The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography?

Miloš Jakubíček[1,2], Michael Rundell[1]

[1]Lexical Computing, Czechia & United Kingdom
[2] Faculty of Informatics, Masaryk University, Brno, Czechia
E-mail: milos.jakubicek@sketchengine.eu, michael.rundell@gmail.com

Abstract

In this paper we present a small English dictionary consisting of 99 sample entries generated fully automatically using the ChatGPT engine. We briefly introduce ChatGPT and the underlying machinery (an autoregressive transformer-based neural network) but primarily focus on discussing the performance of the system, factors that influence the quality of the output and limitations that we have established. We show that while the system clearly represents part of the state-of-the-art of automatic generation for some entry components, it also has significant limitations which the lexicographic community should be aware of.

1. Introduction

Lexicographic tasks have been subject to automation efforts since the inception of corpus-based lexicography (see Rundell & Kilgarriff, 2011; Rundell et al., 2020). Methods and tools for automatic production of word lists, example sentences or collocations were developed, alongside of large corpora (Jakubíček et al., 2013). Those tools were typically task-specific and were applied individually or collectively to draft a complete dictionary entry and thereby streamline the process of dictionary-making. In this paper we elaborate on the use of ChatGPT, a chatbot based on a very large language model (LLM), which may be perceived as a system that – seemingly – combines all the tools so far produced for lexicography into one, and, when prompted with a simple natural language query such as "Can you give me a dictionary entry for the word table?", answers with a natural language response mimicking a typical entry structure (with arbitrary components).

In this paper we discuss the advantages and disadvantages of using such a system for lexicographic tasks, based on our observations and on an experiment we carried out on a small set of very heterogeneous English headwords. We introduce the system's principal properties and their implications as well as contemporary features that might or might not change in the near future. While our experiment was carried out for English only, we address the multilingualism of ChatGPT right at the beginning.

The purpose of this paper is in the first place educative and speculative, rather than recommending or judgmental. The evaluation we carried out was done mainly for illustrative purposes and is of very limited reproducibility. As with any new technology, or rather in this case, an emerging technology (large language models) used in a new context (lexicography), it is of the utmost importance for the lexicographic community to be aware of all the issues around LLMs, the principal caveats and practical questions to ask, before any decision to apply the technology in their work.

2. ChatGPT and OpenAI's GPT-based models

While we will not attempt a technical description of the system from the NLP point of view, it is necessary to introduce it at a broad level to be able to discuss some of its properties. ChatGPT (Ouyang et al., 2022) is a chatbot based on the GPT-3 language model (Brown et al., 2020) launched by OpenAI in November 2022. GPT stands for Generative Pre-trained Transformer, a type of neural network that is trained on a large unannotated corpus (i.e. plain text), yielding a language model, i.e. a probabilistic distribution over words given prior words. Such a model makes it possible to carry out what is formally called decoding or inference, and in practical terms generates the most likely word sequence given a prompt.

The level of details we can give on how exactly the model was trained and how exactly the inference works is limited. ChatGPT is a closed-source proprietary product of OpenAI, a Microsoft-co-owned company[1]. The aforementioned academic publications discuss many aspects of transformer-based neural network training and usage, yet it is unclear to what extent they describe the actual product. This uncertainty extends to the training corpus data. To understand its level, it is just enough to read page 12 of the very comprehensive report on GPT training data provided by Thompson in March 2022 (Thompson, 2022). All we know is that it was trained on a filtered version of the Common Crawl[2], two unspecified book corpora, one unspecified web corpus and Wikipedia making about 500 billion tokens all together. Unlike the model traditionally used in corpus linguistics, tokens follow the so called subword tokenization – one word typically consists of multiple tokens (or rather, multiple characters form a single token) which – among other benefits – makes it possible for the model to handle morphology. Compared to a corpus-linguistic approach to tokenization, which for English boils down to white-space tokenization where 500 billion tokens would amount to some 450 billion word forms (arbitrarily defined, of course), the subword tokenization approach entails a much smaller word set – the authors estimate two or three times smaller, of the order of "only" about 150–200 billion words.

The tokenizer is one of the very few open components of the system and is available as an online tool[3] as well as a Python package. Different models use slightly different tokenizers though (so that, the GPT-4 tokenizer is not the same as the one used in GPT-3) and OpenAI currently claims that 1 token approximates to 4 characters in English[4]. Perhaps more explanatory is the overall size of the vocabulary used by the various GPT models, i.e. number of different tokens, which was only 50,257 for both GPT-2 and GPT-3.

