The Best of Two Worlds: Exploring the Synergy between Human Expertise and AI in Lexicography

Lars Trap-Jensen
Society for Danish Language and Literature
E-mail: ltj@dsl.dk

Abstract

The arrival of chatbots is probably going to change our daily lives, and it will surely change the conditions for making dictionaries. In this paper, I account for the lexicographic studies that have been carried out in the first year of ChatGPT since its introduction in November 2022, and use them, combined with my own experience as a regular user, as background for an evaluation and discussion of the advantages and disadvantages of using artificial intelligence in lexicographic work. In my own use, I have used the free versions of ChatGPT, versions 3.5 and 4 as a tool in practical lexicography.

Keywords: ChatGPT; artificial intelligence; Large Language Models; dictionary-making; lexicographic tools; prompting; hallucination

1. Introduction

The field of artificial intelligence (AI) is not new. It dates back at least to the summer of 1956, when the term was coined at the Dartmouth conference, and where AI as a discipline was founded. Neither is it new in lexicography, where the subfield of natural language processing (NLP) has long assisted lexicographers in automating some of the routines involved in dictionary-making, such as analysing corpora, extracting examples, and tagging texts with parts of speech. What is new, however, is that the field has expanded from a narrow group of computer scientists and computational linguists to include the general public. This happened when AI chatbots, capable of engaging in humanlike conversations with a human user via NLP, became available. The company OpenAI (dominated by Microsoft) launched their ChatGPT in November 2022, and since then things have developed rapidly: Google introduced Bard in March 2023, OpenAI launched an improved version using GPT-4 instead of GPT-3.5 in the same month, and in November 2023 xAI (founded by Elon Musk) followed suit with Grok.

When I was invited to present a paper at the conference *Lexicography in the XXI Century*, the release of ChatGPT was just two months old, and I had little idea of how it would evolve in the following months. Like many other people, I was impressed with ChatGPT and suspected that it might become a game-changer, also in the field of lexicography.

I decided to use it in my daily work to discover its strengths and weaknesses, starting with ChatGPT-3.5 and later also version 4. However, I did not compare it with Bard and

Grok, as they were released too late for the conference (Bard was not introduced in Europe until the summer of 2023 for legal reasons). My original idea was to use it and evaluate its performance for all the different microstructure elements in a dictionary entry, but since I was not the only one to do this, I will rather try to integrate my own observations with those of other studies and assess how useful chatbots are overall for lexicographic work.

2. About Large Language Models (LLM)

ChatGPT and similar chatbots are so-called Large Language Models. This means that they are based on huge amounts of data that have been pre-trained through machine learning techniques in a way that mimics the neural networks of the brain, hence the term artificial neural networks. ChatGPT is a Large Language Model that uses artificial neural networks called transformers. The name GPT stands for „generative, pre-trained transformer". The type of language model used in ChatGPT-3 is a static model, which means that it does not „know" things that have happened after the time of its latest underlying data. Therefore, it needs to be updated at regular intervals. By contrast, ChatGPT-4 has been integrated into other Microsoft applications, such as Bing or Skype, and can provide feedback by checking real-time information on the internet while answering.

From the user's perspective, communication with the language model takes place via an interface, such as an app, a website, or a dedicated bot function. This may seem a trivial thing to mention, but it is in fact an important argument among those who predict that chatbots will outcompete dictionaries. They claim that this way of communicating is more „natural" than having to turn to an external tool like a dictionary and look up problematic words while reading a given text. Also keep in mind that the prompt does not have to be entered on a keyboard; it could also be given in the form of an oral input.

ChatGPT is a probabilistic model that uses statistical techniques and machine learning algorithms to predict the most likely response based on the input it receives. It does not have any „knowledge" and it does not infer, deduce, or reason anything. It is generative: it generates an output, for example a string of words, based on what is the most likely answer to the input it received, as calculated by the algorithm. ChatGPT in its latest version can also handle images and sound, or, as they say, ChatGPT can now „see, hear and speak", meaning that it is capable of interpreting spoken words and responding with a synthetic voice or by processing images. For the present purposes, however, I limit myself to looking at text in and text out.

In the following, I will present the outcome of existing studies and supply them with my own experience with ChatGPT as a lexicographic tool. The focus will be on two main aspects: the quality and the quantity of the output generated by ChatGPT, and its performance as compared with existing lexicographic tools, notably the use of corpora, as well as the human lexicographer.

My main questions of interest are:

- How accurate and reliable is ChatGPT at producing lexicographic information, such as lemma candidates, pronunciation, definitions, examples, synonyms, antonyms, collocations, and usage examples?
- How efficient and productive is ChatGPT at generating lexicographic information, compared to other lexicographic tools and human lexicographers?

