Bruno Nahod

CAN WE SUBSTITUTE FIELD EXPERTS WITH CUSTOMIZED LARGE LANGUAGE MODEL IN PROCESSING SPECIALIZED LANGUAGES?

A Case Study

Abstract After the Croatian national termbase Struna ceased to receive funding in 2019, we began developing a novel model for compiling terminological collections that will not rely on field experts to provide initial terminological information. A potential solution to our issue of finding a practical and dependable source for obtaining information in the initial stages of processing terminology (i.e., the 'raw definitions') across multiple domains could be the publicly available AI language model developed by OpenAI known as GPT-4. GPT is a substantial language model that offers a range of capabilities, including answering queries, generating text, and executing tasks like translation and summarization.

A custom GPT is currently being devised as an aid module, delivering unprocessed information for terminological units that will be processed in Struna. The initial training phase involved manually providing guidelines for best practices in terminology management, which were designed based on the well-established and successful methodology we used to train field experts in the past. The second phase involves feeding TermAI with modified data that was exported from Struna.

In this paper, we will present the results of the comparative analysis of generated terminological units from TermAI and field experts in the domain of forensic sciences.

Keywords artificial intelligence; GPT-4; terminology management: Struna; definition generation

1. Introduction

In the context of terminology management in small languages like Croatian, the importance of field experts in creating high-quality definitions and hierarchical structures has been frequently stressed. In Struna (Struna, 2024), the Croatian national termbase, the role of field experts has been established as a crucial aspect of the workflow (Bratanić & Ostroški Anić, 2015). Their participation has significantly contributed to establishing and upholding standards for terminological collections in Struna. Due to unexpected combinations of circumstances, the original model of processing terminology in Struna is no longer a viable option, and we were forced to start developing a new model. A new model (yet to be finalized) will no longer include field experts as the main generators of information. To maintain the standard of quality, field experts will be used, but their role will be shifted more towards consultations during the final stages of processing.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

A potential resolution to our challenge of locating a concise and resilient resource for producing information in the initial phases of processing terminology across different domains could potentially be discovered in GPT-4 (OpenAI GPT-4, 2024.). GPT-4 is a family of models that uses deep neural networks to produce natural text (Rees & Lew, 2024) developed by OpenAI. This extensive language model is capable of tasks like responding to queries, generating text, and executing activities such as translation and summarization. The efficiency of GPT-4 in activities that can be comparable to various elements of terminology processing has been supported by studies showing it generating abstracts (Gao et al., 2023; Ma et al., 2023) as well as creating quality essays on different topics (Bašić et al., 2023; Deniko et al., 2015; Nguyen, 2023; Susnjak, 2022). It has been shown that GPT-4 can, with additional external knowledge, be utilized to successfully comprehend cause-effect relationship of Croatian Chakavian dialect (Perak et al., 2024). Furthermore, studies on using GPT to assist in lexicography (Jakubíček & Rundell, 2023; Lew, 2024) have shown that it can be successfully used in various tasks while showing low usefulness in others. The identification of senses is by far the weakest element of the tool's performance (Rees & Lew, 2024). Furthermore, Jakubíček & Rundell (2023) called attention to the difficulty GPT-4 has with figurative elements of the general language. The consensus seems to be that GPT can be a useful lexicography tool if its limitations are recognized and compensated accordingly (de Schryver, 2023). In the context of terminology management, the utilization of AI, such as GPT-4, can offer a potential solution to the challenges faced when processing terminology across different domains (Shahriar & Hayawi, 2023).

We have hypothesized that the utilization of AI in terminology management can play a pivotal role in overcoming the challenges faced in small languages like Croatian (Nahod et al., 2017). By utilizing AI, terminology management processes can be automated to a certain extent, reducing the reliance on field experts and ensuring consistency and accuracy in definitions and hierarchical structures. Our assumption is based on the observation that terminology processing seems to be exempted from just about all tasks that were recognized as problematic for GPT-4 in lexicography(de Schryver, 2023). The concepts of senses and figurative use are not part of the classical approach to terminology processing (Nahod, 2011). The tasks that GPT-4 exceeds, such as answering direct questions, summarizing, and translating, do appear to be closely matched to tasks the field experts had in the workflow of Struna. Specifically, giving exact and concise definitions for a given concept, recognizing hierarchical relations, and providing translation equivalents (Rees & Lew, 2024).