Both training and use of such a model is extremely hardware intensive. It is not easy to quantify this accurately, but it is safe to say that both are far from what is affordable as commodity hardware (broadly defined as any platform costing less than $100,000), the training being of course more demanding than the inference which is more influenced by the particular workload scenario (number of users, size of the prompts and responses etc.).

Multiple new versions of ChatGPT have been released based on the same language model (GPT-3). First, an updated model (GPT-3.5) appeared, and then in March 2023 a new language model (and new version of ChatGPT using it) labeled GPT-4 was released

[1] See https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/
[2] https://commoncrawl.org/
[3] At https://platform.openai.com/tokenizer
[4] https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

too. While OpenAI provided an API to the actual GPT models, there was no API for ChatGPT initially. Our first experiments therefore used a very tedious workaround based on simulating browser access[5] and were later replicated using the official API for ChatGPT based on the GPT-3.5 model. The newly released GPT-4-based ChatGPT is not yet generally available through the API and is extremely rate-limited through browser access[6], thus it was not used for generating the dictionary described in this paper.

```
Certainly! Here is a dictionary entry for the word "table":

Noun:
1. A piece of furniture with a flat top and one or more legs, used as a surface for working at,
   eating from, or on which to place things.
2. A set of facts or figures systematically displayed, especially in columns.
3. A list of contents or topics covered in a publication or document.
4. A group of people sitting at a table for a meal or meeting.
5. A level area, typically at the top or bottom of a slope.

Verb:
1. To postpone consideration of (a motion, bill, or other matter).
2. To enter (a proposition) in a table or list.
3. To present formally for discussion or consideration at a meeting.
4. (British) To offer (someone) a meal or entertainment.

Example sentences:
- I'm going to set the dishes on the table.
- The data is presented in a table for easy comparison.
- The table of contents lists all the chapters in the book.
- We sat around the table and discussed our plans for the future.
- We need to table this motion until we have more information.
- Let's table that idea and come back to it later.
- I'd like to table a motion for the next meeting.
- We'd be happy to table you for dinner tonight.
```

Figure 1: Sample output of GPT-3.5 for the prompt “Can you give me a dictionary entry for the word table?”

Sample output for the prompt “Can you give me a dictionary entry for the word table?” generated by ChatGPT based on GPT3.5 and GPT4 is given in Figures 1 and 2. It is important to realize that because of the non-deterministic nature of the system (which means that the same question may generate a different answer), no reliable conclusions can be drawn from the comparison. But regardless of these uncertainties and of ongoing product development (which invalidates findings particular to certain versions of the system), we briefly enumerate here what appear to be durable properties shared across all current ChatGPT versions:

1. **language modelling – not language reasoning**
   Perhaps the most important point to recognise about how large language models work is that the actual usage (inference) of the model strictly follows the probabilistic nature of the system as learned from the training data. The model does not perform any *logical* inference in terms of formal reasoning, nor does it build any kind of

[5] Using https://github.com/mmabrouk/chatgpt-wrapper
[6] By the time of writing this paper this was 25 messages every 3 hours only.

```
table (noun)

1. A piece of furniture consisting of a flat, horizontal surface supported by one or more legs,
   used for eating, writing, working, or displaying objects.
   Example: We sat down at the table to have dinner together.

2. A systematic arrangement of data or information, usually in rows and columns, that presents
   facts or figures for easy reference and analysis.
   Example: The table in the report shows the monthly sales figures for the past year.

3. A list or register of topics or items, especially in a book or document,
   for reference or information.
   Example: The periodic table organizes chemical elements based on their atomic number
   and properties.

4. In computing, a structured set of data elements, often consisting of rows and columns,
   used to store and manage information in a database.
   Example: The customer table in the database contains information about all the clients.

table (verb)

1. To present (a proposal, resolution, or other matter) for consideration or discussion.
   Example: The senator tabled a bill to increase funding for public schools.

2. In British English, to postpone or suspend the consideration of (a matter) for a later time.
   Example: The committee decided to table the discussion until next month's meeting.
```

Figure 2: Sample output of GPT-4 for the prompt "Can you give me a dictionary entry for the word table?"

knowledge base of facts. Many recent controversies[7] around the GPT models (both in and outside of the academic community) result from ignoring, misunderstanding or simply not being aware of this basic principle.
In the context of generating dictionary entries, it is important to emphasize that the overall structure of the entries is also completely learned from the training data. There is no explicit information the system can use to determine which entry components to generate, how to typeset an entry, how to visualize homographs or polysemous entries or that they should be presented in a numbered list. All of that comes through seeing existing entries of existing dictionaries that were part of the training data.

