https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

Customised GPTs for Lexicography

Gilles-Maurice de Schryver, *BantUGent & LT³,*
Ghent University, Ghent, Belgium;
and Department of African Languages,
University of Pretoria, Pretoria, South Africa
(gillesmaurice.deschryver@UGent.be)
(https://orcid.org/0000-0001-7272-9878)

Abstract: About a year and a half ago, lexicographers realised that the enthusiasm they had when using ChatGPT-3.5 for English lexicography, could not be carried over to other languages, not even big ones like Portuguese. A year ago, the lexicographic community was made aware that ChatGPT-3.5 could not be used at all for small languages, and especially not for oral ones like Lusoga. In mid-2025, we notice that the models have improved drastically. Both Gemini-2.5 and especially GPT-4o perform miracles on languages like Portuguese, and for undocumented languages like Lusoga, customising GPT-4o by feeding it corpora as well as entire dictionaries is a major step forward, down to tackling neologisms. We are not yet there in full but moving forward.

Keywords: GENERATIVE ARTIFICIAL INTELLIGENCE (GENAI), LARGE LANGUAGE MODELS (LLMs), GENERATIVE PRE-TRAINED TRANSFORMER (GPT), CUSTOMISED GPTs, LEXICOGRAPHY, DICTIONARIES, LEXICOGRAPHERS, PORTUGUESE, LUSOGA

Obufunze: GPT emmumbe eya GenAI mw'isomo ly'okuwandiika amawanika. Mu ibbanga nga ely'omwaka mulala n'ekitundu eliweile, abawandiisi b'amawanika baatandiika okuteebeleza nti eitutumu lye baatuukaku nga bakozesa ChatGPT-3.5 ku lulimi Olungeleza, tilisobola kwongezebwayo kukola ku nnimi dhindi ng'Olupootigo. Omwaka oguweile, abawandiisi b'amawanika bano beene b'amala b'ayolekebwa nti ChatGPT-3.5 tesobola kukozesebwa ku nnimi dhitamanhiibwa inho, nailala ng'edho edhikaasinga okuba obw'omu ndhogela ng'Olusoga. Ebyafaayo by'ensonga dhino dhombi waigulu byandhulwa mu kitundu ky'olupapula luno ekisooka. Kino n'ekitundu eky'okubili ekilaga ebibailewo okutuuka mu magatigati g'omwaka 2025, ekilaga nti enteeko ya tekinologiya akozesebwa mu ChatGPT-3.5, ayongeile okwiluulibwamu. Kino kiidhie oluvainhuma lw'okugezesebwa okulaze nti, olwaleelo luno, enkola dha Gemini-2.5 ni GPT-4o dhiyumbwilwa inho mu ngeli ye dhikola ku lulimi ng'Olupootigo okusingila ilala. Bwe kiila ku nnimi edhitali mu buwandiike ng'Olusoga, okuliikiliza GPT-4o ebiwandiiko n'ebigambo ebiva mu itu ly'Olusoga, awalala n'ebyo ebiva mu nkenga dh'Eiwanika ly'Olusoga, kusenvwilwa kinene inho kw'aba nti kusobozesa n'ennondoola y'ebigambo ebiyaaka ebyakaingila mu lulimi luno. Waile ng'enkola eyeefaanaanhiliza kw'eyo eliwo buti ku Lungeleza n'Olupootigo ekaali kutuukibwaku mu Lusoga, aye, olugendo olukiluubilila lutandiikiibwaku.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

Ebigambo ebikulu: KALIMAGEZI-KISIBUKILA, OKUTETENKELEZA KW'ENNIMI ENNHINGI, KISIBUKILA AKAALI KUKENGULWA, GPT EMMUMBE, EISOMO LY'OKUWANDIIKA AMAWANIKA, AMAWANIKA, ABAWANDIISI B'AMAWANIKA, OLUPOOTIGO, OLUSOGA

1. Out-of-the-Box GPTs

1.1 Chatbot Arena

Even though Chatbot Arena[1] now exists where one can currently pair any two of about two-hundred LLMs — and thus choose from models like **Gemini** (Google), **GPT** (OpenAI), **Claude** (Anthropic), **DeepSeek** (High-Flyer), **Grok** (xAI), **Qwen** (Alibaba), **LeChat** (Mistral), **Hunyuan** (Tencent), **Command R+** (Cohere), **Gemma** (Google), **GLM** (Zhipu), **Step** (StepFun), **LLaMA** (Meta), **Yi** (01.AI), **Amazon Nova** (Amazon), etc. — OpenAI's GPT models still have the strongest conversational abilities, content creation, and summarisation, while Google's Gemini models are top for real-time web searches, complex reasoning, and research tasks. OpenAI's GPT and Google's Gemini models will therefore be used in this article.

1.2 Semi-visible Portuguese

The availability of the Chatbot Arena thus means that any of the prompts designed in earlier lexicographic research may now be re-used to test the quality of the output of over 200 models, to see which one works best for which type of dictionary, which language, etc. While this clearly provides for near-endless fun, all of these, as offered, are still out-of-the-box (see De Schryver 2025). Apart from the obvious improvement to have easy access to so much variety, one should check whether some of the recent LLMs are still as bad for non-English as those from a year and a half ago, as was the case for Portuguese when the then ChatGPT-3.5 was used at the end of October 2023 (De Schryver 2023b).

As examples, Figures 1 to 4 show a comparison between the Gemini-2.5 model and the GPT-4o model, for the prompts shown at the top-right in each case, to get to proper dictionary articles for: (1) the verb of communication *recommend* in an English–Portuguese learners' dictionary, (2) the lemma *table* in an English–Portuguese learners' dictionary, (3) the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in English, and (4) the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in Portuguese. The full texts offered by both the Gemini and GPT models, received as-is and thus left untouched, are presented in Addendum A, available online, as Figures 1 to 4 only show the opening sections. Click here for Addendum A.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

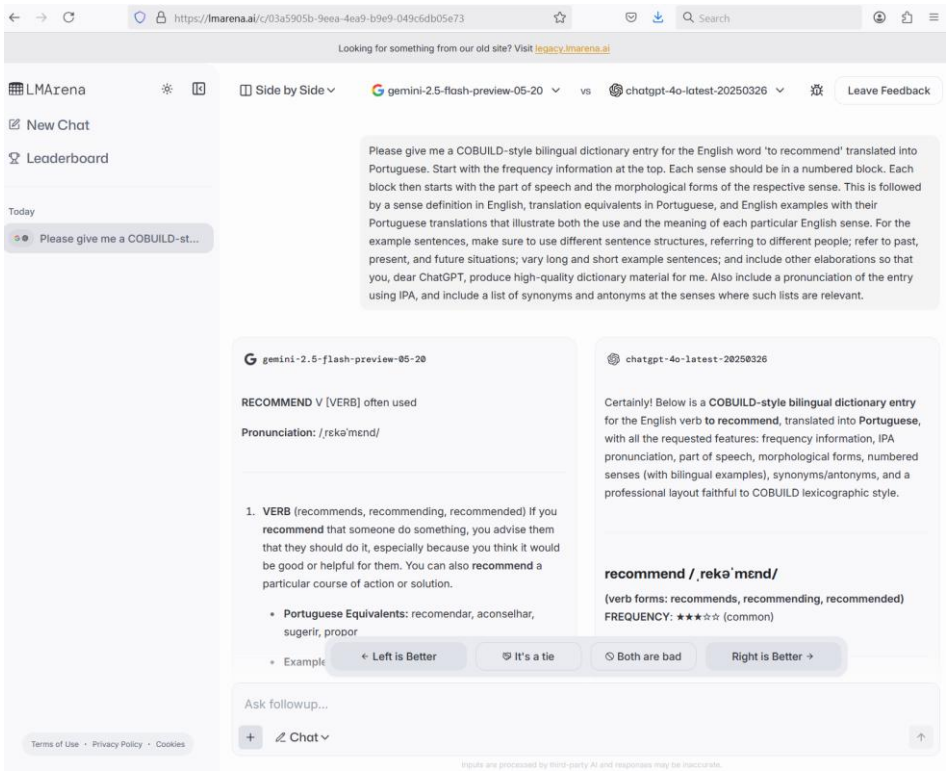

Figure 1: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the verb of communication *recommend* in an English–Portuguese learners' dictionary