- What are the advantages and disadvantages of using ChatGPT as a lexicographic tool, in terms of usability, reliability, and legal/ethical issues?
- To answer these questions, I will account for the observations of the previous studies and compare the results with my own experience from daily use.

3. Studies of ChatGPT

Since the release of ChatGPT in 2022, we have seen several conferences and publications addressing the impact of ChatGPT in lexicography. Gilles-Maurice de Schryver started the debate in Tokyo at a seminar held by the Center for Open Data in the Humanities (CODH) in February 2023, where he and his colleague David Joffe welcomed the new technology and demonstrated how ChatGPT had been integrated into their dictionary writing system, Tlex, with promising results (de Schryver & Joffe 2023).

At the conferences held by the regional lexicographical associations in North America and Asia, several presentations were specifically dedicated to the use of AI in lexicography, and this was especially evident at the eLex conference in the Czech Republic, where several keynotes, regular papers, and a round table discussion addressed the topic.

Finally, two papers must be mentioned from the autumn of 2023: de Schryver's summary article in International Journal of Lexicography (de Schryver 2023), and an article by Robert Lew in which he tests ChatGPT's ability to compete with human lexicographers for various information elements in a dictionary entry (Lew 2023).

In the following, I will examine the potential applicability of ChatGPT in dictionary writing. In principle, a chatbot may be used for any microstructural element of an entry, but it is worth investigating whether its performance is better for some aspects than others. I will therefore go through each element and note the assessment as reported by the studies in the presentations and articles just mentioned. I will refer to these collectively as „the experiments", and I will add my own experience from using ChatGPT when relevant.

3.1 Lemma selection

The investigation into lemma selection proves to be a challenge, and only one source addresses this issue in the context of the experiments. In their Asialex paper, Erin McKean and Will Fitzgerald conclude that ChatGPT's performance is rather poor, resulting in what they call „disappointing results" (McKean & Fitzgerald 2023:19). My own experience partly agrees with this observation. It is important to note, however, that I used only the free versions of ChatGPT, which have a limit on the number of words they can process. Therefore, requesting a comprehensive list of headwords for a large dictionary is not realistic. However, one can make more specific inquiries, focusing on particular fields. I tried to do this using the following prompt:

> Give me a list of headwords from the subject 'track and field sports' to be used for a medium-sized dictionary aimed at graduate students. The words should be fairly frequent and established but not found in most dictionaries. A headword may consist of one or more words. Provide as many headwords as possible.

This prompt produced a list of headwords that, in my opinion, are quite acceptable. Examples include:

> *decathlon, discus, shot put, triple jump, pentathlon, relay race, shot putter, starting block, starting pistol, photo finish, steeple chaser, decathlete, long jump pit, heptathlon, pole vault, hurdles, high jump, racewalk, ultramarathon, race-walker, takeoff board, track spikes, split times, lap counter, race bib, finish line tape, javelin, steeplechase, hammer throw, long jump, cross country, Fosbury flop, discus thrower, cross-country course, hurdler, high jump bar, Paralympics, relay baton, shot put circle*

However, the second requirement of the prompt turned out to be more problematic: only about 40% of the items are not already present in dictionaries, as checked against Collins English Dictionary (33%), Oxford Dictionary of English (33%) and Cambridge Advanced Learner's Dictionary (46%). Whether one labels this outcome as disappointing is a matter of temperament, but it is reasonable to suggest that there are other, perhaps more obvious, corpus-based methods for identifying candidate headwords. It is noteworthy, though, that the premium version of ChatGPT allows users to install plugins and upload their own files into the language model. As a result, one can create customized ChatGPT versions based on one's own data, such as existing dictionary entries and corpus texts. While I have not personally tested this feature, it offers the possibility for a more tailored approach, with the bot having detailed knowledge of the user's specific resources, including what is included and what is not. Further exploration of this aspect may reveal insights beyond the scope of this discussion.

3.2 Pronunciation

McKean & Fitzgerald (2023:24) report that IPA pronunciation is „wildly inconsistent", while de Schryver (2023:19), on the other hand, concludes that it is „good", so we find conflicting evaluations for this element. In my own experience, ChatGPT generally performs rather well, making only few mistakes. By default, American pronunciation is used, but one can specify another regional standard in the prompt if preferred.

During my testing, I encountered a few problematic cases. Version 3.5 had some difficulties with handling primary and secondary stress correctly in polysyllabic words, as shown in table 1.

Word	***ChatGPT-3.5***	***ChatGPT-4***
flabbergasted	/flæˈb.ə.gɑːstɪd/	/ˈflæb.ə.gɑː.stɪd/
perpendicular	/pərˈpɛndɪkjələr/	/ˌpɜː.pənˈdɪk.jə.lər/
dandelion	/ˈdændɪlaɪən/	/ˈdændɪlaɪənz/

Table 1: *ChatGPT pronunciation of selected words*

This issue was improved in version 4, which in turn introduced other bugs. Here the pronunciation of *dandelion* is in the plural although this was correct in the previous version.

