

The Road towards Fine-Tuned LLMs for Lexicography

Keywords LLMs; ChatGPT; out-of-the-box GPT; custom GPT; fine-tuned GPT; English; Portuguese; Bantu; Xhosa; Zulu; Lusoga; Luganda

Ever since the first LLM has been put into the hands of the general public, by means of the release of a nifty interface to ‘chat’ with it, the world has been in a frenzy. We are referring, of course, to the release of ChatGPT in November 2022. Already in February 2023, during a live demo from Tokyo, the first uses in the field of lexicography were extolled. Convinced that the future had arrived, AI functionality was added to the dictionary writing system TLex using OpenAI’s GPT-3 API with which a public online dictionary was updated in real time with new entries generated by the AI a few minutes prior (de Schryver & Joffe, 2023). In the months that followed, colleagues on all continents joined the fray, at the conferences of the DSNA (Barrett, 2023), ASIALEX (McKean & Fitzgerald, 2023; Rundell, 2023), and eLex (de Schryver et al., 2023; Jakubíček & Rundell, 2023; Nichols, 2023; Tran et al., 2023), followed by the first publications in professional journals (de Schryver, 2023b; Lew, 2023).

By and large, the most successful experiments were reported for applications using English,¹ with critical notes regarding the usefulness for other languages mostly missing from the conversation. A notable exception is Jakubíček and Rundell (2023, pp. 522-523) who briefly looked at Czech (11 million speakers) and derided the use of ChatGPT for it. The results of the first in-depth and detailed attempt to study the lexicographic potential of using ChatGPT for a language other than English, were presented at the inaugural Americalex-S conference in São Paulo in October 2023 (de Schryver, 2023a). Although the assumption was that ChatGPT would be able to do an excellent job for Portuguese (260 million speakers), it was shown that it got things very wrong for the compilation of both monolingual Portuguese as well as bilingual Portuguese-English dictionaries. By way of illustration, Figure 1 displays one of the slides from that talk. While this Portuguese material may at first appear rather convincing (it fooled two professors!), it is riddled with errors. The reason? One is actually looking at English lexical analysis through the eyes of Portuguese. Reformulated, everything one sees there was created internally *in* English, *for* and *about* English, and was merely translated back into Portuguese (the prompt was in Portuguese). If a further analogy is allowed: Given the overwhelming amount of English in an LLM, the model is simply incapable to ‘think’ in any other language, let alone analyse it linguistically.

¹ See for instance the supplementary material for a COBUILD-style monolingual English dictionary compiled by ChatGPT here <https://academic.oup.com/ijl/article/36/4/355/7288213#supplementary-data>.

3. Verbo

- **Formas Morfológicas:** banco (3ª pessoa do singular do presente), bancou (pretérito), bancando (particípio presente), banco (3ª pessoa do singular do presente)
- **Sentido:** Depositar ou armazenar algo em um banco, ou confiar ou armazenar algo para uso futuro.
- **Padrão de Uso:** Alguém **bancou** algo em um banco, ou eles **bancam** algo para um propósito específico.
- **Definição:** Quando alguém **deposita** algo em um banco, está armazenando em uma instituição financeira ou **confiando** em algo **para uso futuro**.
- Equivalentes de Tradução em Inglês:** Deposit, Rely on
- **Exemplo 1 em Português:** Ela **deposita** suas **economias** em um banco local.
- **Tradução 1 em Inglês:** She banks her savings in a local bank.
- **Exemplo 2 em Português:** Ele está **contando com** sua vitória na competição com suas habilidades.
- **Tradução 2 em Inglês:** He's banking on winning the competition with his skills.
- **Pronúncia:** ['bẽku]
- **Sinônimos:** depositar, confiar
- **Antônimos:** sacar, desacreditar

Fig. 1: The 3rd sense for the entry **banco** in a COBUILD-style bilingual Portuguese-English dictionary as proffered by ChatGPT-3.5. This sense, 'to bank money', simply does not exist in Portuguese, but ChatGPT not only 'created' this sense, it went as far as inventing conjugations for it to then use them as well (**bancou*, **bancam*), proceeds with a convincing COBUILD-style full-sentence definition but forgets to use the lemma and instead uses the (correct!) verb '*depositar*', and finally does the same when providing the example. Likewise for the phrasal verb 'bank on' which also does not use **banco** in Portuguese, but rather '*contar com*' as seen in the second example. ChatGPT's cover is blown when it offers its list of synonyms, repeating what it had used all along in the microstructure already.