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

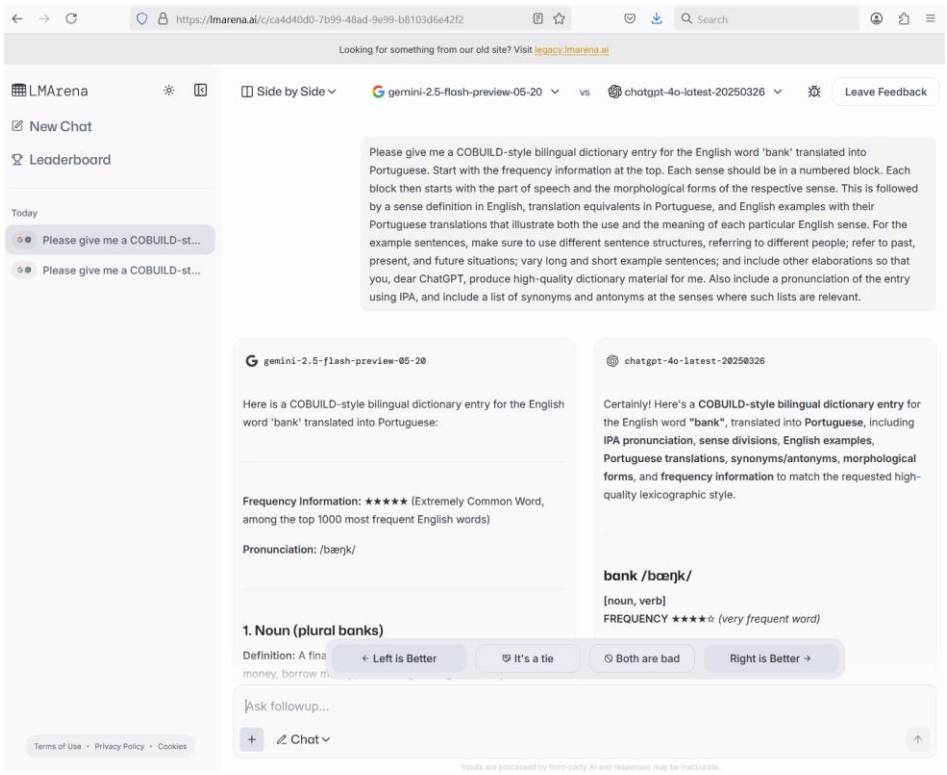

Figure 2: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in an English–Portuguese learners' dictionary

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

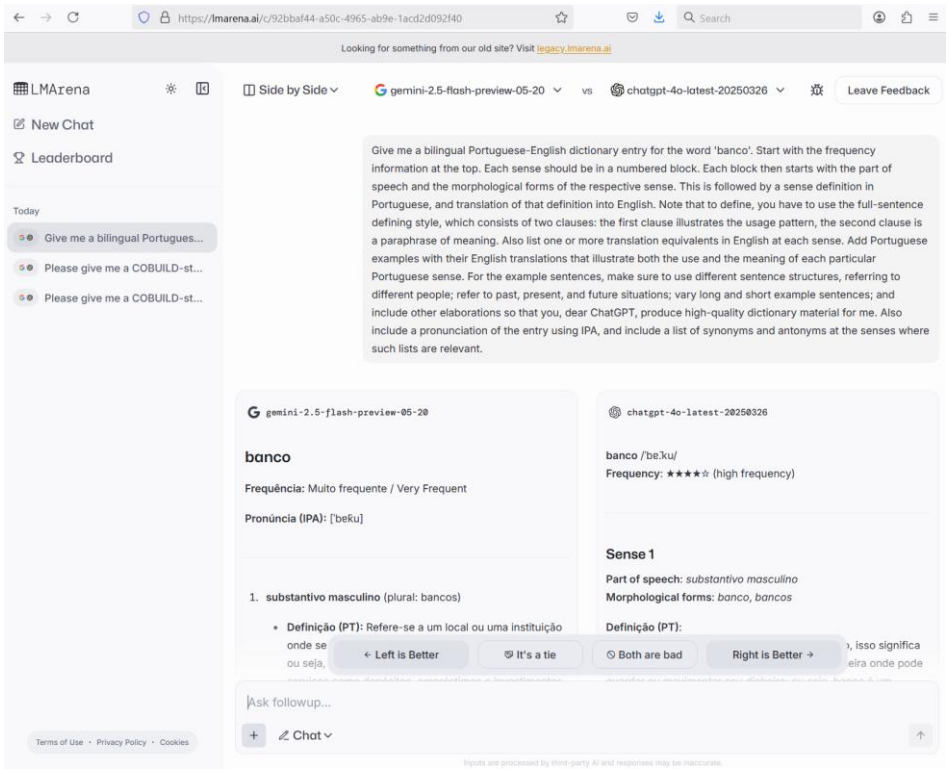

Figure 3: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in English

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

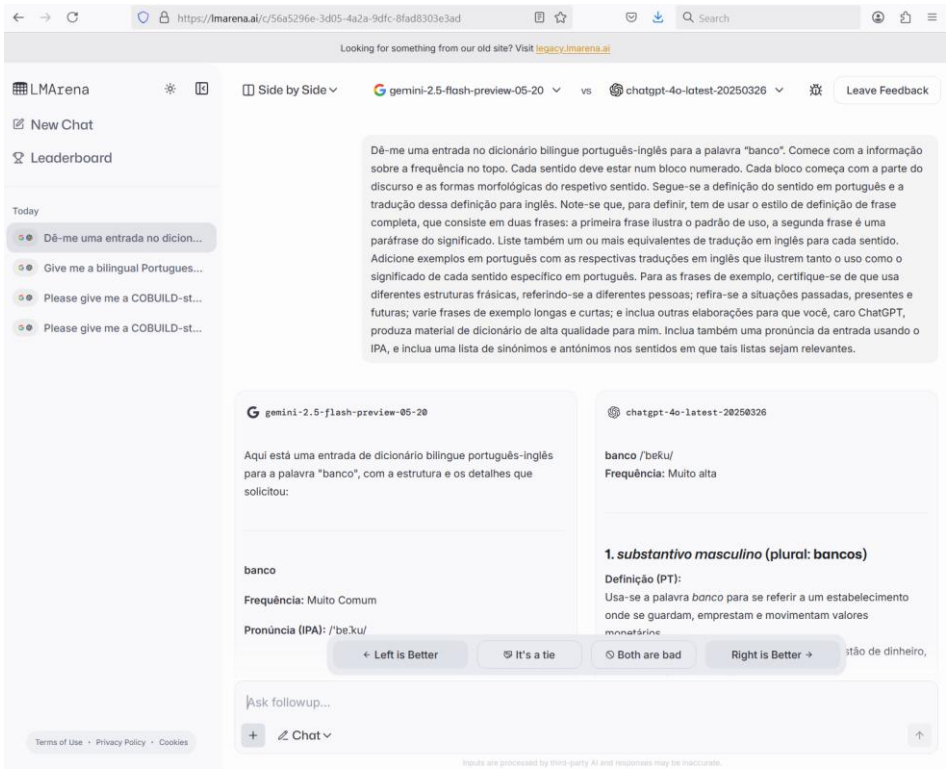

Figure 4: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in Portuguese

Astonishingly, all the problems ChatGPT-3.5 had when trying to treat non-English like Portuguese for lexicographic purposes (see De Schryver 2025) are simply gone. This is most evident from the treatment of the lemma *table* when prompting GPT-4o, no matter the direction of the dictionary, no matter whether the prompt is in English or in Portuguese. As of June 2025 no more hallucinations are evident, and only noun senses are given, and thus no more inventions of non-existent verbal uses of Portuguese *banco*. Given that there are four prompts, submitted to two LLMs, there are a total of 8 answers in Addendum A, available online, numbered A1 to A8. Click here for Addendum A. More interesting summaries now follow for each of the eight dictionary articles produced via GenAI.

In **A1**, Gemini-2.5 for *recommend*, we immediately note excellent COBUILD-style full-sentence definitions (FSDs). On top, at sense 3, following the FSD, a second sentence starts with '*This usage often implies that …*' — brilliant, as indeed, meaning is the result of usage, with meaning potentials only triggered by the

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

context, as Patrick Hanks reminded us throughout his life (Hanks 1979, 1988, 2000, 2002, 2012, Hanks and Franklin 2019, Hanks and Može 2019, Grefenstette and Hanks 2023). The Portuguese translations of the examples at sense 3 are also brilliant, as the various translation equivalents offered are indeed used (the underlinings, both full and broken up, were added by the present author). Lastly, the separation into three uses, and thus three senses, makes perfect sense.

In **A2**, GPT-4o for *recommend*, and thus even though the same verb is treated, the three uses/senses are not the same as in A1. In a way, the uses here are even further apart, as is also obvious from the additions of [VERB + that] at sense 2, and [VERB + itself, be + VERB] at sense 3. All of this implicitly refers to Patrick Hanks's Corpus Pattern Analysis (Hanks 2004, Hanks and Pustejovsky 2005, Hanks 2008, Hanks et al. 2018, Hanks and Ma 2021), and more specially his (unfinished) *Pattern Dictionary of English Verbs*[2]. As impressive are the Usage Notes, which fully delve into linguistics, and the Related entries linking the verb to the noun and adjective.

In **A3**, Gemini-2.5 for *bank*, the first sense (as a financial institution) and second sense (as the sloping land alongside a river) mention that there are 'no direct antonyms' with a reason why — smart. At the second sense the mistake is not made to include **banco* as one of the translation equivalents, and as to the synonym 'levee' the context is added between brackets: '*for a raised bank to prevent flooding*' — very good. At the third sense (as a mass or pile of something) the translation equivalents are provided with context — excellent. At the fourth sense (as the verb *to bank*) the definition is sadly not COBUILD-style, but the mistake is not made to include **banco* as a translation equivalent. The fifth sense (of an aircraft or vehicle) is even far-more detailed than as seen in the actual COBUILD.[3] The sixth sense (a row or series of similar items) correctly illustrates two of the translation equivalents. At the seventh sense, we see all sorts of correct compounds. At the eight sense the phrasal verb *bank on* is treated, but alas no FSD. The examples here use two different translation equivalents.