2. **non-deterministic learning and inference**
As with many other neural networks, training of the transformer language model is non-deterministic, mostly because some model parameters are initialized at random. This means that repeated training on the same training data creates a (possibly substantially) different model.
Moreover, the inference carried out by ChatGPT through the GPT models is by default non-deterministic too, i.e. it yields different answers for repeated prompts. This results from the fact that finding the optimal answer for a given prompt (or, in other words, the most probable sequence in the model) is not tractable in a model of this size. Different inference heuristics are being applied[8] to mitigate this issue, and it is not absolutely clear which one is used by ChatGPT[9]. In the API,

[7] Such as https://www.bbc.com/news/technology-65202597
[8] See https://huggingface.co/blog/how-to-generate for a very reader-friendly introduction to this topic.
[9] although based on the API parameters it is likely a variant of nucleus sampling

the so called temperature parameter may be used to tune the greediness of the inference, and by setting it to 0 one gets deterministic outputs – at the cost of a (possibly substantially) worse output quality, obviously, because a greedy search is rarely the optimal one.

3. **static model**
   Once the model is trained, it is static and it is in principle not possible to make any incremental updates easily. This is why the system frequently informs users that the training data are not newer than 2021. Including newer data means retraining the model completely (which in principle always has unforeseen implications). ChatGPT plugins introduced recently do not change the model in any way, but are rather used as part of the prompting mechanism.[10]
4. **no source data reference**
   In its current forms, the model does not keep references to training sources and it is not straightforward to implement this[11]. Obviously, the model is unable to "cite" whole sentences verbatim as this is not the way it operates. If the output contains whole sentences that occur in the training data, it is a random (and unintentional) artifact of the training process.
5. **limited prompt and response length**
   ChatGPT's input (prompt) and output (response) length is limited because the model inference is hardware-consuming too. Depending on the model, it varies between 2,048 and 4,096 tokens (so, approximately 500 and 1,000 words). For many applications this is not a problem, but generating long responses that might sometimes be needed in lexicography could be affected.
6. **translation through multi-lingualism**
   The vast majority (92%) of GPT-3 training data is English only.[12] The ability to carry out high-quality translations without parallel data (which are hard to obtain for many low resource languages, with no prospect of this changing in the near future) is certainly one of the biggest achievements of the transformer architecture but it also has certain disadvantages, particularly for lexicographic use cases. Translation using ChatGPT also works through language modelling (like anything else) and is achieved by being able to match language-specific contexts cross-lingually. This means that prompts formulated in other languages than English are answered using also (and mainly) English data.
   This gets particularly manifested when prompting the system with a cross-lingually polysemous word. To give a simple example, we asked ChatGPT (based on GPT-3.5 and GPT4) in Czech to generate a dictionary entry the Czech noun *stát* which means *state* in the sense of a *country*. The result is given in Figures 3 and 4. Both versions responded with an entry that includes a non-existing sense coming from the translation into the English *state* and one of its English (but not Czech) senses being *status*. Moreover, the example sentence in Czech says *Our company is in a good financial status and is not in an endangered country.*, demonstrating that another translation error occurred in the other direction by translating the English *status* back as a *country*.

[10] https://openai.com/blog/chatgpt-plugins
[11] Although there are similar attempts, see https://www.longshot.ai/features/longshot-fact-gpt or perplexity.ai
[12] An overview by OpenAI is provided at https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_character_count.csv

```
Určitě! Zde je slovníkové heslo pro podstatné jméno "stát":

stát (mužský rod)

Významy:
1. Politická a teritoriální jednotka řízená vládou a tvořící svrchovanou moc na určitém území.
2. Země, území, region, oblast nebo teritorium.
3. Mocnost, suverénní autorita nebo řídící orgán, který má kontrolu nad určitým územím
   nebo skupinou lidí.
4. Stav nebo podmínka, ve které se něco nachází.

Příklady:
- Česká republika je státem ve střední Evropě.
- Ve státě Texas v USA je mnoho ropných vrtných plošin.
- V takovémto státě by nebyla svoboda projevu.
- Naše firma je v dobrém finančním stavu a není v ohrožení státu.

Poznámka: Slovníkové heslo může být upraveno a doplněno v závislosti na slovním spojení
nebo kontextu, ve kterém se slovo vyskytuje.
```