However, during the remainder of 2023, and the start of 2024, colleagues in lexicography (and terminography) continued to sing the praises of ChatGPT, and fabulous results indeed kept coming in, see for instance Rees and Lew (2023, December 13), Franceschi and Pinnavaia (2023, December 20), San Martín (2024, February 25), Ptasznik and Lew (2024, March 25) and Cai et al. (2024, April 9) — needless to say, all focused on English.

With the belief that an LLM like ChatGPT is 'fantastic' for English, and the knowledge that it is often 'fake' for bilingual lexicography, the next step was to show that it simply 'fails' for exotic languages. This is exactly what was done in April 2024 during a talk at the University of Missouri (de Schryver, 2024b), where various attempts were made at having ChatGPT compile proper dictionary articles for a variety of Bantu languages, amongst others Swahili (100 million speakers) and Xhosa (19 million speakers). Figure 2 reproduces one of the slides from that talk, from which it is clear that ChatGPT cannot even pick 20 frequent words from Xhosa, as it includes a word from the neighbouring and bigger language Zulu (spoken by 27 million people). In current off-the-shelf LLMs English not only crushes other big languages, but small languages crush the even smaller ones, making them useless for lexicography.

More recently, colleagues have also begun to give attention to lexicographic applications in languages other than English, such as for Spanish (600 million speakers) (Fuertes-Olivera, 2024; Tarp & Nomdedeu-Rull, 2024), as well as exotic varieties of English such as Singlish (4 million speakers) (Chow et al., 2024). Concurrently, the pendulum is swinging back, and the need for the human touch in lexicography is being stressed again (Lew, 2024, pp. 6-7).