However, these are minor irregularities, and in most cases ChatGPT gets it right. Remember also that version 4 has internet access and provides links to some sources. Whether the exact source for its own pronunciation is always among them, I have not been able to establish with certainty.

3.3 Morphology

The result for morphology is comparable to that of pronunciation. Only de Schryver (2023) investigated inflection, and he concludes that ChatGPT's performance is „good". My testing reaches a similar conclusion. English inflectional morphology is, of course, fairly simple compared to other languages such as Georgian or Icelandic, but it nevertheless seems to get it right most of the time. I checked a number of irregular English words, and the only errors I found were in the comparison paradigm of the adjectives, for example *little, littler, littlest*; *peculiar, peculiarer, peculiarst* (along with the periphrastic forms *more* and *most peculiar*).

3.4 Definitions

The area where the different authors agree the most is definitions. They all report that ChatGPT is doing a fine job as a defining lexicographer, and I can confirm this. It is surprisingly eloquent, concise, and often includes aspects that you might not have thought of immediately. It is probably the most impressive thing about ChatGPT and a good reason to use it regardless of your opinion of it otherwise.

Three authors have examined the style labels, and they all conclude that ChatGPT does well, not only for controversial words related to sex, race, and minority issues, but also for formal, archaic, or old-fashioned words. This came as a surprise to me, but I must admit that it does hit the mark rather well. However, one caveat is that the results are not consistent: when I checked the examples given by Jakubiček & Rundell (2023), ChatGPT did not give me the same labels as they report. For *half-caste*, Jakubiček & Rundell report the label 'offensive', but in my case, ChatGPT suggested 'informal'. For their example *betimes*, they report 'archaic', whereas I got 'formal'. This is not unusual; in fact, it is quite normal. So, when I repeated my requests, it returned the label 'highly offensive and outdated' for *half-caste*, and 'old-fashioned' for *betimes*, where it had 'formal' the first time. This is what is sometimes called the non-deterministic nature of the LLM (same question, different answer) and will be discussed in more detail below.

3.5 Other parts of the meaning description

The reports are mixed on synonyms and other word relations. Nichols (2023) claims that ChatGPT cannot easily deal with synonyms, while Jakubiček & Rundell and de Schryver maintain that it is good. My own experience is somewhere in between. In my experience, its suggestions are sometimes quite apt and at other times it offers what is more a paraphrase than a real synonym. A case in point is *pony*, for which it suggested „little horse" or „small horse" as synonyms. In other cases, it offered very generic synonyms that are hard

to use as substitutions except in a few contexts. In the case of *they*, it returned „them“, „those“, and „others“ that are not very useful as substitutable synonyms.

Expressed in more general terms, it seems reasonable to assume that chatbots are not good at making evaluations as they do not have any intuition about language in the same way as lexicographers have. And as choosing the most suitable synonyms requires evaluative skills, you would not expect ChatGPT to perform particularly well for this task. But it is not that simple.

In figure 1, for example, I asked for synonyms of *treacherous*, and the suggestions given are not bad at all.

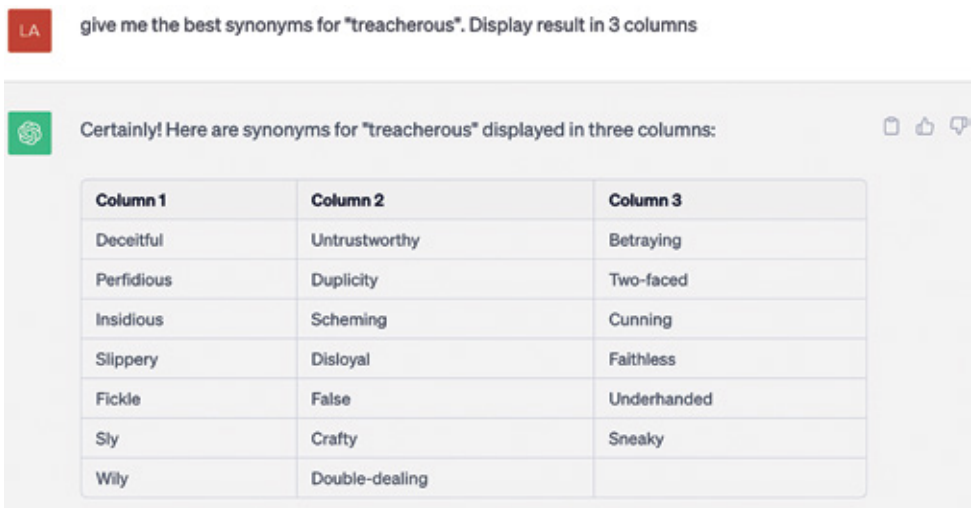

Figure 1: *synonyms of treacherous, ChatGPT-3.5*

These results were provided by version 3.5. For comparison, we can look at the response to the same prompt in version 4. This is given in figure 2.