Figure 3: Sample output of GPT-3.5-based ChatGPT for the prompt "Můžeš mi vygenerovat slovníkové heslo pro podstatné jméno stát?" (Can you give me a dictionary entry for the noun country?). The erroneous sense and example sentences are the last (fourth) ones.

```
Stát (podstatné jméno, rod mužský)

1. Územně a politicky organizovaný celek, který má svou vládu, území, obyvatelstvo a zákony;
   země, republika, království.
   Příklad: Česká republika je stát ve střední Evropě.

2. Instituce a orgány, které vykonávají moc a správu v daném území; veřejná moc, vláda.
   Příklad: Stát zajišťuje základní veřejné služby pro občany.

3. Obecně, stav nebo poloha, ve kterém se něco nebo někdo nachází; kondice, situace.
   Příklad: Přítel mi pověděl o svém zdravotním stavu.

Synonyma: země, republika, království, veřejná moc, vláda, území, situace, kondice

Související pojmy: vláda, ústava, politika, zákony, suverenita, území, obyvatelstvo,
                   státní správa
```

Figure 4: Sample output of GPT-4-based ChatGPT for the prompt "Můžeš mi vygenerovat slovníkové heslo pro podstatné jméno stát?" (Can you give me a dictionary entry for the noun country?). The erroneous sense and example sentences are the last (third) ones. Correspondingly the list of synonyms contains erroneous items too (*situace*, *kondice*).

7. **prompting** The very nature of the system consists of natural language prompts and natural language responses. While this is a big advantage for everyday usage, it represents an obstacle for more formal approaches, as the optimal design of the prompts becomes crucial for high quality output. Moreover the best prompt design for a particular task may change unnoticed between versions or languages. All we know is that finding the 'right' prompt is critical to getting the best response, but there is no reliable way of knowing how to find those best prompts. On top of this, the non-deterministic nature makes it hard to evaluate even a single prompt.

3. Making an English dictionary using ChatGPT

We generated two monolingual English mini-dictionaries: one using the January 9, 2023 version of ChatGPT (based on the GPT3.5 model) by the time of submitting the extended abstract of this paper; and one using the March 23rd, 2023 version of ChatGPT (based on the same model) by the time we were preparing the full paper. Both dictionaries are publicly available with the Lexonomy platform (Měchura et al., 2017) at https://lexonomy.eu/chatgpt and https://www.lexonomy.eu/chatgpt35. The former was, for reasons explained earlier, done by simulating browser access in the user interface, the latter through the official API that became available meanwhile.

The entries of these two dictionaries were generated for 99 English single- and multi-word headwords which are listed in full as Appendix A. Because the limited availability of the system made it impossible to create a bigger dictionary sample while preparing this paper, we wanted the dataset to be very diverse and therefore adapted a sample headword list used in the preparation of the DANTE lexical database for English (Convery et al., 2010).

The sample covers words of varying complexity and several parts-of-speech, as well as some multi-word expressions. We presented ChatGPT with each headword with no additional information (such as part-of-speech) and collected the response. Because the system is fine-tuned as a chatbot, we asked the following three questions for each headword H:

1. What does the word H mean?
2. Generate a dictionary entry for H.
3. Generate a dictionary entry for H including possible word forms, word senses, pronunciation, collocations, synonyms, antonyms and examples of usage.

These three questions were asked in this particular order in one conversation. As the inference of the system is generally not deterministic, we repeated this whole conversation three times independently in a new ChatGPT context, so that there would be no influence between the three runs. Altogether we thus obtained 297 entries consisting of verbatim answers to the three questions composing each conversation. In Lexonomy, entry names bear the `.a1`, `.a2` and `.a3` suffixes for the first, second and third run, respectively.

4. Investigating the mini-dictionary: a lexicographic evaluation

The simplest way to evaluate ChatGPT's responses in this task is to see how well it handles each of the principal components in a dictionary entry. We will therefore consider its performance across the following elements: word-sense disambiguation, definitions, grammatical information, 'marked' items (such as words which are formal, archaic, or offensive), and example sentences. In each case, we compare ChatGPT's output with equivalent entries in two high-quality 'human-produced' dictionaries: the *Oxford Dictionary of English* (ODE), which is now the default source for a Google search on the lines of 'define X'; and the *Macmillan English Dictionary* (MED). We refer to these as our 'reference dictionaries'.

4.1 word-senses

The challenges here are well known. Establishing a set of word senses for a given headword is generally considered the hardest task in lexicography – not least because meaning is so contextually-determined that 'it makes sense to ask whether words do in fact have meaning at all' (Hanks, 2013: p. 65). The discrete numbered senses in dictionaries are in reality a lexicographic construct, and in many cases no two dictionaries will present the same inventory of senses for a polysemous word. Nevertheless, within this conventional paradigm, we can still judge whether a given dictionary's analysis of a word's meanings is a fair – and practically useful – reflection of the way the word is used in real communicative situations.