In this study, we conducted a comparative analysis between the outputs generated by TermAI and those provided by human experts in the domain of forensic sciences. This analysis focused on two key aspects: the form of definitions and the semantic accuracy of the generated content. Our investigation led to the formulation of the following hypotheses:

Hypothesis 1

H1: TermAI will perform at a comparable level of accuracy to domain experts in the form of definition test.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

H1a: There will be no significant difference in the accuracy of the form of definition test results between queries given in English and those given in Croatian.

H1b: There will be no significant difference in the accuracy of the form of definition test results across different subfields within the domain of forensic sciences.

Hypothesis 2

H2: TermAI will perform at a comparable level of accuracy to domain experts in the semantic match test and the equivalent match test.

H2a: There will be no significant difference in the accuracy of the semantic match test and the equivalent match test results between queries given in English and those given in Croatian.

H2b: There will be no significant difference in the accuracy of the semantic match test and the equivalent match test results across different subfields within the domain of forensic sciences.

These hypotheses reflect our expectation that TermAI, through its customization and training, will match the performance of human experts in most cases. However, the results of this study will also highlight areas where further refinement of the TermAI may be necessary, particularly in specialized or interdisciplinary subfields. Our ultimate goal is to implement GPT-4 into our workflow, where it would be used as the main generator of 'raw definitions', conceptual relations, and translation equivalents in the early stages of the workflow, mimicking the tasks performed by the field experts in the late model. The information generated by the GPT-4 would be processed by Croatian language experts and trained terminologists, followed by conformational editing performed by field experts.

2. Materials and Methods

Following the intensive stage of testing the capabilities and limitations of the general GPT-4, we have initiated the development of custom GPT (GPTs OpenAI, 2024) that allow the customization of the GPT for specific purposes by following the list of personalized prompts and instruction sets defined by the user.

We are currently developing a single purpose custom GPT that we have named TermAI. OpenAI is offering GPTs service (GPTs-OpenAI, 2024) that allows users to customize GPT-4. Where one can, relatively simply, customize a GPT to their own specifications for specific purposes. The main purpose of TermAI is to give a definition and term equivalents for the prompt given in the proscribed form that we are calling a query.

The development and customisation of the TermAI is planned to go through three main stages: defining the rules, manual training, and finally batch training. Naturally, all the stages of customization include multiple revisions by both trained terminologists and our project coordinator. The first stage of customization, which included the

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

logical and theoretical parts of the training, was based on the workshops we used to train field experts over the last 15 years. During this phase, we have defined the Cardinal Rules—a set of instructions and directives that can't be sidestepped or broken. Currently, there are 9 rules that TermAI must follow when interacting with a user. Most of them concern the proper form of the terminological definition, the structure of the answers, and the acceptable form of the queries. For example, Rule 1: "definition must be one sentence in the form of *genus proximum et differentia specifica*"; Rule 3: "any additional information about the term or the concept must be in the "napomena/note" part of the answer"; and Rule 6: "acceptable forms of the query are; term, term/domain, term/language and term/domain/language."

Stage two of the training covers the practical part of producing 'raw definitions' and is performed by training TermAI with a term-definition form of information sourced from Struna. We decided that the first milestone would be 1000 terminological units from the domain of forensic sciences (Table 1). The domain of forensic sciences was chosen for two reasons: it was the last project in the old model, so most of the team members were familiar with terminological collection and were able to pick the "best" terminological units for training TermAI. Secondly, forensics science was by far the most interdisciplinary collection we have processed in Struna so the terms and definitions processed have a wide semantic field, covering over 20 domains of knowledge.