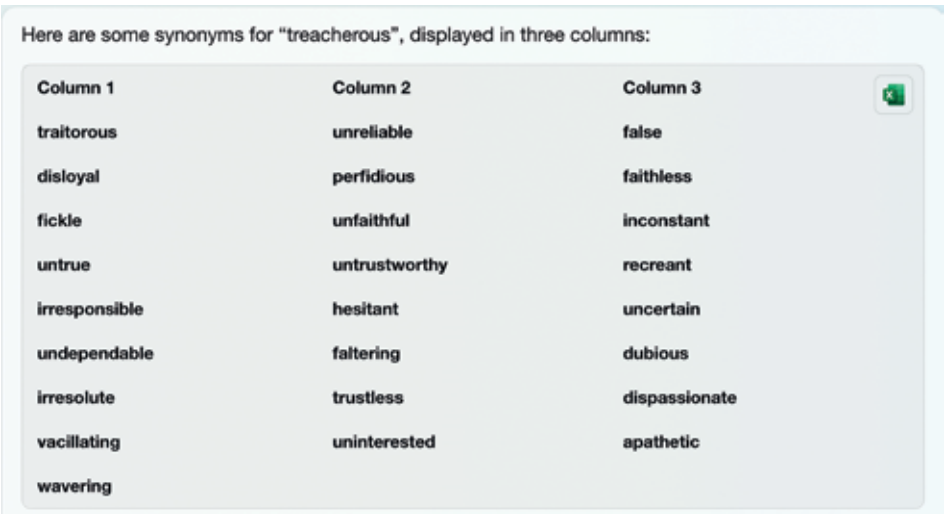

Figure 2: *synonyms for treacherous, ChatGPT-4*

Surprisingly, the results given by version 4 are not better, in my opinion. Disappointingly, the top candidate for version 3.5, *deceitful*, is not even on the list, and the reverse is also true: *traitorous*, which tops the second list, is not on the list by version 3.5.

But overall, ChatGPT is doing quite well with synonyms, and the same is true for register labels ('formal', 'derogatory', 'slang', etc.). For other elements, the conclusion is less clear. For citations and example sentences, Michael Rundell (2023:14) concluded that it is „consistently bad", while Lew is less categorical but still critical. In Lew (2023), four human experts were asked to evaluate the output of selected entries from the COBUILD dictionary with the corresponding articles generated by ChatGPT, not knowing which entries were which. Lew is led to conclude that the AI-generated examples are less satisfactory but, by rephrasing his prompt, he was able to produce a result that was more elaborate and varied. So, the picture is not clear. It may be going too far to conclude that ChatGPT is not doing well at all. It is certainly not always poor, and in some instances, it is doing a good job.

Word sense division is another area where the experiences are mixed. Jakubiček & Rundell are critical, reporting that ChatGPT does not do well, as it either splits too much, omits existing senses, or even invents non-existent ones. They mention the entry for *climate* where they found that ChatGPT listed six different senses that were actually all aspects of the same sense, while at the same time neglected an existing sense. Lew is more positive but acknowledges that further investigation is needed for sense identification. He reports that some of his evaluators noted that some of the senses needed to be either split or merged. My own use produced mixed results: sometimes ChatGPT was accurate, sometimes it split too much, and I also experienced examples of both missing and non-existing senses.

The same blurred picture appears for the remaining information elements: grammar, example sentences, and etymology: ChatGPT is neither hopeless nor perfect.

4 Evaluation and discussion

4.1 Advantages and strengths of ChatGPT

ChatGPT is a user-friendly computational linguistic tool with many advantages. It requires minimal computer skills, making it easy to start and use, even for users without special prerequisites. It is cheap, and there is also a free version available for everyone to use. The answers it provides in response to the prompts are usually well-formed, straightforward, good, and idiomatic. Communicating with it is uncomplicated and natural, which makes it easy and simple to use. This may explain why all the testing authors agree that definitions are where ChatGPT performs best. It is often as good as or better than a human lexicographer at defining, and sometimes it adds relevant details that a lexicographer might not think of. Some authors claim that it is superior to the human.