Even allowing for the inherent difficulty of the task, ChatGPT does not perform well in this area. Furthermore, our sample did not include any headwords of great complexity (words with, say, six or more senses in a traditional dictionary), so we can assume that – in its current form, at least – it would be defeated by any highly polysemous headword.

A recurring problem is what we might call 'false polysemy', where the system enumerates multiple senses, with different definitions, in cases where there is really only one. A standout example is its treatment of the word *climate*. In both our reference dictionaries, *climate* has two main senses: the weather-related one, and a metaphorical use encoded in expressions like 'in the current economic climate' or 'a climate of fear'. ChatGPT (in this case response `a1`) gives the following senses:

1. The long-term patterns of temperature, humidity, wind, and precipitation in a particular region.
2. The overall weather conditions of a place over a period of time, typically 30 years or more.
3. The typical or average weather conditions of a place.
4. The general set of weather conditions of a planet or region.
5. The state of the atmosphere in a region in terms of temperature, humidity, wind patterns and precipitation.
6. The average of weather conditions over a period of time, typically 30 years or more.

This goes far beyond what is known in the lexicographic trade as 'splitting' (as opposed to 'lumping'): there is essentially just a single meaning here, explained in six different ways. What is worse, the system fails to take account of the second, very common, metaphorical use identified in our reference dictionaries.

While this is the most egregious instance of false polysemy, there are very few cases in the sample where the system performs adequately (*empty* is probably the best of a bad lot).

At *command*, we find a similar tendency to split one sense unnecessarily while completely missing another common meaning. The IT-related noun use ('the "insert block" command is executed') is correctly identified, but the system posits an equivalent verb use (with the implausible example 'To shut down the computer, you need to command it to shut down'), for which there is little evidence. In response `a3`, the core sense of 'giving an order' is needlessly split to cover the case of pets: 'To control or direct a pet, animal, or machine through the use of specific commands.' At the same time, other frequent usages are overlooked, with nothing to account for sentences such as 'truffles command a high price' or 'the fort commands a panoramic view of the coast' – all well covered in our two reference dictionaries.

Even simple concrete nouns do not escape these problems, with the word *potato* given no fewer than five 'senses' in response `a2`:

1. (Botany) A starchy, tuberous crop from the perennial nightshade Solanum tuberosum, native to the Andes in South America.
2. (Food) A staple food in many parts of the world, often boiled, baked, or fried.
3. (Industry) Used in the production of various food products, such as potato chips and French fries.
4. (Alcohol) Also used as an ingredient in the production of alcohol, such as vodka.
5. (Variety) Can come in various varieties with different colors, shapes, and textures.

These are all legitimate things to say about potatoes and their use, but this treatment suggests that the system does not really understand what humans mean by a 'dictionary word sense'.

Identifying word senses is rarely straightforward, but when even a simple word like *ameliorate* ends up with three senses, it is clear that ChatGPT is not up to the task.

4.2 definitions

Here the news is more promising, and definitions are in general one of ChatGPT's stronger points. Definitions such as 'An order or instruction given by a person in authority' (*command*, noun use), or 'Capable of producing desired results with a minimum of effort or energy' (*efficient*) give the right information in an accessible form, and compare favourably with those in the reference dictionaries. Some say too much and end up being longer than is desirable: the entry for *bargain* (response `a1`) includes 'an agreement between two or more parties in which each party agrees to certain terms, often used to refer to a transaction where goods or services are exchanged for an agreed-upon price that is typically lower than the market value.' A tweak to the prompt question might resolve this, specifying a maximum word count (as some dictionary styleguides do).

Other definitions employ familiar lexicographic formulae: *closure* (response `a3`) has 'The act or process of closing or the state of being closed', and one version of *slavish* (the others are better) includes 'Resembling or characteristic of a slave'. Styles like this, which are unhelpful for users, were widespread in older publications but are less often found in good contemporary dictionaries. Occasionally a definition will fail to include a key meaning component: thus *garden* ('A piece of land used for growing plants, flowers, or vegetables.')

does not mention that gardens are typically attached to houses; similarly, one version of *beach* describes it as 'a place of recreation or relaxation, where people go to swim, sunbathe, and engage in other outdoor activities', without noting its adjacency to the sea or a lake. But there is plenty that is positive. The system seems to perform especially well when defining technical terms. All versions of *carbon cycle*, for example, are well (and clearly) defined (if sometimes over-long), and duly mention the key related terms photosynthesis and respiration. This is important because, of all the components in a dictionary entry, definitions have so far proved the least tractable in terms of automation. ChatGPT may be at least part of the answer.

