Generating bilingual example sentences with large language models as lexicography assistants

Raphael Merx **Ekaterina Vylomova** **Kemal Kurniawan**
School of Computing and Information Systems, The University of Melbourne
rmerx@student.unimelb.edu.au
{vylomovae, kurniawan.k}@unimelb.edu.au

Abstract

We present a study of LLMs' performance in generating and rating example sentences for bilingual dictionaries across languages with varying resource levels: French (high-resource), Indonesian (mid-resource), and Tetun (low-resource), with English as the target language. We evaluate the quality of LLM-generated examples against the GDEX (Good Dictionary EXample) criteria: typicality, informativeness, and intelligibility (Kilgarriff et al., 2008). Our findings reveal that while LLMs can generate reasonably good dictionary examples, their performance degrades significantly for lower-resourced languages. We also observe high variability in human preferences for example quality, reflected in low inter-annotator agreement rates. To address this, we demonstrate that in-context learning can successfully align LLMs with individual annotator preferences. Additionally, we explore the use of pre-trained language models for automated rating of examples, finding that sentence perplexity serves as a good proxy for "typicality" and "intelligibility" in higher-resourced languages. Our study also contributes a novel dataset of 600 ratings for LLM-generated sentence pairs, and provides insights into the potential of LLMs in reducing the cost of lexicographic work, particularly for low-resource languages.

1 Introduction

Example sentences in bilingual dictionaries play a crucial role in language learning, helping L2 speakers to understand the meaning of headwords (words that mark a separate entry in the dictionary), and their usage in context (Potgieter, 2012; Nielsen, 2014; Caballero, 2024). What makes candidate sentences good as examples is the subject of linguistic research, with Kilgarriff et al. (2008) proposing the GDEX (Good Dictionary EXample) framework, which qualifies good examples as typical ("exhibiting frequent and well-dispersed patterns of usage"), intelligible ("avoiding gratuitously difficult lexis and structures"), and informative ("helping to elucidate the definition"), as illustrated in Table 1. In bilingual setups, the accuracy of translation between source and target examples also contributes to example quality.

Typical: Show how the word is commonly used.
Yes The business was highly successful, turning a profit in its first year.
No The successful completion of his puzzle took months.
Informative: Provide additional clarity beyond the word definition.
Yes Her marketing campaign was successful, resulting in a 50% increase in sales.
No They were successful.
Intelligible: Easy to understand, not overly complex.
Yes The students were successful in completing their group project on time.
No Notwithstanding the exigencies of the situation, the team's herculean efforts proved successful.

Table 1: GDEX criteria definitions and English example sentences for the word "successful", with one sentence that fulfils the criterion and one that does not.

The extensive work required to come up with example sentences increases the cost of compiling lexicographic resources (He and Yiu, 2022). This has prompted research into the automatic selection of example sentences from existing corpora (Kilgarriff et al., 2008; Frankenberg-Garcia, 2014). However, existing corpora might not always contain sentences that are suited to language learning, as their text can be overly complex, fail to further explain the meaning of the headword, or not be licensed for reproduction. As a result, researchers have begun exploring models tailored for the generation of dictionary example sentences from a headword and its dictionary definition (He and Yiu, 2022).

Large language models (LLMs) trained on a wide range of texts (Gao et al., 2020) might be well suited to formulate generic and informative example sentences that benefit language learning.

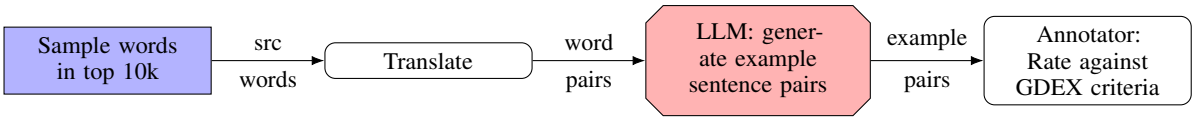

Figure 1: Overview of our process for generating example sentence pairs using LLMs.

In particular, their capacity to adapt to new, unseen tasks (Radford et al., 2019; Kojima et al., 2023) means that they might be well suited to generate sentences against specific criteria. However, questions about the quality of the sentences they generate, and their ability to understand what makes a good example, remain.