Table 1: Examples of terminological data from Struna used in the second stage of TermAI training

Croatian term	English term	Definition
poligrafsko ispitivanje	polygraph testing	ispitivanje osobe i praćenje njezinih fizioloških reakcija s pomoću poligrafa radi utvrđivanja istinitosti iskaza 'the testing of a person while monitoring their physiological reactions using a polygraph to determine the truthfulness of their statement'
očevidac	eyewitness	osoba koja je svojim osjetilima neposredno opažala počinjenje kaznenoga djela 'person who directly perceived the criminal offense with their senses'
petlja	loop	crtež papilarnih linija s jednom deltom u kojemu linije teku u smjeru središta crteža gdje se uvijaju i vraćaju natrag u smjeru odakle su došle 'fingerprint pattern with one delta, in which the ridges flow towards the center of the pattern, where they loop and return in the direction from which they came'

After the first milestone of 1000 terminological units was reached, we initiated a series of tests to evaluate the current state of TermAI and identify problems and areas we need to address in future training. Following the promising results of the first series of tests, we designed a study to test TermAI performance. TermAI was given the task of generating definitions for the 20 new concepts from the domain of forensic sciences. We have selected concepts that were not previously processed in Struna and have not been used in TermAi's training so far.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The study was designed as a two-group comparative, where the same terminological units were generated by forensic sciences experts.

As a source for our concepts a small corpus was compiled using 150 thesis papers provided by the University Department of Forensic Sciences, University of Split. The extracted term candidates were filtered against the terms edited in Struna and the final collection of 20 terms were selected to be used in this analysis. Considering how forensic sciences is extremely interdisciplinary we have tried to select the concepts from the broader domain of forensic sciences that closely represented the semantical broadness of the terminological units processed in *Forensic and Criminal Investigation Project* FuNK (Bašić, 2024). Selected terms varied on multiple levels: single-word (n = 7), multi-word (n = 13); by subfields law and legislation (n = 7), security and defence sciences (n = 8), information science (n = 2), and sociology, ethics, psychiatry one (n = 1) each.

The TermAI was given queries in the form of 'term/domain' which is the form defined by one of the rules that state that for this form of a query, it should give the response representing the specific semantic field corresponding to the domain in the query, both in English and Croatian. Queries were divided into two sets: set A had all the terms in the Croatian language and set B had all the terms in English (Table 2).

Table 2: Queries by set

Set A	*neprihvatljiv dokaz* 'inadmissible evidence', *teško ubojstvo* 'aggravated murder', *digitalni dvojnik* 'digital twin', *krunski svjedok* 'crown witness', *posebna dokazna radnja* 'special investigative measure', *ugrožena osoba* 'endangered person', *kvalificirano kazneno djelo* 'serious crime', *kriptografija* 'cryptography', *nedominanta ruka* 'non-dominant hand', čestitost 'integrity'
Set B	treachery, grooming, para-suicidality, socioeconomic status, hybrid warfare, hybrid threat, national security, trinitrotoluene, espionage, counterintelligence

The main goal was to evaluate TermAI's ability to generate definitions in the domain in which was trained. That would, hopefully, give us a better understanding of its current state of development and allow us to identify its weaknesses. Our secondary goal was to determine the quality of term translation, both from Croatian to English and from English to Croatian.

For the purpose of this study, we formed a group of forensic sciences experts (n=3), all of whom were previously involved in the FuNK project and had undertaken a full terminological and linguistic training 5 years ago. They were asked to provide definitions and English equivalents for the same 20 concepts. For the ease of the data analysis, we have named them Group A, while TermAi was named Group B.

3. Results

The data generated by Group B (TermAI) was evaluated and scored against the data generated by Group A (Table 3). Generated terminological units were parsed

into three categories: **equivalent match** – where we were looking if the generated Croatian and English equivalents correspond to their pairs from the Group A, **semantic match** – where we were looking how closely the definition corresponds to the definitions from the Group A, and **form score** – where each definition was scored based on the terminological principles enforced in Struna. And finally, following the same rules, we have also scored the definitions generated by the Group A. The scoring for the **form** was 1–5, where 5 was a perfect definition, 4 needing minimal terminological and/or language intervention, 3 – a considerable intervention is needed, 2 – definition needs a full rewrite with consultations, and 1 – a new definition is needed.