To my surprise, ChatGPT is often good at finding synonyms and other word relations, though not always. Jakubiček & Rundell note that ChatGPT „generates decent (though not perfect) responses" when asked to classify a list of words into synonyms and antonyms. It is also important to remember that the technology is constantly evolving. What de Schryver & Joffe reported in February, or what was discussed at the eLex conference in June 2023, may be out of date by winter 2024. Given that ChatGPT is less than a year old, it is easy to imagine how much better it will be in the future. One should remember that it improves every day. Finally, it is easy to integrate and customise ChatGPT to a specific project. I have

not tried this myself, but de Schryver & Joffe report that within a few months of its release they integrated a version of ChatGPT within their TLex dictionary writing system. In the software, a standardised prompt is used by default but you can also specify your own prompt if you prefer to do so.

4.2 Problems and concerns of ChatGPT

4.2.1 Reliability

Perhaps the most striking outcome of the comparison between the different testers is the low degree of author agreement. There may be several reasons for this. One thing is that the input given in the different tests is not the same – the testers have used different prompts. Another reason is that the tests were done at different times, and, as we just saw, the technology is constantly evolving. But even if we controlled for these factors, the results would still be variable and inconclusive. In fact, this is something that is inherent in the model itself. Large Language Models are probabilistic rather than deterministic in nature. This means that if different people ask the chatbot the same question, they will get different answers. This feature is even built into the interface in version 3.5 which offers a „regenerate“ option: if you are not satisfied with the response received, you can ask again and get a different answer. At the same time, ChatGPT can learn from the feedback, and you can evaluate the output by giving thumbs up or down.

For the user, this may be a nice feature, but from a research perspective, it poses a challenge. If the same question generates different answers, it is impossible to draw any conclusions from a comparison of different answers, and this is also what Jakubiček & Rundell conclude. Usually, we expect the results and conclusions of scientific studies to be intersubjectively transferable, so that the results can be confirmed or falsified if the study is repeated. These requirements are essential to ensure that research is credible and can have practical applicability. Other researchers should be able to replicate the study by following the same methods and get similar results. This is crucial for verifying the validity of their findings. But with AI technology of this kind, this is not the case: we will not get the same result even if we use the exact same method.

This problem is related to the „black box“ of the underlying algorithms in the Large Language Models used by OpenAI, Google, and other players on the market. They will not disclose the algorithms used, and they will probably never be going to. All we have is the input and the output. We may be able to judge whether one answer is better or more plausible than another, but we will not be able to know how and why it produced the answer it did. So, the obvious question is: can we trust it?

We are in uncharted territory here, and everyone must make their own judgment. I think it is a problem, and I am not alone. At the eLex round table in June, Michael Rundell called it a „deal-breaker“. In our current way of working, we can always verify the information given in a dictionary by going back and consulting the underlying data: what was the evidence for this in our corpus or in the collections of slips? With AI chatbots, this is not possible.

4.2.2 Hallucination

Another and more obvious problem is what is called 'hallucination': the fact that ChatGPT sometimes provides imprecise, incorrect or inappropriate information among otherwise acceptable responses. This is especially problematic if the information is accepted uncritically, for example as part of an automated process.

Consider the following example, shown in figure 3, where I asked for an entry outline of the lemma „spoon" and received a fine definition of the kitchen utensil. However, for the second sense, it gives „a term of endearment used towards a romantic partner", along with two example sentences: „Good morning, my love. You're my favourite spoon" and „Let's cuddle; you're the best spoon ever". We cannot know where this comes from. It is eloquently phrased and sounds very convincing. But it is not in any dictionary because it is not part of the language.