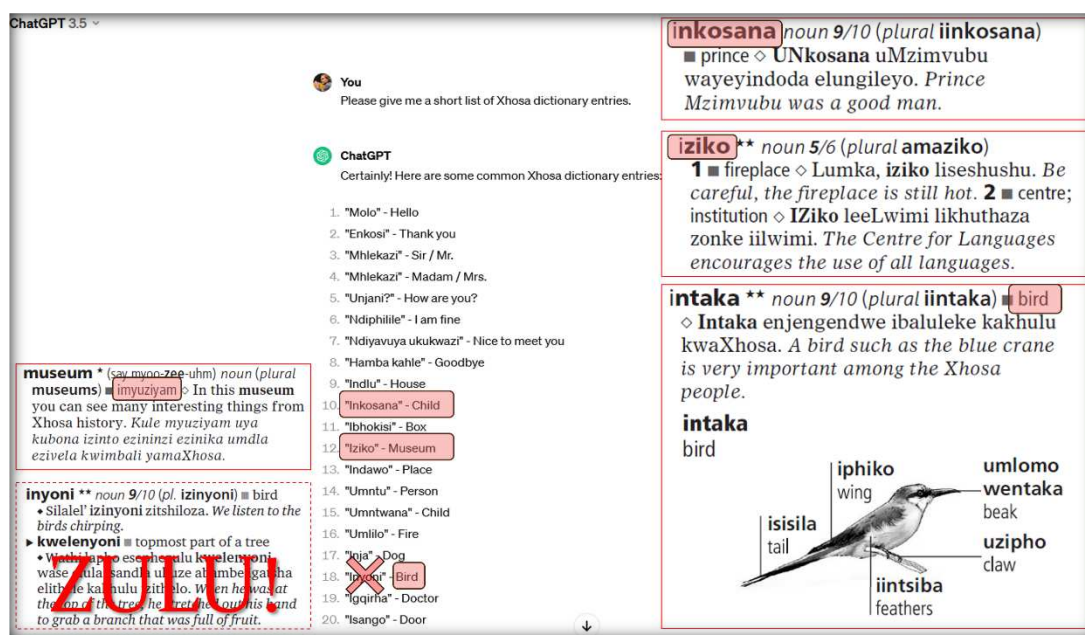


Fig. 2: Asking ChatGPT for a short list of Xhosa dictionary entries is underwhelming. Of the 20 'common' words two are wrongly translated: The noun *inkosana* is a 'prince', not a '*child', and *iziko* is a 'fireplace; centre, institution', not a '*museum' for which the loanword '*imyuziyam*' is used — as may be seen from the top three dictionary articles taken from OUP's Xhosa-English School Dictionary (de Schryver & Reynolds, 2014). Additionally and worryingly, ChatGPT even includes the Zulu word *inyoni* for 'bird' where it is actually *intaka* in Xhosa — as may be seen from the bottom two dictionary articles, respectively taken from OUP's Zulu-English School Dictionary (de Schryver, 2010) and the said Xhosa dictionary.

Even though Chatbot arenas now exist (e.g. <https://lmsys.org/>) where one can currently pair any two of up to a 100 LLMs — and thus choose from models like GPT (OpenAI), Claude (Anthropic), Gemini & Bard (Google), Llama (Meta), Command R+ (Cohere), Mistral, etc. — OpenAI's GPT models are still 'winning' overall, so it is not surprising that lexicographers have stuck by ChatGPT. Also, all the studies reviewed so far took LLMs straight 'out-of-the-box', but since the release of GPT-4o in mid-May 2024, easy customisation has been put into the hands of users. That is why, one week after the release of GPT-4o a serious attempt was made to customise this model by trying the impossible: Could it compile dictionary articles for neologisms in the extremely under-documented Bantu language Lusoga (3 million speakers, with no written tradition), by feeding GPT-4o everything that exists on Lusoga, viz. a corpus of 3.7 million tokens (de Schryver & Nabirye, 2022), and a full comprehensive monolingual dictionary in XML, the *Eiwanika* (Nabirye, 2012)? The outcome, presented at LREC-COLING in Turin (de Schryver, 2024a), was tantalising: whereas GPT-4o generated only nonsense out-of-the-box, it started to make sense with this customisation, see Figure 3. While there is still a long way to go to get to proper lexicography, and while the neighbouring language Luganda (spoken by 20 million people) still creeps in, seeing that GPT-4o can actually learn a new language and generate material in it is promising. Moreover, when asked to explain where the data for the concocted examples come from, GPT-4o is able to point to the various corpus sections from where it sourced and summarised information, meaning that the LLM truly stops being the black box feared by lexicographers. At the same time, customisation for an unknown exotic language is not good enough, so we need more.