In this paper, we review LLMs capability to generate and rate example sentences in a bilingual lexicography context, against the GDEX criteria. We work with three language pairs, with English on the target side, and source sides that cover a range of language resource levels: French (high-resource), Indonesian (mid-resource), Tetun (low-resource). The paper makes the following contributions:

- An evaluation of LLMs capability to *generate* bilingual example sentence pairs, across languages of different resource levels;
- An evaluation of pre-trained models and LLMs capability to *rate* the generated bilingual example pairs, both against the GDEX criteria (qualitative), and against an overall rating (quantitative, 1-5);
- A novel dataset of 600 sentence ratings for LLM-generated example sentence pairs in French, Indonesian, and Tetun as source, and English as target. Each pair is rated against 5 criteria, resulting in 3,000 individual annotations.[1]

2 Background

LLMs for synthetic data generation. While hallucinations can make LLMs unreliable for tasks that require factual accuracy (Azamfirei et al., 2023), the text they generate can be of high quality, in some cases preferred over human-generated text by human annotators (West et al., 2023; Almeman et al., 2024; Cai et al., 2024a). LLM generation of synthetic data has several downstream applications, including the creation of corpora for subsequent training of specialised models (Li et al., 2023; Whitehouse et al., 2023) and the generation of examples to aid learning (Jury et al., 2024; Nam et al., 2024). In lower resource scenarios, LLMs exhibit an increased tendency to generate inaccurate or poor quality information (Cahyawijaya et al., 2024; Benkirane et al., 2024). However, this limitation is not entirely prohibitive; recent research has demonstrated that LLMs can be leveraged to generate synthetic resources when authentic materials are scarce (Santoso et al., 2024). This dual nature of LLMs in low-resource contexts—their proneness to hallucination and their potential for synthetic data generation—presents both challenges and opportunities for their application in bilingual lexicography.

Automated extraction and generation of dictionary examples. The identification, rating, and generation of dictionary examples has been the subject of previous research. Using the GDEX criteria, Almeman and Anke (2022) found that many WordNet examples (Miller, 1995) are of poor quality, often because they are too short, in comparison with those from the Oxford English Dictionary (1989). A subsequent study found that ChatGPT-generated examples are rated higher by human annotators than those from the Oxford Dictionary (Almeman et al., 2024). Cai et al. (2024a) further introduced OxfordEval, an evaluation metric defined as the win rate between generated sentences and the Oxford Dictionary, and found that LLM-generated examples have over 80% win rate. They also introduced the selection of candidate sentences through a masked language model to marginally improve the win rate. In non-English settings, results were found to be more mixed: working with Japanese, Benedetti et al. (2024a) found human examples were still preferred by annotators, with high rates of disagreement between annotators about example quality. In a low-resource setting, working with Singlish, Chow et al. (2024) found that ChatGPT could be leveraged to produce draft dictionary entries, including example sentences, but authors did not rate the examples independently of generated definitions.

[1] https://github.com/raphaelmerx/llm-bilingual-examples

Lang	Src	Tgt	Src sentence	Tgt sentence	GDEX ratings	Overall rating
tdt	rai	country	Timor-Leste mak rai ida ne'ebe iha laran kultura barak.	Timor-Leste is a country rich in culture.	Typical: Yes Informative: Yes Intelligible: Yes Transl. correct: No	3 - Average
ind	meriam	cannon	Meriam itu ditempatkan di atas bukit untuk melindungi kota dari serangan musuh.	The cannon was placed on the hill to protect the city from enemy attacks.	Typical: Yes Informative: Somewhat Intelligible: Yes Transl. correct: Yes	4 - Good
fra	on	we	On va au cinéma ce soir.	We are going to the cinema tonight.	Typical: Yes Informative: Yes Intelligible: Yes Transl. correct: Yes	5 - Very good

Table 2: Example LLM-generated sentences and annotator ratings for languages covered in this study.

Research gap. Despite the growing body of research on LLMs in lexicography, several areas remain unexplored. First, there has been no structured evaluation of LLM capabilities in generating example sentences for bilingual dictionaries, where additional challenges arise compared to monolingual dictionaries, such as maintaining GDEX criteria across languages while ensuring translation accuracy. Second, the potential of LLMs to help assess the quality of examples in a bilingual context, which could assist with example selection and with the setup of self-improvement pipelines for generation, has not been systematically investigated. Lastly, we have not found comprehensive studies examining LLM-based optimisation techniques—such as prompt engineering, fine-tuning, and in-context learning—for the specific task of generating dictionary examples. Addressing these research gaps could advance our understanding of how to effectively harness LLMs for creating high-quality, contextually appropriate example sentences in bilingual dictionaries, across languages of varying resource levels.

3 LLM generation of bilingual example sentences

This section describes our methodology for generating bilingual example sentences using LLMs, and results from human annotation of these generated sentences.

3.1 Methodology for generation

Figure 1 provides an overview of our proposed methodology for generating and rating examples.