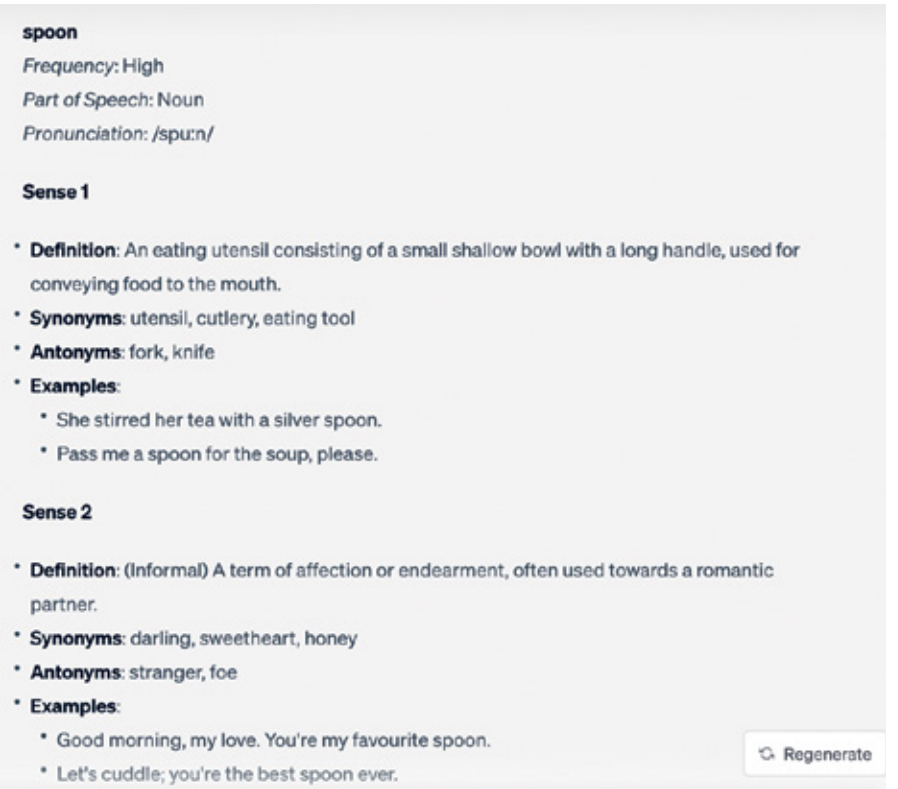

Figure 3: *ChatGPT, entry for spoon*

In order to be sure, I tried to ask if it was absolutely certain about the answer, as it sometimes makes an excuse, admitting that it made a mistake. The answer is shown in figure 4.

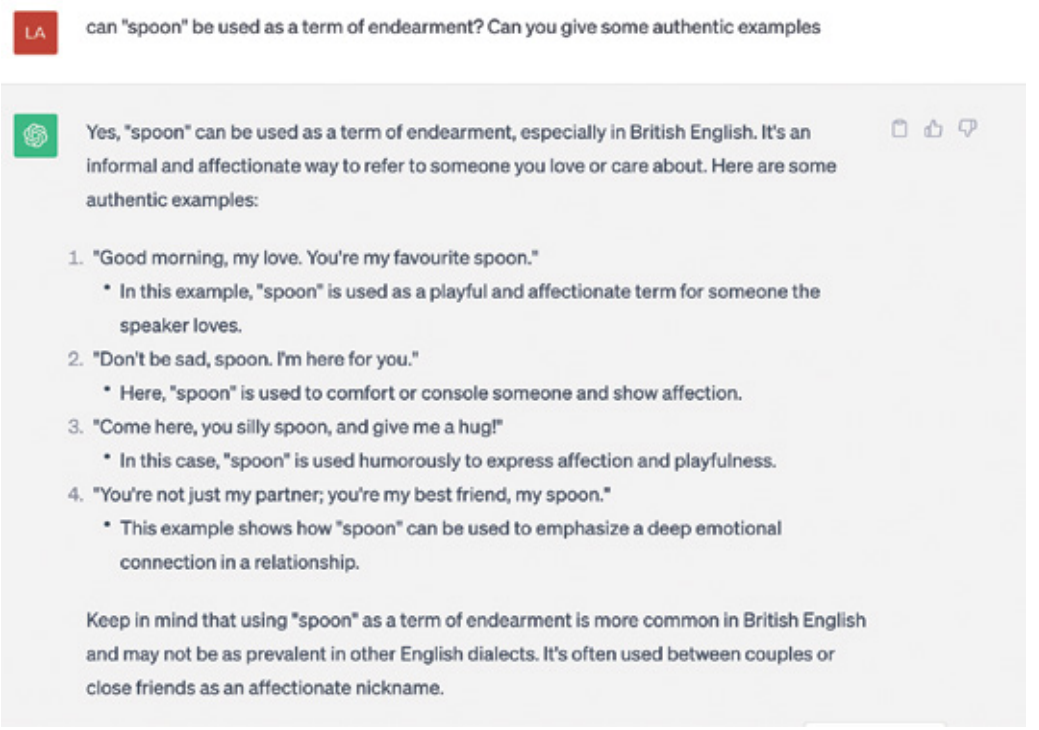

Figure 4: *clarifying question*

To be fair, I should emphasize that this example was from version 3.5. I have not encountered similar hallucinations with version 4.

4.2.3 Other issues

A related but different problem is the quality of the texts that ChatGPT has access to. The model does not distinguish between the quality of the texts it processes, as far as I know. This means that statements and views that are frequently repeated in the data may be given more weight in the answers we get from ChatGPT, regardless of their appropriateness. This can result in the inclusion of stereotypical content or fake news in the answers provided by chatbots.

Finally, it is important to note that ChatGPT is currently biased towards English. It is an American tool, and more than 90% of the training material was in English. As a result, it performs much better for English than for any other language. However, this does not mean that it cannot handle other languages. For example, my own language, Danish, has only about 4% of the training data compared to English. Although ChatGPT works for Danish and many other languages, it is not as good as it is for English. It is hard to estimate exactly how much worse it is for other languages, but it definitely hallucinates more, it is less eloquent, and it makes more mistakes.

