable importance of

terminology work, the question arises: How well do LLM tools perform in one of the most time-consuming aspects of translation and interpretation work—terminology extraction (TE)?

1. Terminology Extraction

Terminology work consists of collecting, describing, processing, and presenting concepts and their corresponding terms, as defined by ISO 1087:2019, 3.5.1. In this study, the focus is on the first step: the extraction of terms from a designated text. Historically, this process has been manual and time-consuming, with translators reviewing lengthy texts and identifying words that "sound professional." Efforts have been made to automate this process with tools such as Sketch Engine, TermSuite, and others.

In this experiment, Sketch Engine, considered one of the leading terminology extraction tools, was chosen as the traditional tool. ChatGPT, using the GPT-5 model, was selected as the LLM tool for comparison. The Chinese and German versions of the "Government Work Report 2020" were chosen for their complexity and density of terminology, making them suitable for extraction.

2. Introduction

Sketch Engine has developed its own algorithm for terminology extraction, which combines statistical and linguistic methods. It uses the Simple Maths Score, called "Keyness," to identify term candidates by comparing their frequency in the target corpus with a reference corpus. This statistical method helps extract relevant terms by analyzing and normalizing word frequencies. A linguistic approach is applied through tools for tokenization, lemmatization, and POS-tagging, though specific grammatical rules are not disclosed (Kilgarriff, 2014, p. 2). Users can adjust extraction settings to refine results.

In multilingual TE, users can work with non-aligned or aligned corpora. After selecting the source and target languages, results are generated quickly. Sketch Engine defines "Keywords" as single-word terms and "Terms" as domain-specific multi-word terms that follow grammatical structures in the language (Sketch Engine, 12.09.2024).

In contrast, for generative artificial intelligence like GPT-5, the generation of the term list is primarily based on statistical predictions.

3. Experiment Description

This experiment compares the traditional terminology extraction tool Sketch Engine with ChatGPT, focusing on various aspects such as accuracy, ease of use, learning curve, integration into workflows, and flexibility. The objective is to evaluate their strengths and weaknesses in practical translation environments, particularly for extracting political terminology from Chinese and German texts of the "Government Work Report 2020"

The experiment involves the following procedures:

1. **Golden Dataset**: Manually extract a golden dataset of relevant political terms for comparison.
2. **Text Alignment**: The Chinese and German texts are aligned at the sentence level using Trados Studio 2022 and saved in a format compatible with both tools.

3. **TE with Sketch Engine**: Upload the aligned text into Sketch Engine, extract terms, and calculate the recall score by comparing with the golden dataset.
4. **TE with ChatGPT**: Use a specific prompt to instruct ChatGPT to extract political terms from the text, and then compare the results with the golden dataset.
5. **Comparison**: Analyze the accuracy (recall), efficiency, and required manual intervention for both methods.

The following metrics are primarily used to evaluate the performance of both tools: accuracy, recall, and F1-score. These metrics categorize extracted terms into four categories: true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN).

- **True positives (TP):** Terms correctly identified as relevant.
- **False negatives (FN):** Relevant terms that were missed by the tool.
- **False positives (FP):** Terms incorrectly identified as relevant.
- **True negatives (TN):** Terms correctly identified as irrelevant.

Accuracy reflects the proportion of terms correctly identified out of all extracted terms. It indicates the overall precision of the tool. A high accuracy value suggests that the tool is effective at minimizing irrelevant terms. However, it does not fully capture the tool's ability to retrieve all relevant terms.

The formula for accuracy is as follows:

$$Accuracy = \frac{\text{Number of Correct Terms}}{\text{Number of All Extracted Terms}}$$

Recall measures the proportion of actual relevant terms successfully extracted by the tool, providing insight into how well the tool retrieves all relevant terms. A higher recall indicates fewer missed relevant terms, even if it occasionally includes irrelevant ones.

The formula for recall is as follows:

$$Recall = \frac{\text{Number of Correct Terms}}{\text{Number of All Relevant Terms}}$$

In this study, recall is prioritized because the goal is to extract as many relevant political terms as possible, even at the cost of introducing some false positives. While the F1-score often balances accuracy and recall, it is not applicable here, as the nature of LLMs like ChatGPT tends to generate extensive lists of terms, making precision and F1-score impractical. Therefore, this study focuses on recall as the primary performance measure.