In **A4**, GPT-4o for *bank*, the first sense (as a financial institution) *does* include antonyms — even smarter than Gemini-2.5. At sense two (as the side or edge of a river) only *margem* is (correctly) offered as a translation equivalent. Sadly, at the third sense (as the verb *to bank*), no proper FSD, and all examples use only one of two possible translation equivalents. At the fourth sense (a row or series of similar things) the third example uses a translation equivalent not listed; the other two (correctly) use two different ones. All is well with the fifth sense (the noun in aviation or driving). Idem for the sixth sense (the verb in aviation or driving), and seventh sense (as a mass or layer of something). The 'Summary Usage Note' is most-excellent and reminds the user of the highly polysemous nature of this lemma, as it is used both literally and metaphorically, across various fields, pointing out that the resulting translation equivalents in Portuguese truly depend on context. Brilliant lexicography.

In **A5**, Gemini-2.5 for Portuguese *banco* (prompt in English), at the first sense (as a financial institution) the translation equivalent and use in example sen-

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

tences are all correct. No antonym is provided, although we know, thanks to GPT-4o under A4 sense 1 that there are antonyms (namely *borrower*, *debtor*). At the second sense (as furniture for sitting), three translation equivalents are provided and two are used in the examples (*bench*, *seat*). At the third sense (as a sandbank, or school of fish), four translation equivalents are provided and all four are used in the examples (*sandbank*, *school*, *bank*, *shoal*) — good lexicography in a learners' dictionary. At the fourth sense (as a blood or organ bank), only bank is offered as a translation equivalent, but all examples combine it (once each with *blood*, *organ*, *gene*, *tissue*). None of senses 2 to 4 include antonyms; but GPT-4o does for all, see the next example.

In **A6**, GPT-4o for Portuguese *banco* (prompt in English), the definition at the first sense (as a financial institution) is a good attempt at an FSD, which is always tougher for a noun than for a verb. Two translation equivalents are offered, but only one is used in the three examples (*bank*). As antonyms of this use of *banco*, Portuguese *cofre* 'safe' and *colchão (em uso figurado para guardar dinheiro em casa)* 'mattress (figuratively for storing money at home)' are offered — very good. At the second sense (as a bench in a park), the FSD is near-excellent. Two translation equivalents are offered, of which only one is used in all three examples (*bench*). Here too, the antonyms are head-on, Portuguese *chão* 'floor' and *cadeira individual* 'single chair'. At the third sense (as a substitute bench), the FSD is again near-excellent. While two translation equivalents are offered, again only one is used in all three examples (*bench*), though here it makes more sense, as the expanded *substitute bench* is of course rarer. The antonyms are again very good, Portuguese *titular* 'starter' and *escalado* 'line-up'. At the fourth sense (as a database), we see the usual near-FSD, two translation equivalents of which one used in the three examples (*database*), while the antonyms are original, Portuguese *ausência de dados* 'no data' and *sistema manual* 'manual system'. At the fifth sense (as a sandbank), again a near-FSD, three translation equivalents, but in the three examples only one of them (twice *sandbank*, plus even *bank*, which is not among the translation equivalents), and as antonyms the debatable Portuguese *canal* 'canal' and *profundidade* 'depth'. The concluding summary at the end, focusing on the polysemy of Portuguese *banco*, found with financial, physical, sportive, digital, and geological uses — listed in the order of the five senses — is most precise and helpful. That note continues with the claim that this GenAI entry is *'offering a comprehensive bilingual reference for language learners and lexicographers alike'*. Compared with the ridiculously short entry for *banco* as offered by Collins,[4] this is undoubtedly better, near-perfect even. Near-perfect, as the FSD does indeed contain the two parts, but is followed by a semi-colon and then a more traditional definition on top, making the FSD a bit too long. Other than that, let GPT-4o compile an entire dictionary in this format overnight, for the full lexicon, and we might have (the draft of) a good bilingual Portuguese–English learners' dictionary right away.

In **A7**, Gemini-2.5 for Portuguese *banco* (prompt in Portuguese), even more is presented in Portuguese as a result of prompting in Portuguese. At the first

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

sense (as a financial institution), the definition consists of two sentences, so not an FSD, but if the two had been merged and shortened, it comes close. Of the two translation equivalents, all four examples use one (*bank*). Two synonyms are offered, no antonyms. At the second sense (as a bench in a park), we see the same type of non-FSD. Of the three translation equivalents that are offered, two are used in the four examples (*bench*, *pew*). Three synonyms are offered, no antonyms. At the third sense (as a sandbank), again a non-FSD. Of the four translation equivalents that are offered, three are used in the three examples (*sandbank*, *mudflat*, *shoal*). Three synonyms are offered, no antonyms. At the fourth sense (as a blood or data bank), another non-FSD. Of the four translation equivalents that are offered, only one is used in the four examples (*bank*). Four synonyms are offered, no antonyms. At the fifth sense (as a school or shoal of fish), the same issue with the non-FSD is evident. Of the two translation equivalents that are offered, only two are used in the three examples (*school*, *shoal*). Only one synonym is offered, no antonyms.

In **A8**, GPT-4o for Portuguese *banco* (prompt in Portuguese), there is again more in Portuguese, so here too the result of the Portuguese prompt. The first sense (as a financial institution) presents a definition in the FSD-style, use vs. meaning, followed by a second sentence that is a variation on the meaning only. Surely, the two sentences could have been merged, to achieve the FSD-goal of a learners' dictionary. One translation equivalent is offered and used in all three examples (*bank*). Two synonyms follow, with this time an indication that there are no antonyms in this sense (even though GPT-4o offered some for this sense in A4 and A6). The second sense (as a long seat) presents a similar FSD cum extra sentence. Two translation equivalents are offered, but only one is used in all three examples (*bench*). Three synonyms follow, and two antonyms (*cadeira* 'chair' and *poltrona* 'armchair'). The third sense (as a subaquatic formation) again has the pseudo-FSD and then a second sentence on a reformulation of the meaning. Four translation equivalents are offered, of which two are used in two examples (*sandbank*, *bank*) and then one not listed under the translation equivalents (*school*). Three synonyms follow, and two antonyms (*mar profundo* 'deep sea' and *leito rochoso* 'bedrock'). The fourth sense (as a vehicle seat) once again has the pseudo-FSD and second sentence. Two translation equivalents are offered, and both are used in all three examples (*car seat*, *seat*). Two synonyms follow, and two antonyms (*espaço livre* 'free space' and *chão do veículo* 'vehicle floor'). The fifth sense (as a collection of information) repeats the pseudo-FSD plus extra sentence on meaning. Two translation equivalents are offered, and both are used in all three examples (*database*, *data bank*). Two synonyms follow, and two antonyms (*informação dispersa* 'scattered information' and *fonte não estruturada* 'unstructured source'). At the very end, a pronunciation is offered for the lemma (both in European and Brazilian Portuguese, although the same here), followed by an invitation to treat GPT-4o as a professional lexicographer, as it invites the user with: *'If you'd like examples of compositions with 'banco' (such as data bank, square bank, image bank), I can draw up more specific blocks. Would you like to continue?'*.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

Even a cursory look — but especially the more detailed study just presented — of the lexicographic output of Gemini-2.5 vs. GPT-4o, makes clear that for lexicography GPT-4o is the better tool. Comparing the output under A6 (with the prompt in English) vs. the output under A8 (with the prompt in Portuguese) further indicates that prompting in English remains better than prompting in another language. It is therefore not surprising that over the past two years and a half, lexicographers not only started with but also mainly stuck by ChatGPT, as seen from the detailed study of a sample of 100 contributions on 'GenAI in lexicography' analysed in De Schryver (forthcoming-a, forthcoming-b). In any case, for large languages like Portuguese there is thus no need anymore to customise an LLM, at least not for general lexicographic purposes.

2. Customised GPTs

The necessary move for large non-English languages — so from a 'hyper-visible' to a 'semi-visible' uptake in technology like GenAI — has been made: Recent LLM models are now also achieving top lexicographic performances for languages such as Portuguese with out-of-the-box tools. Of course these same tools won't have improved much for low-visible languages such as the Bantu languages, and very likely not for a local-only visible language such as Lusoga: As far as we know, no extra data was added anywhere online in recent years for that language, so for exotic oral languages like Lusoga it becomes important to now know whether a tool like GPT-4o can be used to take the next step for them, namely *customisation*.

2.1 Local-visible Lusoga

Since the release of GPT-4o on 13 May 2024, customisation has been put into the hands of users. Within days of 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 oral — so extremely under-documented — Bantu language Lusoga, by feeding GPT-4o everything that exists on Lusoga, viz. a corpus of 3.7 million tokens (De Schryver and Nabirye 2022), and a full comprehensive monolingual dictionary in XML, the *e-Eiwanika ly'Olusoga* (Nabirye et al. 2012)[5]? The outcome, presented on 20 May 2024 as a keynote at the CogALex workshop of LREC-COLING 2024 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.