Table 3: Values by category used in evaluation of the data

Equivalent match	**Semantic match**	**Form score**
Perfect match	Match	1–5
Synonym	Different aspect	
Fail	Fail	

The first variable analysed to test Hypothesis 1 was the terminological form of definitions produced by both Group A (domain experts) and Group B (TermAI). The descriptive statistics for the form scores across both groups are summarized below:

Statistic	Group A form score	Group B form score
Count	20	20
Mean	4.42	4.84
Standard Deviation	0.77	0.37
Min	3.00	4.00
25th Percentile	4.00	5.00
50th Percentile (Median)	5.00	5.00
75th Percentile	5.00	5.00
Max	5.00	5.00

To evaluate the difference in form scores between Group A and Group B, an independent t-test was conducted. The results were as follows:

T-statistic: -2.147

P-value: 0.039

Given that the p-value is less than the significance level of 0.05, we can conclude that there is a statistically significant difference between the form scores of Group A and Group B. Specifically, TermAI (Group B) demonstrated a higher mean form score compared to the human experts (Group A), indicating that TermAI was able to produce terminological definitions that adhered more closely to the standardized form criteria.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

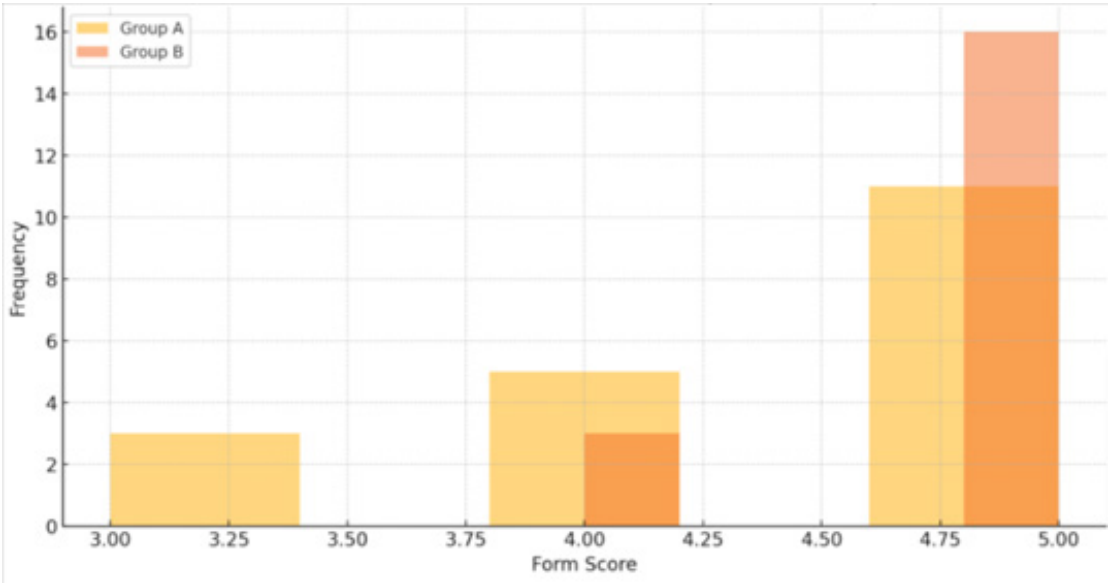

Fig. 1: Distributions for the **form score** for Groups A & B

The overall distribution analysis of the form scores for TermAI indicates a high level of terminological accuracy, with all scores falling within the 4 to 5 range on a 1–5 scale. This suggests that TermAI consistently produced well-formed definitions. A slight difference in performance was observed between queries given in Croatian and those in English, with the Croatian queries achieving a 90% ratio of scores at 5, compared to an 80% ratio for English queries. This observation is supported by a weak negative correlation (-0.140) between the query language and the form score, suggesting only a minimal impact of language on TermAI's performance.

To test Hypothesis 2 and assess whether TermAI performs at a comparable level of accuracy across different subfields, we analyzed the correlation between the Equivalent Match Score and the Semantic Match Score with the various subfields included in this study. The correlations are summarized below:

Subfield	Correlation with Equivalent Match Score
IT	0.249
SDS	0.440
law	-0.815
rest	0.316
Subfield	Correlation with Semantic Match Score
IT	-0.094
SDS	0.131
law	-0.095
rest	0.031

The data reveal a notable negative correlation between the law subfield and both the Equivalent Match Score (-0.815) and the Semantic Match Score (-0.095). This suggests that TermAI's performance is less accurate in the law subfield compared to others, indicating that higher scores are less likely to be associated with legal terminology.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