To assess the performance of the tool using these metrics, a reference dataset, known as the Golden Dataset, is required. This dataset consists of terms that have been manually extracted and serves as the benchmark for comparing the tool's output. The creation of the Golden Dataset follows a rigorous methodology based on the principles outlined in the DIN standard 2330 (2022, pp. 27-28), which stresses that multilingual terminology work should be conducted independently for each language, while ensuring consistent standards and methods across languages. This approach guarantees that the terms extracted in different languages are conceptually equivalent, preventing the terminology of one language from unduly influencing that of another.

For the purpose of this experiment, the terminology extraction focuses on terms related to government policies, political strategies, and the naming of institutions and official bodies. The Golden Dataset is manually compiled, keeping in mind the specific characteristics of political terminology, such as interdisciplinarity, the use of abbreviations, and the normative nature of the language.

4. Experiment Result

ChatGPT provided with only texts in both languages extracted 53 terms that were not present in the text, but when provided with an aligned text, this number was reduced to 43 and it extracted 30 true positive terms, with a recall rate of 28.30%.

Sketch Engine, also performed better when provided with aligned text, extracted a total of 873 terms. For calculating the recall score in the experiment, only the first 106 terms, ordered by term frequency from highest to lowest, were used. As there are 106 terms in the Golden Dataset. The recall score for Sketch Engine is 4.71%.

5. Summary

Sketch Engine showed weaknesses in extracting complex, multi-part terms, particularly when working with Chinese texts. Issues such as incorrect segmentation and the omission of important semantic elements occurred frequently. The tool also relies heavily on statistical methods, which limits the quality of results when working with smaller corpora. Additionally, the outdated Chinese dictionary posed challenges in recognizing and processing more complex terms, making the application of Sketch Engine's results for translation tasks more difficult.

In contrast, ChatGPT performed better in terms of semantic precision and flexibility. While some extracted terms did not exactly match those in the corpus, they exhibited close semantic relationships. However, ChatGPT struggled with extracting terms in their base forms, which led to grammatical inaccuracies. Moreover, the tool required more specific instructions to yield optimal results, making the preparation and task formulation more complex.

With this understanding, it is clear that manual work in TE still plays a significant role. LLMs could, however, has the potential to offer a fully automatic solution for TE, as they are capable of recognizing semantically related phrases. Looking at the development of TE—from purely statistical methods to hybrid approaches, as seen with Sketch Engine, which requires regular maintenance and updates—it is likely that TE will increasingly rely on statistical methods at a higher level.

In practice, particularly in freelance translation projects, as it is rare for translators to receive large volumes of reference texts that can be processed in tools like Sketch Engine, the ability to extract terms from limited resources is crucial for translators. In this context, LLMs offer a promising approach.

In summary, traditional tools like Sketch Engine are likely to be gradually replaced from the workflow of translators by modern, TE-specialized LLMs. This shift is primarily due to the ability of LLMs to precisely capture semantic relationships and efficiently extract relevant terms without the need for extensive manual adjustments or large amounts of reference texts.

However, traditional tools such as Sketch Engine remain valuable, particularly

in the research field. Their strength lies in providing detailed statistical analyses, such as term frequency and keyness, which enable deeper analysis of corpora. This makes them indispensable for linguistic studies, creating specialized glossaries, and investigating language phenomena. As such, traditional tools will likely continue to play a central role in academic and research-oriented contexts, while LLMs are gaining prominence in the practical aspects of TE due to their user-friendliness and efficiency.

The study highlights the need to better adapt existing traditional TE tools to the linguistic nuances and specific demands of translation work, particularly in optimizing extraction algorithms to improve their applicability in translation tasks. This presents several exciting avenues for future research.

A central research topic could be the further development and refinement of LLMs to enhance their efficiency in TE. Specifically, with the resource required to train LLMs (Yuan et al., 2025), it would be interesting to explore how LLMs can be integrated more deeply into specific fields and languages to enable even more precise and context-sensitive TE. Research could also focus on how such models can be adapted to small, specialized corpora without compromising their performance.

Furthermore, research could investigate how LLMs can be better combined with existing traditional TE tools. A hybrid solution that leverages the strengths of both approaches—detailed statistical analysis from traditional tools and semantic capture from LLMs—could significantly shape the future development of TE. A major challenge would be designing these tools to be intuitive and easy for translators to use, while still capturing highly complex linguistic structures.

Another interesting aspect for future research would be examining how the training efficiency and data requirements of LLMs can be improved. The use of transfer learning methods, which allow models to be efficiently further developed using pre-existing training data, could greatly reduce the time and resources needed for training.