4.3 grammatical information

In other experiments we have specifically prompted ChatGPT to identify the syntax patterns that typically follow a given word – in the way that pedagogical dictionaries usually do. (Results have been patchy.) This was not done in the case of the mini-dictionary, so our focus here is on the way grammatical features are dealt with at a general level. Transitivity is not always handled well. Thus the entry for *empty* (verb) fails to cover intransitive uses like this (from MED): 'the stadium began to empty'.

More worryingly, some words are wrongly categorised in terms of word class. In one version of *aside* (`a3`), a sense which is explained as 'to one side: He pushed the plate aside' is labelled as a preposition. In other cases, the form of a definition does not match the word class, as in sense 2 of the verb *haunt* (response `a3`), defined as if it was both an adjective and a noun:

1. Visit frequently, or reside in as a ghost or spirit.
2. Constantly present in one's mind; an obsession.
3. To frequent a place or places frequently.

Problems like these are pervasive, and significantly compromise the value of ChatGPT's output.

4.4 'marked' items

Most lexical items are 'unmarked', but some are specialised in terms of their distribution across text-types. Dictionaries typically use 'labels' (such as *formal*, *offensive*, or *old-fashioned*) to draw users' attention to these features, though other strategies are sometimes employed too. Some of the words in our sample list were specifically included in order to see how well ChatGPT coped with this aspect of language.

In general, the system performed well on this topic. Its response to the word *half-caste* (once a common word for a person of mixed race, but now universally regarded as offensive) was exemplary. In its response, the explanation of meaning was preceded by a warning that it is 'considered to be a derogatory term used to describe a person of mixed racial heritage'. And this definition is followed by further advice: 'It is now considered offensive and outdated and it is better to use terms such as "mixed race" or "multiracial" instead.' It would be difficult to improve on this. Similarly, *betimes* was correctly identified as an archaic word whose 'usage is rare in modern English'. Unsurprisingly, it failed to recognise *bockety*, an Irish-English word meaning unstable or rickety. Though this does appear in ODE (but not MED) its frequency in a general English corpus is very low. Its response to *ameliorate* was somewhat disappointing. European cognates of this word (such as French *améliorer*) are typically unmarked, but in English it is a rare and rather formal word, and is marked as such in MED. However, it carries no label in ODE or Merriam-Webster, so it would be unfair to criticise ChatGPT for this omission.

4.5 example sentences

As prompted, all of our sample entries included at least one example sentence for every word and sense covered. An unexpected feature of these examples – given that the system is based on such a large corpus – is that they often look as if they have been made up by a rather unimaginative human editor. A persistent and very noticeable issue, identified in every experiment we have made with ChatGPT – is that examples predominantly follow the formula '3rd person subject with simple past verb', typically opening with a definite article. One of the entries for *aside* (`a2`) ends with this example set:

- "She put aside her book and listened to the music."
- "The judge set aside the verdict and ordered a new trial."
- "He whispered something aside to his friend before he began to speak."
- "The actor broke character for a moment and delivered an aside to the audience."
- "The singer added an aside to the melody, making the song more interesting."
- "The author inserted an aside in the text to comment on the society of his time."

In pre-corpus times, this pattern was a reliable predictor of an invented example – to the point that lexicographers working on the MED were explicitly warned to avoid using this formula in examples, unless corpus data showed the pattern to be typical of a word's behaviour. ChatGPT's examples are for the most part unconvincing, and when there is a set of examples, they exhibit far too little diversity in terms of structures and even subject matter. (This is something that skilled lexicographers pay a lot of attention to.) One of the worst instances (at *command*), 'The commander commanded his troops to march forward', looks like something invented by a not very good apprentice lexicographer without access to a corpus. In the current state-of-the-art, lexicographers are offered candidate examples filtered by the GDEX software (Kilgarriff et al., 2008), and in most cases it is easy to find a suitable example, which can either be used as is or with minimal tweaking. At the moment, there is probably more mileage in further refining the GDEX algorithm than in trying to get ChatGPT to produce more natural-sounding examples.

5. Conclusions

5.1 Arguments in favour

Some of the arguments in favour of using ChatGPT in a lexicographic context are very straightforward – ChatGPT is:

- extremely simple to use,

- very likely going to be continuously developed and improved,
- very affordable as for pricing[13],
- naturally multilingual,
- equipped with an API that is easy to integrate into a dictionary writing system.