It is also important to know that ChatGPT is culturally biased towards English. Sometimes, it is clear that a response given has been via English first. For example, when I asked

how birthdays and Christmas were celebrated in my country, ChatGPT answered that it was common to sing „Happy Birthday" to the celebrated person and that it was customary to sing Christmas carols after dinner and the sharing of gifts. However, the traditional Danish birthday song is not the one suggested, and we never sing carols after the sharing of Christmas gifts – it is always done while people dance around the Christmas tree. ChatGPT got both these things wrong. At least for the birthday song, it is obvious that this was due to interference from English. Similar examples are reported for Czech in Jakubiček & Rundell (2023) where the Czech word *state* was confused with the English word *state* when they asked for a lexical entry for that word.

5. Conclusion

It is time to summarise the experience with ChatGPT in lexicography in its first year of existence. It is evident that ChatGPT is a tremendously powerful tool. The combination of big data and strong computing power has led to a significant step forward in technology, and there is no doubt that AI tools will play a significant role in the future, in lexicography as well as in many other areas of society.

The experiments that lexicographers have done so far point in different directions. It may have to do with the fact that people have different inclinations and attitudes towards the technology involved. When you work with probabilistic models, the output is not all right or all wrong, true or false, good or bad, but something in between. Someone who is enthusiastic about the new technology is likely to be deeply impressed that the chatbot got 75% right, whereas the sceptic may be disappointed that it got 25% wrong. But that is not the whole story. The inconclusive results are also due to the fact that this technology is changing rapidly, so that we may not be measuring the same thing: the results of a study from February using ChatGPT version 3.5 are very different from those obtained from a similar study done in June using version 4, for the simple reason that the two versions are very different. Version 4 has much more processing power, it has a broader context window (25,000 words instead of 3,000 in the previous version), meaning that it has a much subtler understanding of the input, so its answers are more nuanced and it is less prone to hallucinations. We are getting new versions all the time, not only from OpenAI but also from its competitors: I have mentioned Bard, released by Google, and Grok, the X chatbot launched be Elon Musk, and there is no reason to believe that it will stop there. New bots will arrive and they will get better and better. Our results will differ depending on when we do our tests and which version we use. Finally, it must be remembered that ChatGPT is non-deterministic and that it does not make sense to compare specific responses because responses change from time to time.

Another important lesson is that you must pay attention to what you feed into the chatbot. The quality of the output depends on the quality of the input. We can improve the answer we get from ChatGPT if we make an effort to think of a smart prompt. If we want to evaluate the outcome of different studies of lexical entries generated by ChatGPT, we need to know what prompts were used in the first place. Otherwise, we may be comparing apples and pears. Learning how to formulate the prompts is important if we want good answers. This is not breaking news, but it is important to bear in mind.

It is also important to be aware of the English bias which we have seen both in terms of language and culture. Therefore, it is problematic to compare the results of studies done on English material with those from similar studies for different languages. English has a huge advantage over any other language, which makes comparison impossible. More than anything this is a political problem that must be addressed locally by the language communities collecting enough training data for their languages. In my part of the world, initiatives have been taken to develop a language model for the Scandinavian languages. But whether this will happen on a wider scale, is a different matter. It is a legitimate concern that the existing global inequality will continue to increase rather than decrease.

There are other ethical questions related to the use of AI in society, not all of which I can address here. Let me just mention one thing: training a Large Language Model and querying chatbots require large amounts of energy. One Google search requires 0.28 watt hours of energy, whereas a query to ChatGPT-4 uses 4 times that amount, or the same amount of energy as it takes to keep an old-fashioned Edison-type 60 Watt light bulb alight for 68 seconds. In a time of global climate change, this poses a challenge for decision-makers if all the daily queries are going to be handled in a sustainable way. ChatGPT had 1.5 billion number of visits in September 2023.

Finally, many people ask themselves if AI technology will outperform human expertise. Everyone is of course entitled to their own answer, and within our small field alone there is a wide range of positions. When giving his talk in Tokyo back in February, G.-M. de Schryver entitled it „The end of lexicography, welcome to the machines“ – embracing the new technology. Michael Rundell (2023:16-17) reached the opposite conclusion in his keynote presentation at Asialex in June: „we must conclude that ChatGPT does not herald ‘the end of lexicography’“.

My own position has changed somewhat during the 6 months I have used ChatGPT intensively. I have witnessed the progress made from version 3.5 to version 4, and as an auxiliary tool or sparring partner I have learned to appreciate it. It is very useful once you make an effort and phrase your prompts in an appropriate way.