Overall, numerous research directions exist that focus on improving TE tools and integrating LLMs into the workflow of translators. The continuous development of these technologies could revolutionize the translation industry and significantly ease the translation process.

References

DIN 2330:2022-07. (2022). *Terminology work - Principles and methods*. Deutsches Institut für Normung e.V.

International Organization for Standardization. (2019). *ISO 1087:2019 Terminology work and terminology science — Vocabulary*. ISO.

Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., & Suchomel, V. (2014). Finding terms in corpora for many languages with the Sketch Engine. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Retrieved from https://www.sketchengine.eu/wp-content/uploads/Finding_Terms_2014.pdf

Sketch Engine. (2025). *Keyword*. Retrieved February 28, 2024, from https://www.sketchengine.eu/my_keywords/keyword/

Yuan, J., Gao, H., Dai, D., Luo, J., Zhao, L., Zhang, Z., Xie, Z., Wei, Y. X., Wang, L., Xiao, Z., Wang, Y., Ruan, C., Zhang, M., Liang, W., & Zeng, W. (2025). Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention. https://arxiv.org/pdf/2502.11089.

Xintian Li
International University SDI München
A Comparative Experiment of Traditional Tools and LLM Tools in Terminology Extraction: Sketch Engine and ChatGPT as Examples

The article compares the performance of traditional terminology extraction (TE) tools, Sketch Engine, with generative large language models (LLMs), ChatGPT in extracting political terms from the "Government Work Report 2020." The research evaluates aspects such as accuracy, efficiency, flexibility, and integration into workflows of both tools.
Key words: *Terminology Extraction, Large Language Models, Golden Dataset.*

Shub-Oseledchik, Joseph
Internationale Hochschule SDI München
Supervisors – Drevs, M. David, MA; Wenzl, Katharina, PhD
https://doi.org/10.33989/pnpu.791.c3289

Tandem, Triade, Kooperation – oder doch lieber solo? Unterschiedliche Konstellationen und Methoden bei der Zusammenarbeit von Literaturübersetzer*innen

Bereits Martin Luther forderte: „Übersetzer sollen nicht allein sein, denn einem Einzelnen fallen die guten und richtigen Wörter nicht immer ein.“ (nach Schneider, 2009). Wenngleich Zusammenarbeit unter Literaturübersetzer*innen seit jeher praktiziert wird, erhält sie in deren Praxis und in der Übersetzungswissenschaft noch wenig Aufmerksamkeit. Die Diskussion darüber ist von Vorurteilen geprägt, die Anwendbarkeit dieser Arbeitsweise wird oft angezweifelt (vgl. Neeb/Schmidt, 2015, S. 141). Auch ist die Terminologie in diesem Bereich bisher nicht vereinheitlicht: Die Rede ist etwa von Zusammenarbeit, gemeinsamer Übersetzung, kollektivem oder kollaborativem Übersetzen. In diesem Beitrag werden die ebenfalls gängigen Begriffe *Kooperatives Übersetzen* und *Tandem* bzw. Tandemübersetzung und Übersetzertandem bevorzugt verwendet.

Einer objektiven Definition der Tandemübersetzung sind personenbezogene, konventions- und situationsbedingte Hürden gestellt. Diese Problematik beginnt bereits bei der Übersetzernennung (vgl. Huss, 2018, S. 389). Zunächst sollte man sich jedoch vom Konzept der Übersetzung als „introspektives Handeln eines vereinsamten Einzelnen“ (Orbán/Kornelius, 2008, S. 491) lösen und den Translator stattdessen als „Element einer logotechnischen Interventionsgruppe“ (Pompeu/Gomes, 2022) betrachten. Kooperatives Übersetzen lässt sich allgemein als Zusammenarbeit zweier oder mehrerer Übersetzer*innen („Agenten“) bei der Erstellung einer Übersetzung ansehen (vgl. O’Brien, 2010, S. 17).

Ausführlich beschriebene Pilotprojekte, Beobachtungen und Modelle des Kooperativen Übersetzens beziehen sich häufig auf neue, fortschrittliche Unterrichtsmethoden im universitären Umfeld (vgl. Pavlović, 2007, S. 81), etwa im Kontext der Filmuntertitelung (Małgorzewicz/Hartwich, 2017), der Buchübersetzung in einer großen Arbeitsgruppe (Mikšić, 2022) oder des Einsatzes von Think-Aloud-Protokollen zur Prozessanalyse (Pavlović, 2007). Besonders hervorzuheben sind die