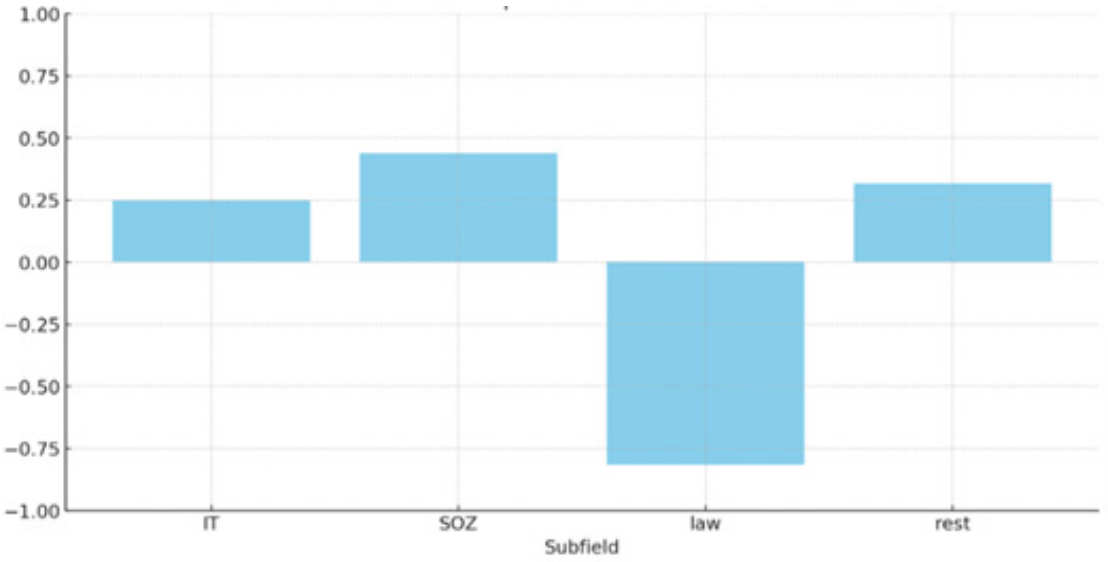

Fig. 2: Correlation between Equivalent match and Subfield

Fig. 3: Correlation between Semantic match and Subfield

A quantitative analysis of the "Equivalent Match" for each "Query Language" was conducted to further explore the performance of TermAI, particularly in relation to Hypothesis 2. The analysis revealed the following distributions:

- Croatian queries: Fail: 20%, Original 0%, Perfect match 50%, Synonym 30%
- English queries: Fail 10%, Original 20%, Perfect Match 70%, Synonym 0%.

The results indicate that for Croatian queries, TermAI produced a perfect match for 50% of the cases, while for English queries, the perfect match rate was higher at 70%. Conversely, the failure rate was higher for Croatian queries (20%) compared to English queries (10%).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

A weak negative correlation (-0.169) between the "Equivalent Match" and the language of the query suggests a slight tendency for the accuracy of equivalent matches to vary with the language. Specifically, as the query language changes from Croatian to English, the likelihood of achieving a perfect or original equivalent match slightly decreases in its numerical value.

4. Discussion

As far as we are aware, this is the first study that tested ChatGPT as a terminology unit generating tool in comparison to human field experts. Our study showed that TermAI preformed beyond expectations for its current state of development. Compared to (subjectively) best terminologically trained field experts in the domain of forensic sciences, TermAI performed on the level that would be acceptable for the terminology processing in Struna. With the mean score of 4.85 for **form score** which represents the level of acceptancy in the definition form it slightly overperformed trained domain experts (mean 4.42).

Notably, the form of the definitions is not the most important aspect of terminology processing, but given our experience reaching this level of performance has generally been a struggle while training domain experts. The accuracy of definitions which were presented with the **semantic match** variable does show observable problems in matching domain experts. Of the 20 definitions generated by TermAI, 13 (65%) definitions that were semantically identical to the corresponding definition given by the domain experts, as far as we were able to judge. The remaining 7 (35%) were not wrong *per se*, but were generated with a noticeable semantic shift. The majority of them (42%) were from the subfield of law and legislation, the results when combined with a negative correlation (-0.815) for **equivalent match** with the same subfield does show that TermAI has performance problems in that subfield when dealing in interdisciplinary domain of forensic sciences. Considering the nature of the law and legislation terminology, which can be described as highly constricted and dependent on specific aspects such as country, tradition, and case-by-case application (Fajfar et al., 2019), this should not come as a surprise.