Indeed, following an opening prompt in which GPT-4o was merely informed that the word *ente* is 'cow' in Lusoga; an out-of-the-box generation of a definition for it produced utter gibberish: the reply seen in Figure 5 is meaningless.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

Figure 5: Trying the out-of-the-box GPT-4o, asking for a Lusoga definition of *ente* 'cow'

A team of three therefore joined forces — programmer David Joffe from South Africa, native Lusoga speaker Minah Nabirye from Uganda, and lexicographer Gilles-Maurice de Schryver from Belgium — and messaged the built-in GPT builder to create a custom GPT, which they called 'Lusoga Linguist'. After agreeing on a profile picture, the refinement of the context could start, with Figure 6 showing one of the goals.

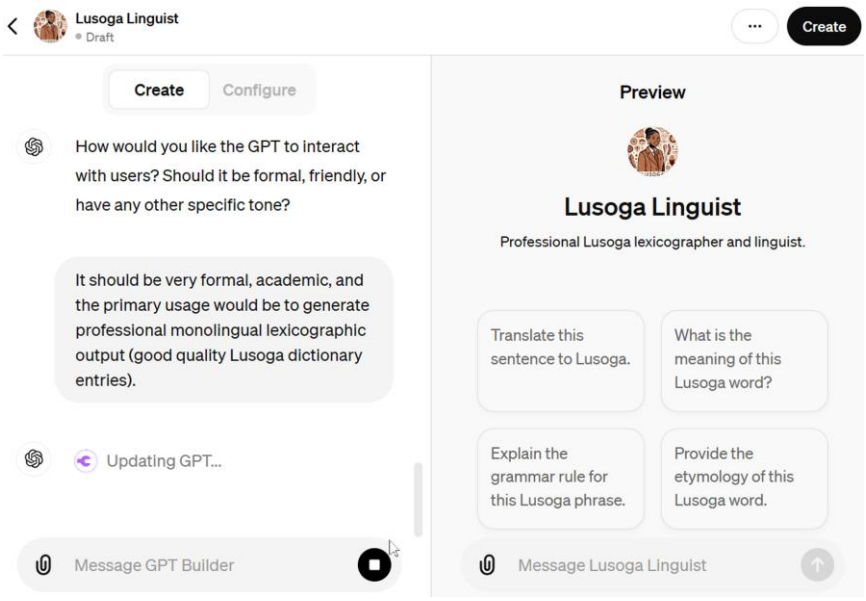

Figure 6: Work on refining the goal of the custom GPT for Lusoga lexicography has begun

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

After that, a first attempt was made at adding Lusoga corpus files to the 'Knowledge base', as seen in Figure 7.

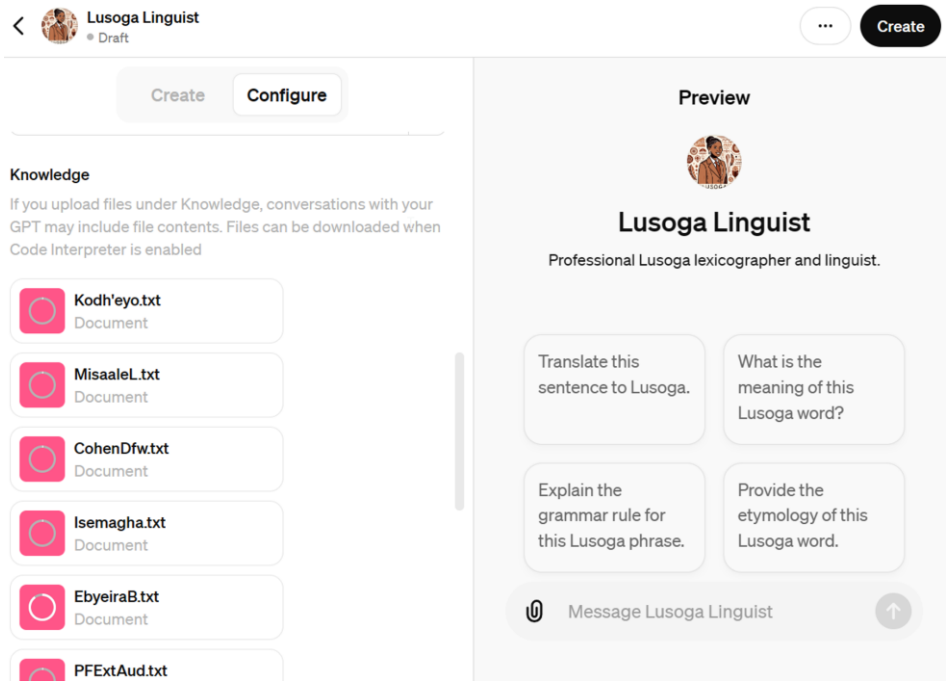

Figure 7: Adding a Lusoga sample corpus to the 'Knowledge base' of GPT-4o

Adding very many files did not work, so a single word file (of about 500KB) with 10,000 lines of corpus texts was added. In there the word *ente* 'cow' occurred 36 times. Once done, GPT-4o was instructed to '*study the uploaded corpus of Lusoga text*' to then use this knowledge before moving to the '*task of generating professional monolingual Lusoga dictionary entries*', as seen in Figure 8.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

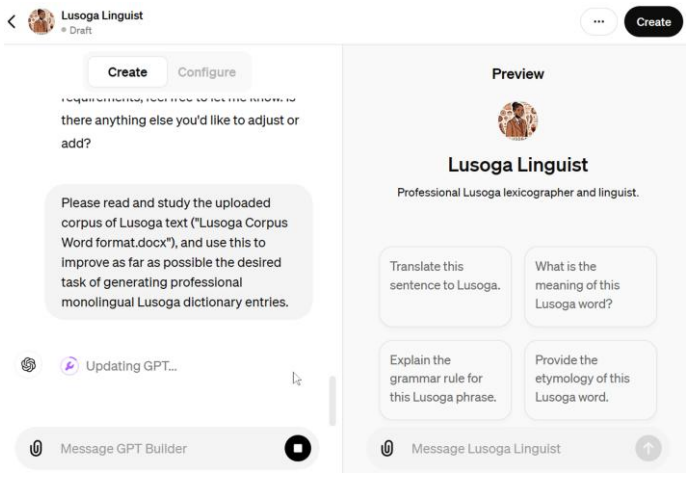

Figure 8: Instructing GPT-4o to study the submitted sample Lusoga corpus

The initial prompt was then repeated, as seen in Figure 9, together with the output.

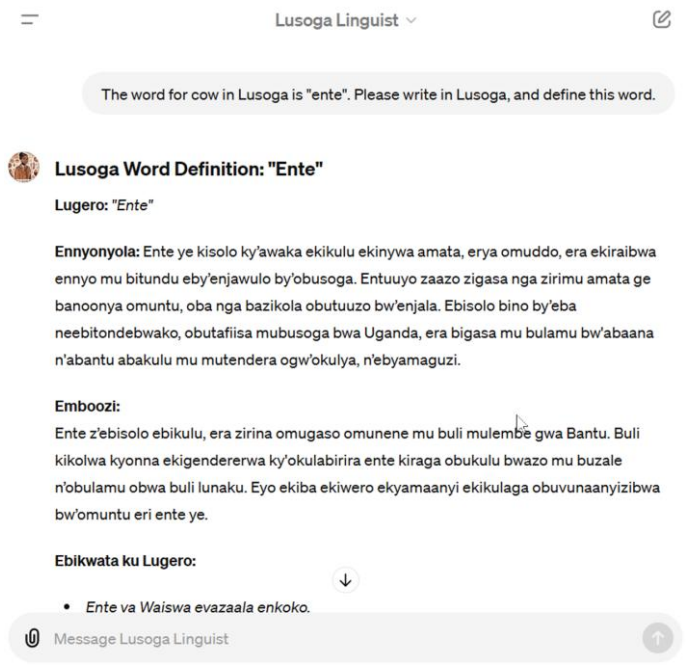

Figure 9: A customised GPT-4o tries to define Lusoga *ente* 'cow', using a sample corpus

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

The output seen in Figure 9 reads more like a story than a dictionary. Plus, most is in the neighbouring and much larger language Luganda, with little in Lusoga. The opening of the definition translates as '*A cow is an animal that drinks milk and eats grass …*' and the last line of the screenshot has an example that translates as '*The cow of Waiswa gives birth to a chicken*'. Further prompting in which GPT-4o was asked to only use Lusoga (and no Luganda) and to compile a proper dictionary (rather than a story) failed to improve the result. The decision was therefore taken to load samples of the monolingual Lusoga dictionary into the 'Knowledge base' of GPT-4o. That work was published as a paper dictionary (Nabirye 2009) as well as an e-dictionary (Nabirye et al. 2012), and for the latter both as a downloadable dictionary app as well as a free online dictionary. See Addendum B, available online (click here).

After uploading the sample with monolingual Lusoga dictionary examples, GPT-4o was instructed to study those sample entries, in the hope that it would 'understand' the lexicographic composition. Figure 10 displays the start of the conversation on this level, while Figure 11 displays the end of the analysis by GPT-4o. Recall that all the data submitted for training of the model is fully in Lusoga, not in any other language, so the task is clearly hard. On the other hand, there is a good lexical overlap between Lusoga and Luganda, a language for which there are considerable amounts of data online.