Word selection For each source language (French, Indonesian, Tetun), we randomly select 50 words from the top 10,000 most frequent words. We use existing word lists for French[2] and Indonesian,[3] and generate that list for Tetun by finding the top 10,000 words in the Labadain 30k dataset (de Jesus and Nunes, 2024), the largest available Tetun dataset audited by native speakers. We then manually translate each of the 50 words to their English equivalent. When words have multiple translations, we select the one that we deem the most frequent. This results in 50 word pairs for each language pair.

Lang	GPT-4o	Llama3.1	t-stat
fra	**4.79** $\pm$ 0.47	4.57 $\pm$ 0.62	3.06*
ind	4.36 $\pm$ 0.82	**4.46** $\pm$ 0.79	-1.04
tdt	**3.86** $\pm$ 1.18	3.61 $\pm$ 1.22	1.55

Table 3: Average overall rating ($\pm$ standard deviation) for LLM-generated examples per language, with paired t-test results, where * represents a statistically significant difference between models ($p < 0.05$). For rating per criteria, see the distribution bar plot in Figure 2.

Example generation We work with two LLMs, GPT-4o (OpenAI team, 2024) and Llama 3.1 405b (Dubey et al., 2024). The former is the highest rated model overall on the Chatbot Arena as of September 2024 (Chiang et al., 2024), the latter is the highest rated among open weights models. For generating example sentence pairs, we use the OpenAI API[4] for GPT-4o, and the Replicate API[5] for Llama 3.1 405b, using a prompt that describes the GDEX criteria and includes the word pairs, shown in Appendix A.1. Both the source and target

[2] http://www.lexique.org/
[3] FrequencyWords/id_full.txt
[4] https://platform.openai.com/
[5] https://replicate.com/

side sentences are generated jointly in the same output.

Annotator selection and training All annotators are native speakers of the source language they rate, and are advanced speakers of English as a second language. We recruit two annotators per source language, one with a computational linguistics background, and one with no background in linguistics or NLP, to get a broad representation of diverse preferences and expectations. Before annotation, we present the task to each annotator, with for each criterion, an explanation of its meaning, along with an example of a sentence that would be rated "Yes" for this criterion, and an example of a sentence that would be rated "No". We explain to each annotator that the "Overall rating" is left to express their general feeling about example quality.

Annotation We ask annotators to rate the generated examples against the GDEX criteria (typical, informative, intelligible), with three options for each criterion: “Yes”, “Somewhat”, “No”. After initial observations (on French) that generated sentences can have translation errors, we add another column "Translation correct", with the same options. We also include an "Overall rating" column, where annotators are asked to give their overall impression of the example pair quality, on a scale of 1 to 5 (1 - Bad, 2 - Pretty bad, 3 - Average, 4 - Good, 5 - Very good).

3.2 Quality of LLM-generated examples

Table 2 shows an example of LLM-generated sentences for each language pair, with their associated ratings.

Per language Mean overall ratings and annotation distribution are presented in Table 3 and Figure 2 respectively. LLM-generated examples get a medium to high overall rating across language pairs. However, there is a clear drop in quality when language is less-resourced. French examples, representing a high-resource language, received the highest ratings (mean 4.68 out of 5), followed by Indonesian (mid-resource, mean 4.41), and then Tetun (low-resource, mean 3.74). This pattern is consistent with previously observed LLM performance degradation on lower-resourced languages (Li et al., 2024), likely due to the reduced amount of training data available for these languages. For example, the MADALAD-400 corpus (Kudugunta et al., 2023), which has documents from Common Crawl tagged by language, has almost 6 times more French documents (∼220M) than Indonesian documents (∼38M), and over 5,000 times more French documents than Tetun documents (∼40k).

Lang	A1	A2	t-stat
fra	4.74 ± 0.56	4.62 ± 0.56	1.830
ind	4.09 ± 0.85	4.73 ± 0.62	−6.273*
tdt	3.62 ± 1.47	3.85 ± 0.88	−1.909

Table 4: Average rating (± standard deviation) per annotator with paired t-test results, where * represents a statistically significant difference between annotators ($p < 0.05$). For each language, A1 is the annotator with a computational linguistics background.

Per LLM Comparing overall rating for the two LLMs used in the study, we find that GPT-4o outperforms Llama3.1 for French (4.79 vs. 4.57), with a statistically significant t-statistic of over 3 indicating a substantial difference between the two models relative to variation in the data. For Indonesian and Tetun however, the paired t-test indicated that the difference between the two models is not statistically significant compared to the variation in the data. We therefore observe variability in LLM output quality that is uneven across languages depending on resource level and shows that performance degradation is not always predictable from resource level.