The analysis of the term equivalents generation TermAI has also performed well with 60% of generated equivalents being an exact match to the ones provided by domain experts. In 3 cases (1.5%) TermAI generated acceptable synonyms, results which we cannot but count as accurate. The rest of the generated equivalents consist of 3 errors, where TermAI generated wrong terms of those 2 were in English and 1 in Croatian, all of them in the subfield of law and legislation: *key witness* for *crown witness*, *vulnerable person* for *endangered person* and *izdaja* for *podmuklost*. Notably the Croatian term *podmuklost* which was given by the domain experts for the English term *treachery* is a strange choice, and (subjectively) we would consider *izdaja* to a better translation. The remaining two equivalents TermAI left in their original English form when generating Croatian terms, which could be considered a better solution than generating a wrong term. Notably, both Croatian terms provided by forensics experts (*mamljenje* for *grooming* and *spufiranje* for *spoofing*) are recent

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

reterminologization and newly coined ones respectively. Therefore, it should not come as a surprise that TermAI wasn't able to provide a matching term equivalent. During this study TermAI hasn't presented a single case of "hallucinations", the fact that we could correlate to the strict and comprehensive theoretical training we provided at the very early stages of development.

5. Conclusion

The outcomes of our study surpassed our expectations. Despite anticipating significant challenges in the early stages of training, our custom GPT TermAI demonstrated remarkable performance. Achieving an overall accuracy exceeding 60% in generating both definitions and term equivalents was more than we initially anticipated at this stage. Notably, if we consider that most of the excluded generated definition-equivalent answers are not fundamentally incorrect but rather not immediately usable, the percentage of potentially useful 'raw definitions' and term equivalents can subjectively be elevated to 90%.

The objectivity of our analysis warrants attention. Evaluating any written information is inherently subjective, especially when dealing with specialized knowledge meanings and definitions. Acknowledging this, we consciously designed the study with certain controls. Definitions generated by TermAI, intended as 'raw definitions' for further refinement by trained terminologists, were compared to those provided by domain experts. These experts, trained by our team, also had experience working on the FuNK project in Struna. Typically, we receive 'raw definitions' from domain experts in fields outside our expertise, such as physics, engineering, and forensic sciences. For this study, we intentionally selected the domain of forensic sciences due to our team members' involvement in the FuNK project, enabling us to promptly verify the accuracy of generated answers. We can consider this a deviation from standard practice, which enabled us to be stricter while evaluating definitions. Considering this bias, the results are even more encouraging.

The subsequent phase involves training TermAI on 20,000 terminological units provided in domain-based batches. This will follow additional studies aimed at identifying limitations and challenges with our custom GPT. We hope that these insights will refine TermAI's capabilities and introduce new features. Our results suggest that additional effort is required to enhance TermAI's handling of interdisciplinary concepts, particularly in integrating specific semantic aspects of definitions within assigned domains. Moreover, it is clear that a synonyms section should be incorporated into TermAI's answers to avoid potential issues for future users. The analysis also indicates a need to improve TermAI's term translation abilities, especially for English-language queries.

At this developmental stage, the most significant progress is observed in the quality of standardized generated definitions. While ChatGPT is generally unpredictable in both form and content, our efforts in training and implementing Cardinal Rules have resulted in a bot that is relatively stable and predictable.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

AI models like GPT-4 are paving new paths in language research. Despite the inherent risks of over-reliance, these models can be valuable tools when their capabilities and limitations are understood. At this stage, despite TermAI's promising performance, we are hesitant to grant it full autonomy in user interactions. We hope that continued research and training will enable us to develop TermAI to a level where it can effectively complement the content of Struna.