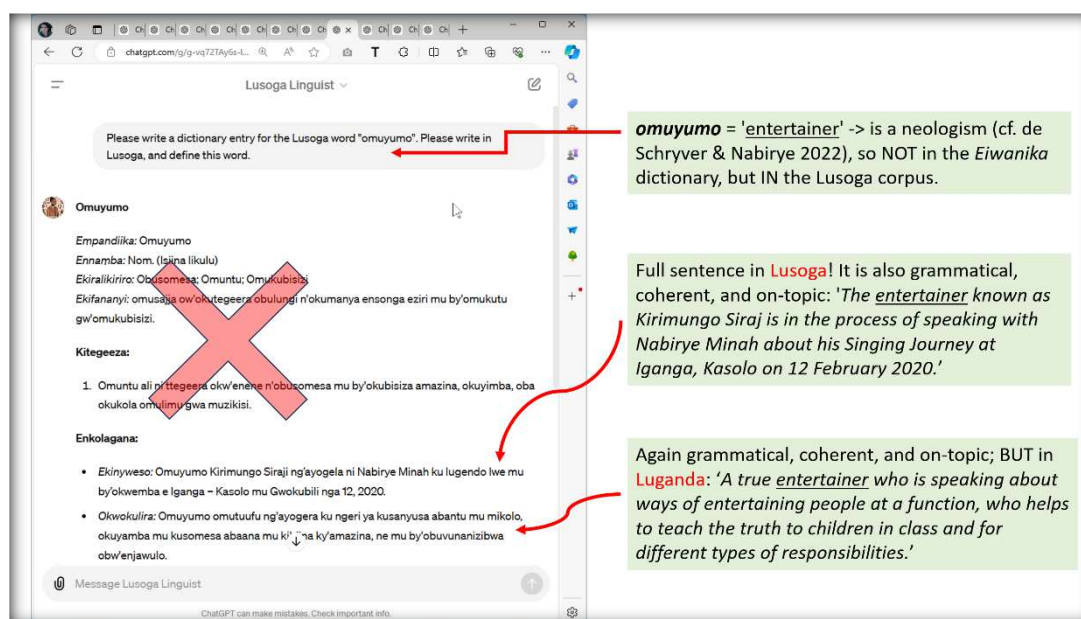


Fig. 3: Customising GPT-4o for Lusoga lexicography, a language for which it generates only rubbish out-of-the-box. Here ChatGPT is asked to compile an article for **omuyumo** 'entertainer', a word absent from the dictionary but present in the corpus – both of which were added to its 'knowledge base'. While the opening section makes no lexicographic sense, the first example is grammatical, coherent and on-topic Lusoga!

Armed with all the facts presented so far, and having proceeded from an out-of-the-box GPT (in February 2023), to a customised GPT for lexicographic purposes (in May 2024), we are now ready and willing to take the next step, namely the fine-tuning of an LLM to compile or even act as a dictionary in its own right (in October 2024). The differences between the three types are summarised in Table 1.

Table 1: Comparing out-of-the-box, customisation and the fine-tuning of GPT models for lexicography

Feature	Out-of-the-Box GPT	Custom GPT	Fine-Tuning
Modifies Core Model	No	No	Yes
Training Required	No	No, leverages pre-trained model	Yes, on a specific dataset
User Input	Text prompt	Instructions and/or relevant documents	New data for the model to learn from
Complexity	Least complex, easy to use	Less complex, potentially no-code/low-code	More complex, requires expertise
Cost	Least expensive	More expensive	A lot more expensive
Output Control	Limited control over the generated text	More control over the direction of the output	High degree of control over the generated text
Suitability for 'exotic lexicography'	Limited, may not be familiar with 'exotic language'	Can be adapted to the 'exotic language' through prompts and instructions	More directly addresses the 'exotic language' through fine-tuning

The most important difference is listed in the first line: we will need to modify the core model. In keeping with seeking a true challenge, as hinted at in the last line: we will proceed with an exotic language about which the model has initially no clue. At this stage, we are still running our experiments. But whatever the outcomes, we intend to report on them during the workshop, warts and all.

References

- Barrett, G. (2023). *Defin-o-bots: Challenging A.I. to create usable dictionary content*. Paper presented at the 24th Biennial Conference of the Dictionary Society of North America, Boulder, CO, USA, 31 May – 3 June 2023.
- Cai, B., Ng, C., Tan, D., & Hotama, S. (2024, April 9). *Low-cost generation and evaluation of dictionary example sentences*. Paper presented at the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Mexico City, Mexico, 16-21 June 2024. <https://doi.org/10.48550/arXiv.2404.06224>
- Chow, S. Y., Shin, C.-U., & Bond, F. (2024). This word mean what: Constructing a Singlish dictionary with ChatGPT. Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI-II), co-located with the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 25 May 2024.
- de Schryver, G.-M. (2010). *Oxford Bilingual School Dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgesi, Esishicilelwe abakwa-Oxford*. Oxford University Press Southern Africa.
- de Schryver, G.-M. (2023a). *Contemporary lexicography: A case study in AI [Keynote Lecture]*. Paper presented at the Inaugural Conference of the Association of Lexicography for the Americas: South, Central, Caribbean, and Mexico (Americalex-S), São Paulo, Brazil, 20-25 October 2023. <https://youtu.be/watch?v=yAxzTx7A2LU>
- de Schryver, G.-M. (2023b). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387 + 10 pp. of supplementary material online.
- de Schryver, G.-M. (2024a). *Customising LLMs for lexicography [Keynote Lecture]*. Paper presented at the 8th International Workshop on Cognitive Aspects of the Lexicon (CogALex-VIII), co-located with the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 20 May 2024.
- de Schryver, G.-M. (2024b). *Generative Artificial Intelligence & Three Lexicographic Fs: FANTASTIC for monolingual English dictionaries, FAKE for translation dictionaries, FAIL for exotic dictionaries [Invited Talk]*. Paper presented at the Department of English, University of Missouri, Columbia, MO, USA, 5 April 2024.
- 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 [Invited Talk]*. Paper presented at the 20th CODH Seminar, Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics, Tokyo, Japan, 27 February 2023. <https://youtu.be/watch?v=mEorw0yefAs>
- de Schryver, G.-M., & Nabirye, M. (2022). Towards a monitor corpus for a Bantu language. A case study of neology detection in Lusoga. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany* (pp. 233). IDS-Verlag.
- de Schryver, G.-M., & Reynolds, M. (2014). *Oxford Bilingual School Dictionary: IsiXhosa and English / Oxford isiXhosa-isiNgesi English-isiXhosa Isichazi-magama Sesikolo*. Oxford University Press Southern Africa.
- de Schryver, G.-M., Rundell, M., Tavast, A., Rychlý, P., Kokol, M., & Krek, S. (2023). *Round table on "Large language models and AI in lexicography"*. Paper presented at the 8th Electronic Lexicography in the 21st Century Conference, Brno, Czech Republic, 27-29 June 2023. https://youtu.be/watch?v=dNkksTDYa_s