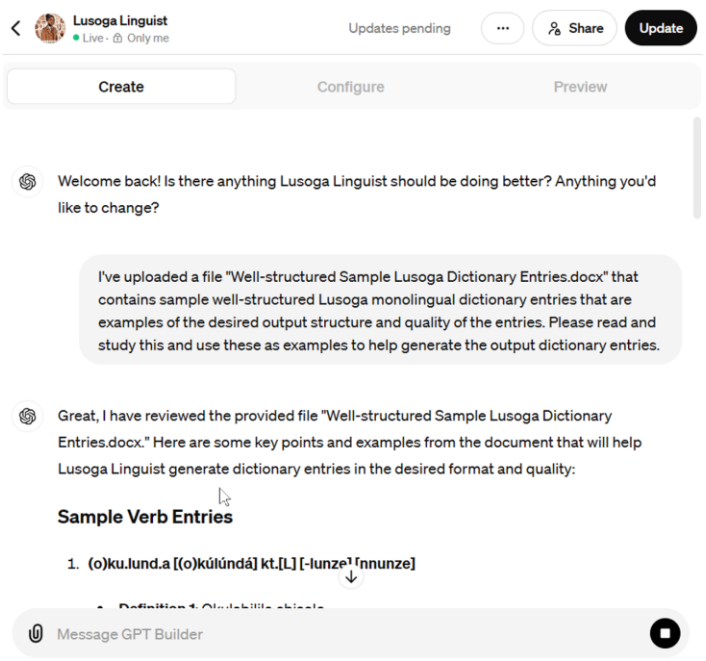

Figure 10: After adding a sample of monolingual Lusoga dictionary entries to the 'Knowledge base', asking GPT-4o to study the structure of that

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

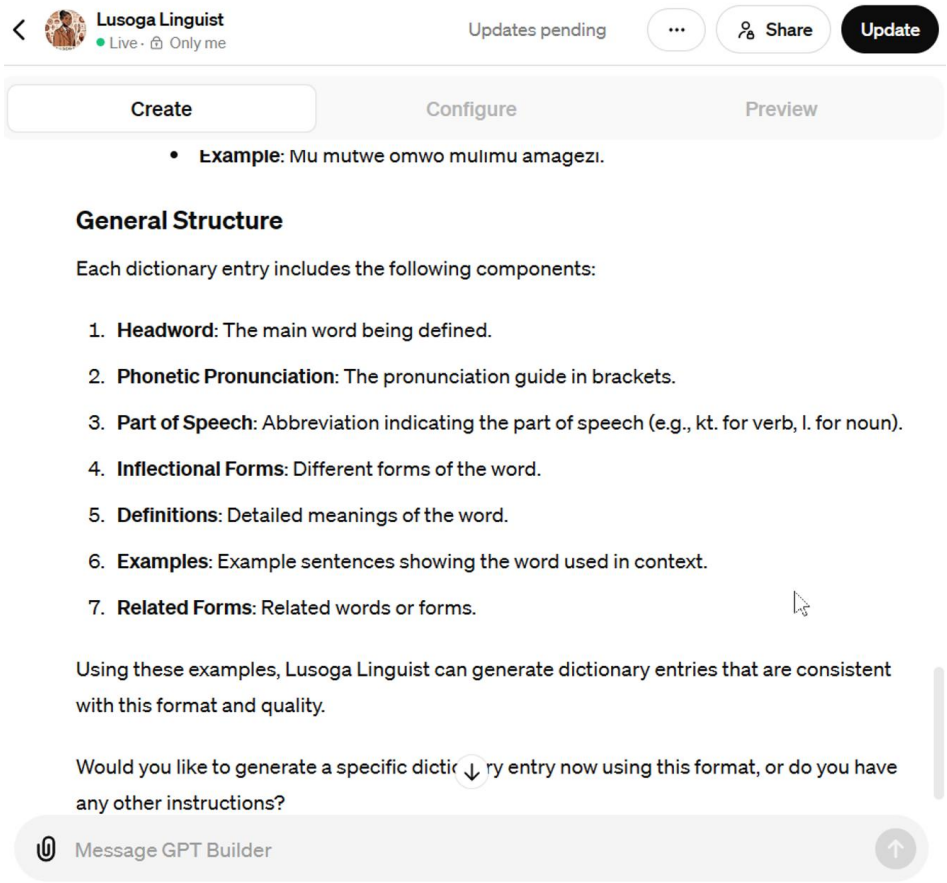

Figure 11: GPT-4o has 'learned' the lexicographic composition of entries in the monolingual Lusoga dictionary

Equipped with a sample Lusoga corpus, and a sample of Lusoga dictionary entries, GPT-4o was asked for a definition of *omukono* 'hand/arm' (not being among the sample dictionary entries), as seen in Figure 12. The overall appearance of Figure 12 is only slightly more dictionary-like. Plus, one continues to notice both Luganda and Lusoga, but whenever it is Lusoga it is still not coherent. Actually, but still surprisingly, only sense 2 is (1) written in Lusoga, and (2) makes somewhat sense, as the translation is '*One of the *drafts of the body ...*'. Is the custom GPT-4o starting to learn Lusoga, or simply lucky in making things up? Two efforts were undertaken to make the custom GPT-4o much better.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

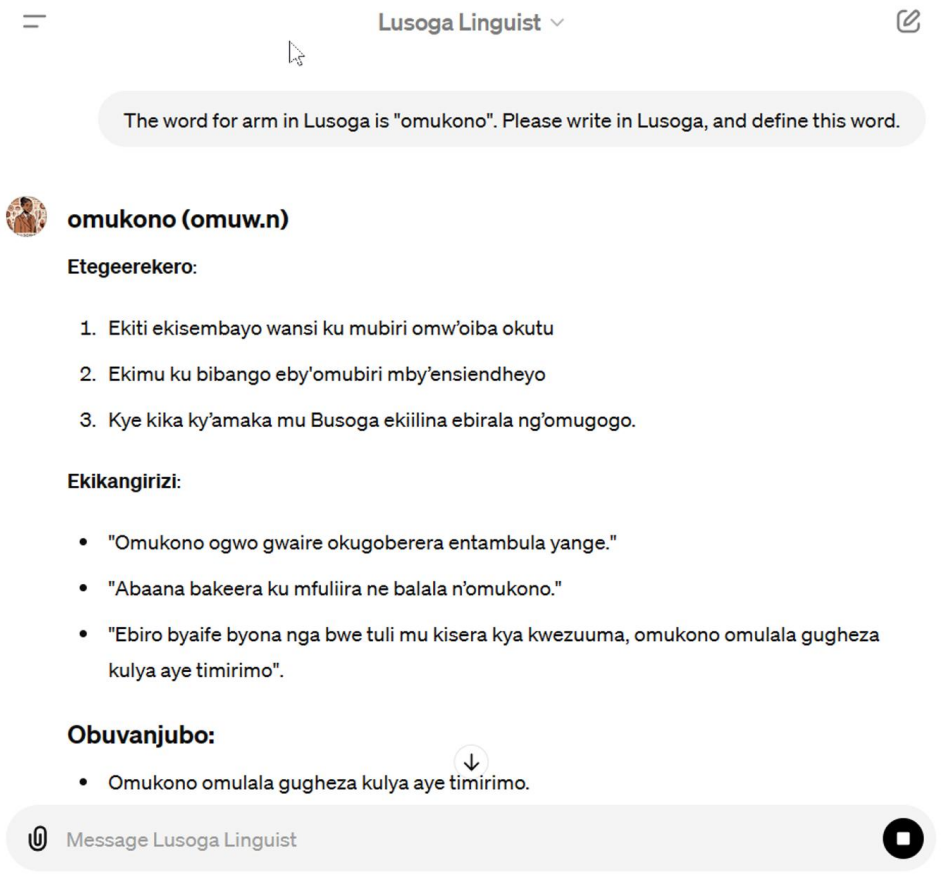

Figure 12: Asking the first version of a customised GPT-4o for the Lusoga entry *omukono* 'hand/arm'

First, given that up to 20 documents may be added to the 'Knowledge base', with each not containing more than 2 million LLM tokens and not larger than 512MB, all the files of the sample corpus were taken away and were now replaced with the full Lusoga corpus of 3.7 million word tokens, cut up in ten parts.[6] That corpus was the result of about two decades of corpus building by Minah Nabirye (De Schryver and Nabirye 2018, 2022), and enlarging it even further is a truly heroic effort.[7] Figure 13 shows how the 3.7 million Lusoga word tokens were added to the 'Knowledge base', and Figure 14 prompts with the same question as in Figure 12, but now assuming that the full corpus 'taught' the customised GPT-4o model more about the Lusoga language.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