And finally, to answer the question from the title – Can we substitute field experts with AI in early stages of specialized languages processing? The results of this study suggests that at this stage we cannot fully depend on AI to substitute real human experts. On the other hand, the results are promising with strong implication that with further research and training AI models could be turned into a useful tool in terminology processing for Croatian language.

References

Bašić, Ž. (n.d.). HRZZ - STRUNA - FuNK. Retrieved June 1, 2024, from https://forenzika.unist.hr/forenzicno-kriminalisticko-nazivlje-struna-funk/

Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications*, *10*(1), 1–6. https://doi.org/10.1057/s41599-023-02269-7

Bratanić, M., & Ostroški Anić, A. (2015). Koncepcija i ustrojstvo terminološke baze Struna. In M. Bratanić, I. Brač, & B. Pritchard (Eds.), *Od Šuleka do Schengena: Terminološki, terminografski i prijevodni aspekti jezika struke* (pp. 57–74). Institut za hrvatski jezik i jezikoslovlje.

Deniko, R. V., Shchitova, O. G., Shchitova, D. A., & Lan, N. T. (2015). Learning Terminology in the Age of Higher Education Internationalization: Problems and Solutions. *Procedia - Social and Behavioral Sciences*, 215(June), 107–111. https://doi.org/10.1016/j.sbspro.2015.11.582

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387. https://doi.org/10.1093/ijl/ecad021

Fajfar, T., Tomazin, M. J., & Karer, M. Z. (2019). Slovenian legal terminology and its presentation in the dictionary of legal terminology. *Jezikoslovni zapiski*, *25*(1), 53–66. https://doi.org/10.3986/jz.v25i1.7565

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *Npj Digital Medicine*, *6*(1), 1–5. https://doi.org/10.1038/s41746-023-00819-6

GPTs-OpenAI. (2024). Introducing GPTs. https://openai.com/blog/introducing-gpts

Jakubíček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, & C. Tiberius (Eds.), *Proceedings of Electronic Lexicography in the 21st Century Conference* (pp. 518–533). Lexical Computing CZ, s.r.o.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, *11*(1), 1-8. https://doi.org/10.1057/s41599-024-02889-7

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). AI vs. Human -- Differentiation Analysis of Scientific Content Generation. January. https://doi.org/10.48550/arXiv.2301.10416

Nahod, B. (2011). Od polisemna leksema do homonimna naziva. In L. Pon, V. Karabalić, & S. Cimer (Eds.), *Aktualna istraživanja u primijenjenoj lingvistici. Zbornik radova s 25. međunarodnog skupa HDPL-a održanog 12.-14. svibnja 2011. u Osijeku* (pp. 15–26). Hrvatsko društvo za primijenjenu lingvistiku.

Nahod, B., Nahod, P. V., & Bjeloš, M. (2017). Three aspects of processing ophthalmological terminology in a "small language": a case of Croatian term bank Struna. *Jazykovedny casopis*, *68*(2), 287–295. https://doi.org/10.1515/jazcas-2017-0038

Nguyen, M. H. (2023). Academic writing and AI: Day-1 experiment. Center for Open Science 2023, January. https://osf.io/kr29c/download

OpenAI. (2024). Introducing GPTs. https://openai.com/blog/introducing-gpts

OpenAI GPT-4. (n.d.). https://openai.com/product/gpt-4

Perak, B., Beliga, S., & Meštrović, A. (2024). Incorporating Dialect Understanding Into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, M. Zampieri, P. Nakov, & J. Tiedemann (Eds.), *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects* (pp. 220–229). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.vardial-1.19

Rees, G. P., & Lew, R. (2024). The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task. *International Journal of Lexicography*, *37*(1), 50–74. https://doi.org/10.1093/ijl/ecad030

Shahriar, S., & Hayawi, K. (2023). Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. *Artificial Intelligence and Applications*, 2(1), 11–20. https://doi.org/10.47852/bonviewaia3202939

Struna. (2024). Struna. http://struna.ihjj.hr/

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? 1–21. http://arxiv.org/abs/2212.09292

Acknowledgements

This paper was created within the framework of the project TermAI: Development and Modernization of Croatian Terminology, which is funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the authors and do not necessarily reflect the official positions of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Contact information

Bruno Nahod
Institute for the Croatian Language
bnahod@ihjj.hr

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.