Moreover the chat environment allows the lexicographer to challenge the response which gets frequently improved in the second round.

But there are also less obvious advantages of ChatGPT. When it comes to generating particular entry components, the system is often on a par with state-of-the-art solutions (though this is very hard to evaluate of course) for some of them, but there are also some where it likely exceeds the state-of-the-art. In our experiments this was most striking on the task of differentiation between synonyms and antonyms. There are many solutions for generating a distributional thesaurus, but discriminating between synonyms and antonyms (and possibly other relations) still remains a challenge. Initial observations however indicate a very low error rate for this task: asking *Can you classify the following list of words into synonyms and antonyms: word1, word2, ...* generates decent (though not perfect) responses (an example of such a prompt is given in Figure 5). This is not surprising though: modelling contexts of words is the very job of a language model!

Similarly, generating dictionary definitions is something that ChatGPT seems to be good at, even though their mapping to a particular sense distribution is problematic (mainly subject to repetitions). A problematic part of the generated definitions might be the inspiration by old dictionaries (using "the act of" definition style, or incorrectly referring to particular dictionaries such as "according to Merriam-Webster X is defined as" and then using an arbitrary generated definition), but this could likely be avoided with better prompts that would instruct the system to avoid such styles.

```
PROMPT:
The following is a list of similar words for the word "fast": slow, quick, efficient, powerful,
expensive, cheap, reliable, simple, complex, small. Can you classify this list into
synonyms and antonyms?

RESPONSE:
Sure! I can classify the list into synonyms and antonyms for the word "fast":

Synonyms: quick, efficient, powerful

Antonyms: slow, simple, complex, small

The remaining words (expensive, cheap, reliable) don't have a direct relationship
with the word "fast," so I haven't classified them as either synonyms or antonyms.
```

Figure 5: Sample conversation with GPT-4-based ChatGPT targeted at synonym/antonym classification.

5.2 Arguments against

Reasons not to use ChatGPT are perhaps less obvious but no less important for that, since some of them are quite fundamental.

[13] See https://openai.com/pricing

1. **Googleology is a bad science. And GPTology too.**
   We borrow the title of Kilgarriff's paper (Kilgarriff, 2007) where he argued against using Google search as a corpus search system. Many of the then-used arguments are valid now as well. ChatGPT is using unknown data sources, with non-deterministic (and very likely soon-to-be-personalized) responses, very limited stability and reproducibility. Using it as a general purpose search system in a scientific context inevitably suffers from all the issues a Google search-based approach does.
2. **Vicious data circle**
   We explained that GPT knows what an entry looks like from existing dictionaries online that formed part of the training data. This represents a challenge: in all likelihood, it is not the best and most up-to-date dictionaries which were freely available for mass download (though CommonCrawl or similar) and which the system learned from. It is notably easy to trigger the kinds of 'lexicographese' ('the act or state of X', 'characterized by Y', etc.) which were once pervasive in dictionaries but are now (thankfully) being abandoned.
   Lexicography has undergone some radical changes in the past 20 years: the arrival of big corpora, NLP analytics, the migration from print to digital dictionaries. All of these have had massive implications on the way lexicographers work and on the range and quality of information that has been uncovered. And these developments are ongoing. Using a system whose training data often pre-dates those changes is somewhat problematic from this point of view.
3. **Evidence generating or evidence observing?**
   Last but not least, a dictionary-making process which relies entirely on the use of tools like GPT implies the abandonment of the lexicographer's current role of scrutinising and verifying the evidence suggested by an analytic system. Most NLP tools for lexicography interrogate a corpus, perform some (often very complex) analysis but track back to corpus evidence in the form of concordance lines, so that the lexicographer can determine whether the automatic results match what is in the corpus (and check the corpus content, metadata, annotation etc.). In the present state-of-the-art, we see this stage as an essential part of the process, and we have significant misgivings about the removal of human actors from the data generation chain. ChatGPT and GPT-like models do not make back-linking evidence possible at the moment, and it is questionable whether this would ever be possible.
   This issue also relates to the whole notion of corpus-driven lexicography. In the case of dictionary examples, for instance, it is generally accepted that they should reflect what the data shows us to be the contexts and patterns in which a word most typically occurs: examples shouldn't be made up, but should be found in the corpus (and shortened or lightly edited if needed). Example sentences generated by ChatGPT cannot be found anywhere. There is no guarantee that they were ever produced by a human writer or speaker, nor (as we have seen above) that their typicality matches what lexicographers would choose.