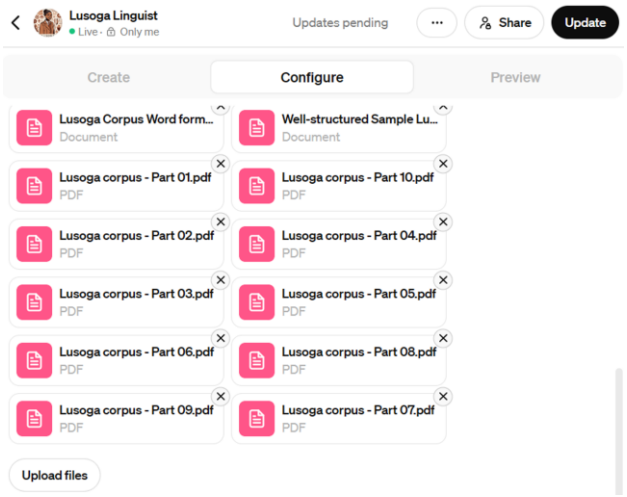

Figure 13: Adding all 3.7m word tokens of Lusoga to the 'Knowledge base' of the customised GPT-4o

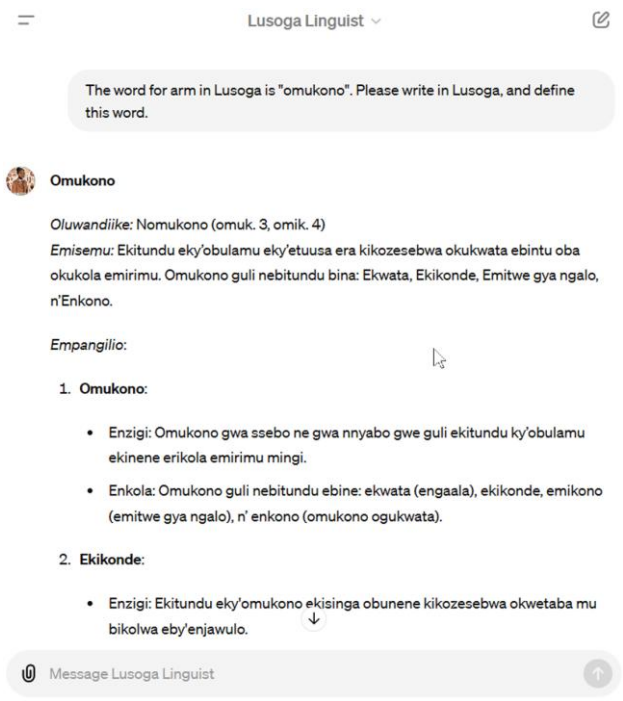

Figure 14: Asking the re-customised GPT-4o for the Lusoga entry *omukono* 'hand/arm'

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

While still far from correct, Figure 14 suddenly comes close to resembling real Lusoga and the analysis of Lusoga. On the first line, for instance, while the singular prefix is (obviously) correct (*omu-* in class 3), for the plural *omi-* in class 4 is suggested, where it should be *emi-* in class 4. Then, the definition of *omukono* 'hand/arm' contains not only too much Luganda, but also Swahili, and translates as '*The part of the body which is close and is used to hold things and to do work*'. A rather 'funny' definition, but it is valid: it is a close body part, and hands are used to hold things and also perform work. Compare this to the actual entry in the *e-Eiwanika ly'Olusoga*, as seen in the first screenshot of Addendum B, available online (click here).

This result logically led to the second effort, namely the addition to the 'Knowledge base' of the XML-file of the entire 600-page *e-Eiwanika ly'Olusoga*, in other words the full monolingual Lusoga dictionary, as seen in Figure 15. (The 'Well-structured Sample' was taken out.)

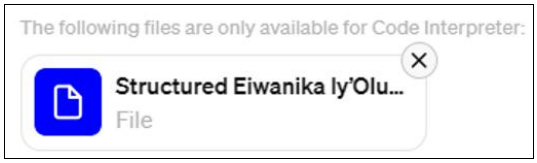

Figure 15: Adding the XML-file of the full monolingual *e-Eiwanika ly'Olusoga* to the 'Knowledge base' of the customised GPT-4o

This was immediately followed by an instruction to the custom GPT-4o to study this full dictionary, as seen in Figure 16.

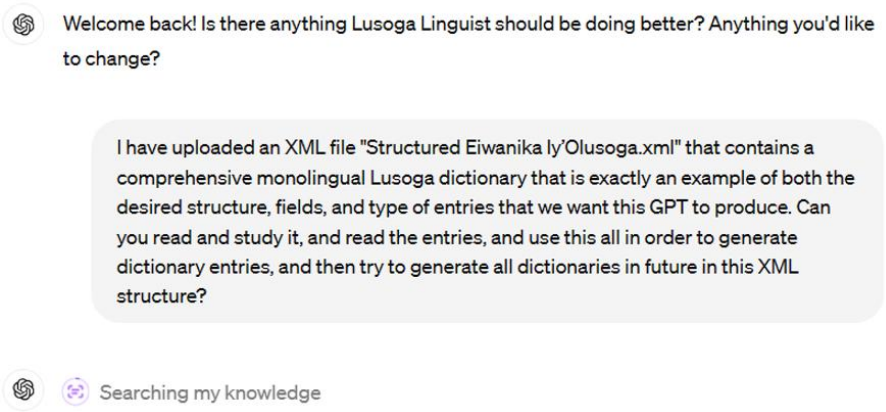

Figure 16: Adding the XML-file of the entire monolingual *e-Eiwanika ly'Olusoga* to the 'Knowledge base' of the customised GPT-4o

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

Working with ChatGPT, Barrett (2023: 10) had noticed that even their GPT-4 model performed some stunts at times, namely to blatantly copy over existing dictionary contents wholesale (in his case for definitions of *virus*). In order to be sure that the customised GPT-4o model does not simply go on a copy-and-paste venture now that the full *e-Eiwanika ly'Olusoga* is in the 'Knowledge base', an additional difficulty was added, namely to focus on neologisms. The lemma *omuyumo* 'entertainer' was chosen, as it is a known neologism, first described in the academic literature in De Schryver and Nabirye (2022: 843), and *not* in the monolingual Lusoga dictionary, but *in* the Lusoga corpus. In Figure 17 one sees the reply to the question to compile a monolingual dictionary article for *omuyumo* 'entertainer'.

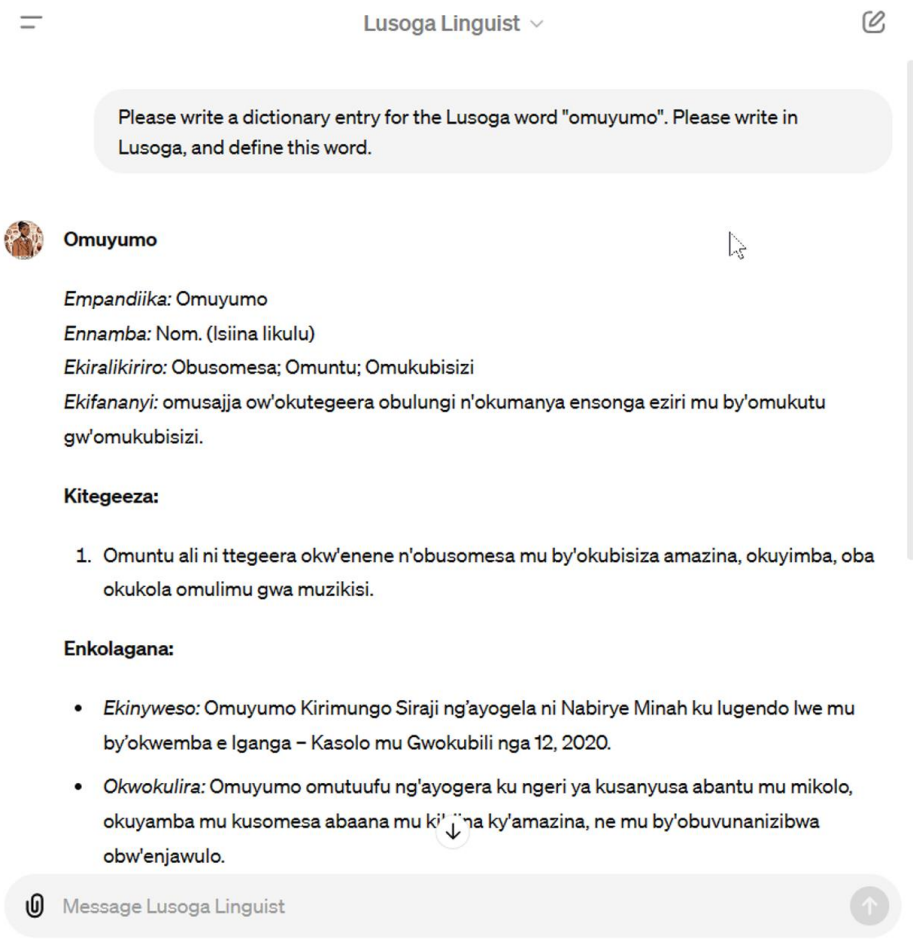

Figure 17: Asking the re-re-customised GPT-4o to compile a dictionary entry for Lusoga *omuyumo* 'entertainer'

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

While the opening section of Figure 17 still makes no lexicographic sense, the first bullet at the bottom provides an example that is fully grammatical, coherent and on-topic Lusoga. The translation of this example sentence is: '*The entertainer known as Kirimungo Siraj is in the process of speaking with Nabirye Minah about his Singing Journey at Iganga, Kasolo on 12-02-2020*'. This sentence does not appear in the corpus, and is clearly an excellent summary of one of the transcribed recordings made in 2020 for a book that came out four years later, ironically called *Obuyumu mu Busoga* 'Entertainment in Busoga' (Nabirye 2024). Unfortunately, while the example at the next bullet is again grammatical, coherent, and on-topic, it is couched in Luganda: '*A true entertainer who is speaking about ways of entertaining people at a function, who helps to teach the truth to children in class and for different types of responsibilities*'.