- Franceschi, D., & Pinnavaia, L. (2023, December 20). New developments in English for Special Purposes (ESP) lexicology and lexicography. *International Journal of English Linguistics*, 13(7), 1–4.
- Fuertes-Olivera, P. A. (2024). Making lexicography sustainable: Using ChatGPT and reusing data for lexicographic purposes. *Lexikos*, 34, 123–140.
- Jakubíček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medved', M. Měchura, C. Tiberius, I. Kosem, J. Kallas, & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 conference: Electronic lexicography in the 21st century* (pp. 508–523). Lexical Computing. <https://youtu.be/watch?v=8e52vvDpdfQ>
- Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(704).
- Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(426).
- McKean, E., & Fitzgerald, W. (2023). The ROI of AI in Lexicography [Keynote Lecture]. In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"* (pp. 10–20). Yonsei University.
- Nabirye, M. (2012). *e-Eiwanika ly'Olusoga [Digital Lusoga Dictionary]*. Menha Publishers / TshwaneDJe HLT. <http://www.menhapublishers.com/dictionary/>
- Nichols, W. (2023). *Invisible lexicographers, AI, and the future of the dictionary [Keynote Lecture]*. Paper presented at the eLex 2023 conference: Electronic lexicography in the 21st century, Brno, Czech Republic, 27-29 June 2023. https://youtu.be/watch?v=xYpwftj_QQI
- Ptasznik, B., & Lew, R. (2024, March 25). A learners' dictionary versus ChatGPT in receptive and productive lexical tasks. <https://doi.org/10.31219/osf.io/9n685>
- Rees, G. P., & Lew, R. (2023, December 13). 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.
- Rundell, M. (2023). Automating the creation of dictionaries: Are we nearly there? [Keynote Lecture]. In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"* (pp. 1–9). Yonsei University.
- San Martín, A. (2024, February 25). *What Generative Artificial Intelligence means for terminological definitions*. Paper presented at the 3rd International Conference on Multilingual Digital Terminology Today. Design, representation formats and management systems, Granada, Spain, 27-28 June 2024. <https://doi.org/10.48550/arXiv.2402.16139>
- Tarp, S., & Nomdedeu-Rull, A. (2024). Who has the last word? Lessons from using ChatGPT to develop an AI-based Spanish writing assistant. *Círculo de Lingüística Aplicada a la Comunicación*, 97, 309–321.
- Tran, H. T. H., Podpečan, V., Jemec Tomazin, M., & Pollak, S. (2023). Definition extraction for Slovene: Patterns, transformer classifiers and ChatGPT. In M. Medved', M. Měchura, C. Tiberius, I. Kosem, J. Kallas, & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 conference: Electronic lexicography in the 21st century* (pp. 19–38). Lexical Computing. <https://youtu.be/watch?v=rQC3Rz04b20>

Contact information

Gilles-Maurice de Schryver

BantUGent (UGent Centre for Bantu Studies), Ghent University

LT³ (Language and Translation Technology Team), Ghent University

GillesMaurice.deSchryver@UGent.be