5.3 Summary

The introduction of ChatGPT has gained huge attention worldwide, often generating excitable or hyperbolic reactions, both positive and negative (see e.g. Beckett, 2023). This paper attempts a more sober-minded evaluation of the potential of this emerging technology, and is cautious about claims that ChatGPT can – to paraphrase a recent talk by de Schryver – handle (almost) all of the lexicographer's tasks (or make us believe it can), with successful results [14].

Our various experiments with ChatGPT (notably but not only with the mini-dictionary described in this paper) have convinced us that it cannot (yet) replace the involvement of lexicographers in the dictionary-making process, and moreover that for some of the requisite tasks (such as sense discrimination and example-writing) its performance is significantly worse than what established technologies can do.

But this certainly does not mean that lexicographers should ignore ChatGPT. For over two decades, we have been adapting lexicographic workflows to emerging technology trends, always with the goal of producing better dictionaries at a lower cost in time and resources. We now need to consider what ChatGPT can contribute to these goals, taking account of the caveats raised in this paper but also of its positive outcomes in some areas. ChatGPT is a general purpose solution and we argue that lexicography needs custom solutions (e.g. through fine tuning of these large language models for particular lexicographic tasks) to mitigate some of the issues discussed in this paper. What these custom solutions may learn from GPT models are all the relevant technological lessons, such as successful application of neural networks as a machine learning computational model and the absolutely crucial role of big datasets. GPT models at the moment represent a highlight of a trend (which has been developing for at least a decade) of using large unannotated datasets for machine-learning purposes. It is up to anyone working in computational lexicography to follow up on this with practical solutions which do not compromise on fundamental principles (above all the idea of a data-driven approach) which have been established over time, since large corpora first became available. And this needs to happen in a workflow model, such as post-editing lexicography, that does not leave the lexicographers sand-blinded as ChatGPT does.

6. References

Beckett, C. (2023). GPT-4 has brought a storm of hype and fright – is it marketing froth, or is this a revolution? *The Guardian*. URL https://www.theguardian.com/commentisfree/2023/mar/17/gpt-4-ai-tools-fashion-architecture.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp. 1877–1901.

Convery, C., Mianáin, P., Raghallaigh, M.Ó., Atkins, S., Kilgarriff, A. & Rundell, M. (2010). The DANTE Database (Database of ANalysed Texts of English). In *Proceedings of the XIV EURALEX International Conference*. pp. 293–5.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. The MIT Press. URL https://doi.org/10.7551/mitpress/9780262018579.001.0001.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference CL2013*. Lancaster University, pp. 125–127.

Kilgarriff, A. (2007). Last words: Googleology is bad science. *Computational linguistics*, 33(1), pp. 147–151.

[14] See https://www.youtube.com/watch?v=mEorw0yefAs&list=PLXmFdQASofcdnRRs0PM1kCzpuoyRTFLmm&index=5&t=566s

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1. Universitat Pompeu Fabra Barcelona, pp. 425–432.

Měchura, M.B. et al. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*. pp. 19–21.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, pp. 27730–27744.

Rundell, M., Jakubíček, M. & Kovář, V. (2020). Technology and English Dictionaries.

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end. *A Taste for Corpora. In Honour of Sylviane Granger*, pp. 257–282.

Thompson, A.D. (2022). What's in my ai. *A comprehensive analysis of datasets used to train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher. LifeArchitect. ai Report.* URL https://lifearchitect.ai/whats-in-my-ai/.

A. DANTE sample headword list

command
echo
empty
haunt
leaf
stomach
Amazon
Amazonian
beach
DJ
echoing
efficient
emptiness
empty-handed
grave
grave
gravely
haunted
haunting
hauntingly
leafy
bargain
butter
camp
camper
camping
slave
slavery
slavish
slavishly
spite
spiteful
spitefully
ameliorate
asleep
azure
Belleek
betimes
bockety
Canada
Canada goose
carbon
carbon cycle
climate
climate change
climate control
cookie
couch potato
DNA
fart
half-caste
Leaving Certificate
moralize
moralizing
ouch
slag
snowboarding
wed
Wed.
wireless
also
always
anyhow
anyway
busy
careful
closure
garden
mackerel
potato
Protestant
suitable
wake
how are you
after
however
might
this
altogether
aside
hereinafter
might
moreover
notwithstanding
nowhere
provided
towards
somewhat
AIDS
anti
can't
chug
-een
gutter
gutters
Shaw
Shavian
meander
speck
swathe
Czech