While there is clearly still a long way to go to get to proper lexicography, and while the neighbouring language Luganda still creeps in, seeing that GPT-4o can actually start learning a brand-new language and generate some dictionary material in it is promising, and, as far as we know, has not yet been tried for any other language within the field of lexicography.[8]

Moreover, when asked to explain where the data for the generated examples come from, GPT-4o is able to point to the various corpus sections from where it sourced and summarised information. For instance, for the example in Luganda at the last bullet in Figure 17, Figure 18 adds the Lusoga sources — in other words, here the example is based on material taken from the Lusoga corpus, but strangely 'translated' from Lusoga into Luganda.

- Ekinyweso: Omuyumo Kirimungo Siraji ng'ayogela ni Nabirye Minah ku lugendo lwe mu by'okwemba e Iganga – Kasolo mu Gwokubili nga 12, 2020.

- Okwokulira: Omuyumo omutuufu ng'ayogera ku ngeri ya kusanyusa abantu mu mikolo, okuyamba mu kusomesa abaana mu kibiina ky'amazina, ne mu by'obuvunanizibwa obw'enjawulo【21:0†source】【21:1†source】【21:2†source】【21:3†source】【21:5†source】【21:6†source】【21:7†source】【21:10†source】.

Figure 18: The custom GPT-4o reveals the Lusoga sources used for the last example in Figure 17

Experimenting more with trying to discover what was used to come up with certain definitions and/or examples, at some point later on, while working on the neologism *obukurisitu* 'Christianity', the custom GPT-4o claims all of what is seen in Figure 19, where the most important claim is: '*Each reference helps trace*

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

back the information used in the dictionary entry to its original location in the uploaded documents. This ensures that the definition provided is well-supported by the source material'. And a bit later yet again, when working on the neologism *omukwampa* 'the direction of', the custom GPT-4o even goes as deep as to claim all that is seen in Figure 20, where the summary is: '*The references are not from a single contiguous line or section but rather from different parts of the text that discuss related topics, such as boundary demarcations, geographical descriptions, and historical or cultural contexts*'. All of this has indeed a lot to do with the meaning of the neologism *omukwampa* 'the direction of', so if all true, this means that a customised LLM stops being the full black box feared by lexicographers, as it becomes somewhat less opaque. At the same time, customisation for an unknown exotic language such as Lusoga is not good enough, so we need more.

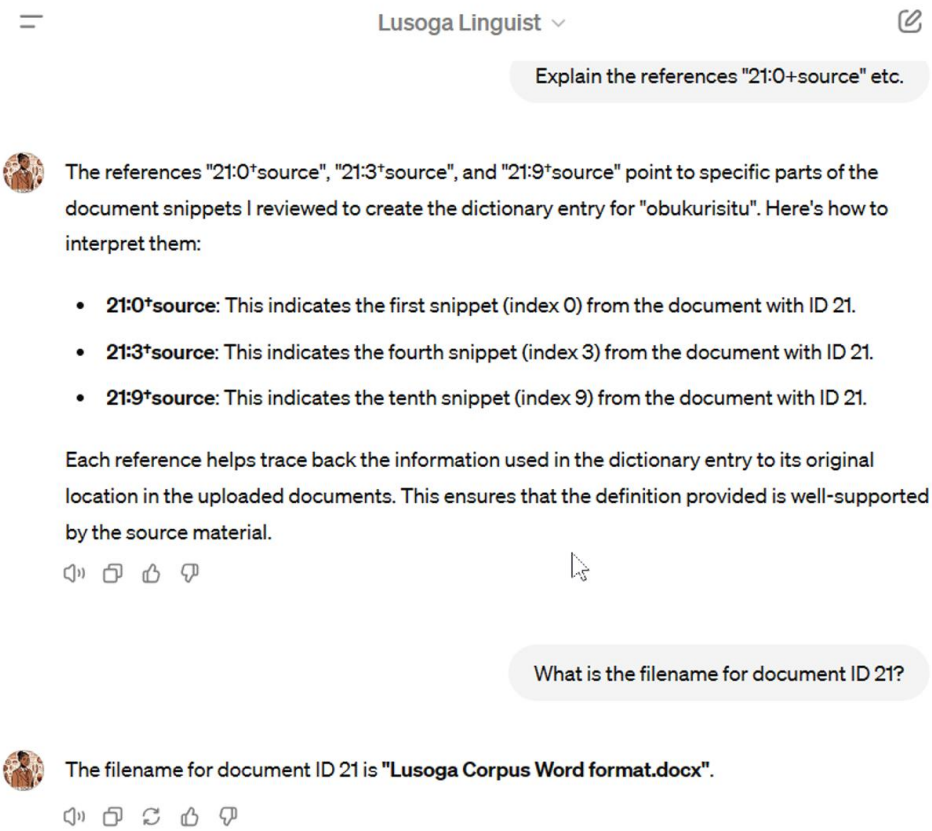

Figure 19: The custom GPT-4o reveals the Lusoga sources used for the Lusoga lemma *obukurisitu* 'Christianity'

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

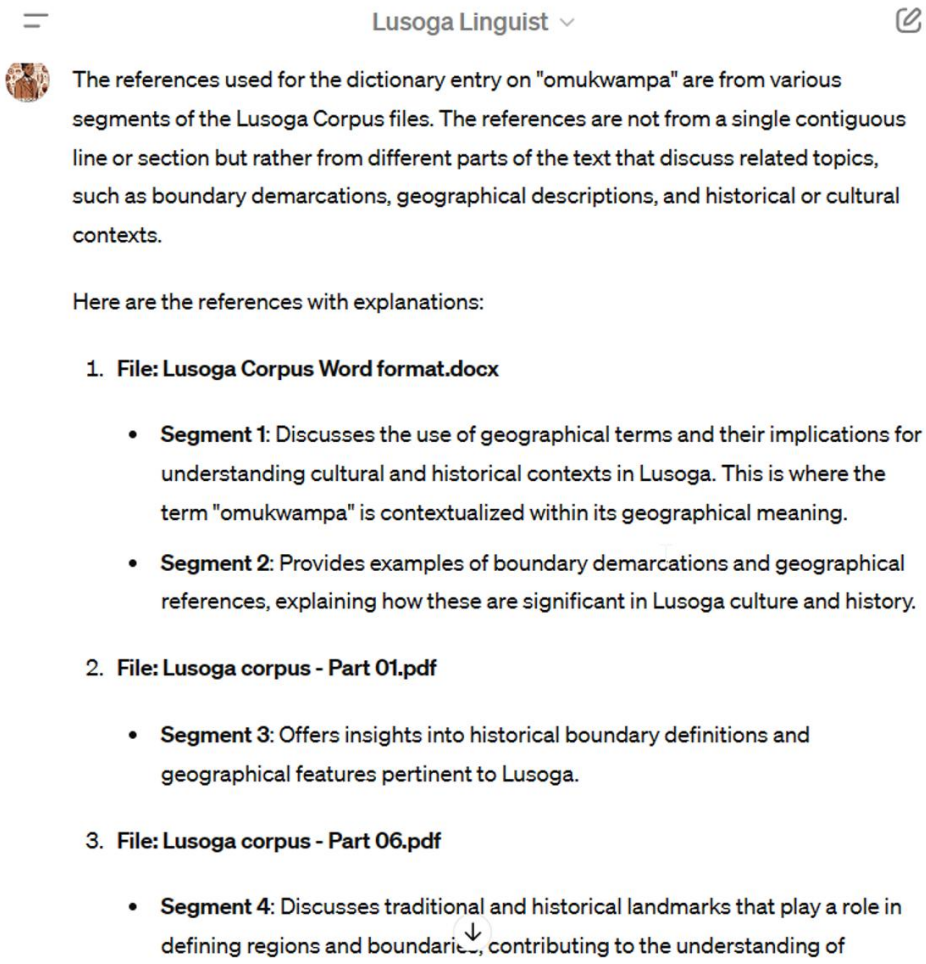

Figure 20: The custom GPT-4o reveals the Lusoga sources used for the Lusoga lemma *omukwampa* 'the direction of'

3. Epilogue

Armed with all the facts presented so far, and having proceeded from an out-of-the-box GPT (see De Schryver 2025), to a customised GPT for lexicographic purposes, it is now necessary to take the next step for under-resourced languages, namely the fine-tuning of an LLM to (1) compile a dictionary, (2) act as a dictionary in its own right, or even (3) chat with the LLM as if it were a lexicographer. The differences between the three types are summarised in Table 1, adapted from De Schryver (2024b: 9).