On the other hand, I am very much in line with Anna Korhonen, Professor of NLP at Cambridge University and director of their AI Centre, when she says that „from a technical standpoint there is no basis for the current hype around AI outsmarting humans or taking over. We are still far from human-level AI“ (Korhonen 2023:44). She is more concerned with the political consequences that I mentioned before: „The biggest risk is not that AI will take over from the human race, but that it will perpetuate global inequalities and increase risks to privacy“ (Korhonen 2023:45). I could not have expressed this better myself. I don’t see how a chatbot could compete with a human lexicographer, and we are certainly not being outsmarted by the chatbots.

However, even if human lexicographers are still superior, this does not necessarily mean that we will continue to be in charge. Because we are not the ones who make the decision: sponsors do, sponsors in the widest sense: publishers, companies, foundations, governments. It is already possible to get more or less automated lexical products from ChatGPT. It may not have the quality of a humanly crafted dictionary, but if it gets 70 or 80 per cent right, maybe this will be enough for a sponsor. If combined with crowdsourcing, it may well be possible to get rid of the worst hallucinations. I think this development is a genuine concern, and it could be or become a threat, in particular for under-resourced languages.

References

Collins English Dictionary. Glasgow: HarperCollins. Accessed at: https://www.collinsdictionary.com [31/01/2024].

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. In *International Journal of Lexicography*, 36(4), pp. 355-387.

de Schryver, G.-M. & Joffe, D. (2023). The End of Lexicography, Welcome to the Machine: On How ChatGPT Can Already Take over All of the Dictionary Maker's Tasks. Paper presented at the 20th CODH Seminar, Tokyo. Accessed at: http://codh.rois.ac.jp/seminar/lexicography-chatgpt-20230227 [31/01/2024].

de Schryver, G.-M., Rundell, M., Tavast A., Rychlý, P., Kokol, M. (panellists) & Krek. S. (moderator) (2023). Round Table on „Large Language Models and AI in Lexicography" at the 8th Electronic Lexicography in the 21st Century Conference (eLex). Brno, Czech Republic, 27-29 June 2023. Accessed at: https://www.youtube.com/watch?v=dNkksTDYa_s [31/01/2024].

Jakubiček, M. & Rundell, M. (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? In M. Marek, M. Měchura, C. Tiberius, I. Kosem, J. Kallas and M. Jakubiček (eds.) *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*. Brno: Lexical Computing, pp. 508-523. Accessed at: https://www.youtube.com/watch?v=8e52vvDpdfQ [31/01/2024].

Korhonen, A. (2023). This idea must die: „AI will outsmart the human race". In *CAM – Cambridge Alumni Magazine*, Issue 98. pp. 44-45.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. In *Humanities and Social Sciences Communications 10*, 704. Accessed at: https://doi.org/10.1057/s41599-023-02119-6 [31/01/2024].

McKean, E. & Fitzgerald, W. (2023). The ROI of AI in Lexicography. In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: „Lexicography, Artificial Intelligence, and Dictionary Users"*. Seoul: Yonsei University, pp. 10-20.

Nichols, W. (2023). Invisible Lexicographers, AI, and the Future of the Dictionary. Paper presented at the eLex 2023 Conference: Electronic Lexicography in the 21st Century. Brno, Czech Republic, 27-29 June 2023. Accessed at: https://www.youtube.com/watch?v=xYpwftj_QQI [31/01/2024].

Oxford Dictionary of English. OUP. Accessed at: https://oxfordreference.com [31/01/2024].

Rundell, M. (2023). Automating the Creation of Dictionaries: Are We Nearly There'. In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: „Lexicography, Artificial Intelligence, and Dictionary Users"*. Seoul: Yonsei University, pp. 1-9.

Appendix

In order to get responses as consistent and comparable as possible, I have used the following prompt, following de Schryver (2023) closely:

Please give me a dictionary entry for 'x'. Use British English as your meta-language throughout. Start with the entry word, followed by frequency information and part of speech. If the word has several senses, consider if there are several words involved: homo-

graphs are words that have different etymologies or belong to different parts of speech. Is this the case, they should be treated in separate entries and marked by a superscript number. After the part of speech information comes inflectional morphology. Also include a British English pronunciation of the entry using IPA notation. Each sense should be in a numbered block. This is followed by a sense definition (maximum of 1-2 sentences), a list of synonyms and antonyms at the senses where such lists are relevant, and sense examples that illustrate both the use and the meaning of each particular sense. Other possible information elements are register markers ('slang', 'informal', 'formal', 'old-fashioned' etc.), collocations, grammatical information (e.g. valency for verbs). For the example sentences given for each sense, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you produce high-quality dictionary material. Finally, include information about etymology in the final section. If needed, include encyclopedic information, common spelling errors, auxiliary verb under the relevant sense or as a note in the end.