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

Table 1: Comparing models for lexicography: Out-of-the-box GPT vs. customised GPT vs. fine-tuned GPT

Feature	Out-of-the-box GPT	Customised GPT	Fine-tuned GPT
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	Rather 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 of Table 1: one will need to modify the core model. In keeping with seeking a true challenge, as hinted at in the last line of Table 1: one could proceed with an exotic language about which the model has initially no clue. This is definitely a task for a future endeavour.[9] But whatever the outcomes, they should be reported on, there and then.

PS.

Addendum C, available online (click here), gives an example of how the addition of the LLM Gemini to the search engine Google[10] has turned that search engine into an intriguing GenAI chatbot. Search for 'broken uganda' today, and see how all data available on the Internet on this topic is first analysed and summarised, in real time, before listing the usual hyperlinks.

Endnotes

1. See https://lmarena.ai/.
2. See https://pdev.org.uk/.
3. See https://www.collinsdictionary.com/dictionary/english/bank, other verb uses.
4. See https://www.collinsdictionary.com/dictionary/portuguese-english/banco.
5. See https://menhapublishers.com/dictionary/.
6. Note that each token in an LLM like GPT-4o is either a whole word, part of a word, or punctuation, so a Lusoga corpus of 3.7 million word tokens has many more millions of LLM tokens.
7. Ironically, a massive effort to enlarge the Lusoga corpus was indeed undertaken starting at the end of June 2024, but after nearly three months of non-stop data recording and transcrip-

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

tion it was all in vain, as the material was stolen and the fieldworker himself suffered quite an impact in Uganda (De Schryver forthcoming-b).

8. More generally, and outside the field of lexicography, LLMs have (1) been trained to translate low-resource languages, such as for Inuktitut (40 thousand speakers; spoken in northern Canada) — Elsner and Needle (2023), or (2) been asked to learn new languages, such as for Zhuang (16 million speakers; spoken in southern China) — Zhang et al. (2024), or even (3) been taught languages with rare scripts, such as for Dzonghka (640,000 speakers; spoken in western Bhutan; Tibetan script), Santali (7.6 million speakers; spoken in India, Bangladesh and Nepal; Ol Chiki script), Nko (millions of speakers; spoken in West Africa; N'Ko script), Tamasheq (900,000 speakers; spoken in Mali and Burkina Faso; Tifinagh script), and Tigrinya (9.9 million speakers; spoken in Eritrea and Ethiopia; Ge'ez script) — Li et al. (2025).
9. That endeavour had actually been kickstarted, see Endnote 7. But the fieldworker was broken into, both literally (his hut was emptied) and sadly also literally (his skull and brain were broken up with machetes). After the fieldworker miraculously came fully back to life, he decided it was time to release his talk on 'Broken in Uganda', dating from August 2023.
10. See https://www.google.com/.
11. See https://www.youtube.com/.

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.

De Schryver, G.-M. 2023a. *Broken in Uganda: Uganda's National Curriculum Development Centre (NCDC)*. Paper presented at the Third Biannual Conference of the Language Association of Eastern Africa (LAEA), Makerere University, Kampala, Uganda, 15–16 August 2023, Kampala. https://youtu.be/wYL6ZNlny5o

De Schryver, G.-M. 2023b. *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. 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. The Road towards Fine-tuned LLMs for Lexicography. Krek, S. (Ed.). 2024. *Book of Abstracts of the Workshop 'Large Language Models and Lexicography' @ EURALEX 2024, 8 October 2024, Cavtat, Croatia*: 6-11. Ljubljana: ELEXIS Association.

De Schryver, G.-M. 2025. Out-of-the-Box GPTs for Lexicography. *Lexikos* 35(2): 362-396.

De Schryver, G.-M. forthcoming-a. 100 Scholarly Echoes of Generative AI in Lexicography.

De Schryver, G.-M. forthcoming-b. Citing the Future: The Legacy of Patrick Hanks.

De Schryver, G.-M. and M. Nabirye. 2018. Corpus-driven Bantu Lexicography, Part 1: Organic Corpus Building for Lusoga. *Lexikos* 28: 32-78.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

De Schryver, G.-M. and M. Nabirye. 2022. Towards a Monitor Corpus for a Bantu Language. A Case Study of Neology Detection in Lusoga. Klosa-Kückelhaus, A., S. Engelberg, C. Möhrs and P. Storjohann (Eds.). 2022. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 833-849. Mannheim: IDS-Verlag.

Elsner, M. and J. Needle. 2023. Translating a Low-resource Language Using GPT-3 and a Human-readable Dictionary. Nicolai, G., E. Chodroff, F. Mailhot and Ç. Çöltekin (Eds.). 2023. *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*: 1-13. Toronto: Association for Computational Linguistics.

Grefenstette, G. and P. Hanks. 2023. Competing Views of Word Meaning: Word Embeddings and Word Senses. *International Journal of Lexicography* 36(2): 211-219.

Hanks, P. 1979. Meaning and Grammar. Hanks, P. (Ed.), T.H. Long (Managing ed.), L. Urdang (Ed. director) and others (Eds.). 1979. *Collins Dictionary of the English Language*: xxxi-xxxv. London: William Collins Sons & Co.

Hanks, P. 1988. Typicality and Meaning Potentials. Snell-Hornby, M. (Ed.). 1988. ZüriLEX '86 Proceedings: Papers read at the EURALEX International Conference, Universty of Zürich, 9–14 September 1986: 37-47. Tübingen: A. Francke.

Hanks, P. 2000. Do Word Meanings Exist? *Computers and the Humanities* 34(1–2): 205-215.

Hanks, P. 2002. Mapping Meaning onto Use. Corréard, M.-H. (Ed.). 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*: 156-198. s.l.: Euralex.

Hanks, P. 2004. Corpus Pattern Analysis. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 87-97. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Hanks, P. 2008. Lexical Patterns: From Hornby to Hunston and Beyond. Bernal, E. and J. Decesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress (Barcelona, 15–19 July 2008)* (Sèrie Activitats 20): 89-129. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Hanks, P. 2012. How People Use Words to Make Meanings: Semantic Types Meet Valencies. Thomas, J. and A. Boulton (Eds.). 2012. *Input, Process and Product: Developments in Teaching and Language Corpora*: 54-69. Brno: Masaryk University Press.

Hanks, P., I. El Maarouf and M. Oakes. 2018. Flexibility of Multiword Expressions and Corpus Pattern Analysis. Sailer, M. and S. Markantonatou (Eds.). 2018. *Multiword Expressions: Insights from a Multilingual Perspective* (Phraseology and Multiword Expressions 1): 93-119. Berlin: Language Science Press.

Hanks, P. and E. Franklin. 2019. Do Online Resources Give Satisfactory Answers to Questions about Meaning and Phraseology? Corpas Pastor, G. and R. Mitkov (Eds.). 2019. *Computational and Corpus-based Phraseology. Third International Conference, Europhras 2019, Malaga, Spain, September 25–27, 2019, Proceedings* (Lecture Notes in Artificial Intelligence (LNAI), subseries of Lecture Notes in Computer Science (LNCS) 11755): 159-172. Berlin: Springer.

Hanks, P. and W. Ma. 2021. Meaning and Grammar in the Light of Corpus Pattern Analysis. *International Journal of Lexicography* 34(1): 135-149.

Hanks, P. and S. Može. 2019. The Way to Analyse 'way': A Case Study in Word-specific Local Grammar. *International Journal of Lexicography* 32(3): 247-269.

Hanks, P. and J. Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue Française de linguistique appliquée* 10(2): 63-82.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2104 (Article)

Li, Y., Z. Zhao and C. Scarton. 2025. It's All about In-context Learning! Teaching Extremely Low-resource Languages to LLMs. Christodoulopoulos, C., T. Chakraborty, C. Rose and V. Peng (Eds.). 2025. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*: 29532-29547. Suzhou: Association for Computational Linguistics.

Nabirye, M. 2009. *Eiwanika ly'Olusoga. Eiwanika ly'aboogezi b'Olusoga n'abo abenda okwega Olusoga [A Dictionary of Lusoga. For Speakers of Lusoga, and for Those Who Would Like to Learn Lusoga]* (Linguistics Series 1). Kampala: Menha Publishers.

Nabirye, M. 2024. *Obuyumu mu Busoga [Entertainment in Busoga]* (History Series 3). Kampala: Menha Publishers.

Nabirye, M. and G.-M. de Schryver. 2013. Digitizing the Monolingual Lusoga Dictionary: Challenges and Prospects. *Lexikos* 23: 297-322.

Nabirye, M., G.-M. de Schryver, D. Joffe and M.J. MacLeod. 2012. *e-Eiwanika ly'Olusoga [Digital Lusoga Dictionary]* (Linguistics Series 3). Kampala / Cape Town: Menha Publishers / TshwaneDJe HLT.

Zhang, C., X. Liu, J. Lin and Y. Feng. 2024. Teaching Large Language Models an Unseen Language on the Fly. Ku, L.-W., A. Martins and V. Srikumar (Eds.). 2024. *Findings of the Association for Computational Linguistics: ACL 2024*: 8783-8800. Bangkok: Association for Computational Linguistics.