Per GDEX criteria Comparing qualitative ratings (typical / intelligible / informative / translation correct), we find a consistent degradation across criteria as the resource level of the language decreased (Figure 2). For example, 95% of examples are rated as "typical" for French, but this decreased to 92% for Indonesian and 69% for Tetun. The trend was particularly pronounced for the "Informative" criterion (fra 95%, ind 77%, tdt 56%), highlighting the challenges LLMs face in maintaining accurate and relevant examples for lower-resourced languages.

Per annotator qualification level Table 4 shows no significant difference in mean ratings between annotators for French and Tetun relative to variation in the data, when measured through a paired t-test. For Indonesian, however, we observe a significant and large difference in mean ratings between annotators, where A1 (the annotator with a computational linguistics background) gave much

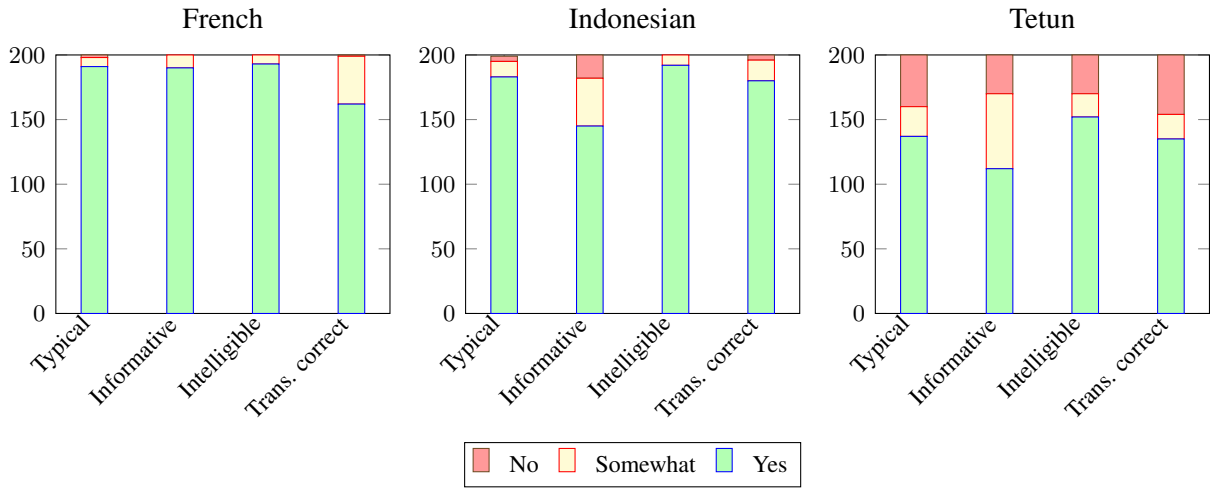

Figure 2: Rating distributions (GPT-4o and Llama 3.1 combined) for GDEX criteria and translation correctness.

Lang	Criteria	Krippendorff's α
fra	Typical	**0.378**
	Informative	–0.047
	Intelligible	0.264
	Translation correct	0.136
	Overall rating	0.136
ind	Typical	**0.517**
	Informative	–0.269
	Intelligible	–0.036
	Translation correct	–0.093
	Overall rating	–0.093
tdt	Typical	**0.548**
	Informative	**0.449**
	Intelligible	**0.519**
	Translation correct	**0.529**
	Overall rating	**0.529**

Table 5: Inter-annotator agreement measured using Krippendorff's alpha for different GDEX criteria and overall rating. Bold indicates $\alpha > 0.35$.

lower ratings than A2.

3.3 A note on inter-annotator agreement

Table 5 shows relatively low rates of inter-annotator agreement for French and Indonesian, measured through Krippendorff's alpha (Castro, 2017), both for overall rating (where individual judgement is encouraged) and for qualitative GDEX criteria (where standard rating is encouraged). For Tetun, however, we observe relatively high inter-annotator agreement across all criteria, including overall rating. We hypothesise that this is due to the more pronounced mistakes in Tetun sentences, which means both that ratings rely less on subtlety of judgement, and that there is more signal to measure. For example, in French, all GDEX criteria are rated "Yes" in over 95% of examples, giving little room to measure disagreement.

We note that low inter-annotator agreement for rating examples was observed in previous studies (Benedetti et al., 2024b). This finding guides our further experiments: (1) when working with in-context learning, we favour aligning LLM rating with one annotator's judgement at a time, rather than aligning with contradicting ratings coming from multiple annotators (Section 4.1); (2) when working with pre-trained language models, which are not fine-tuned to annotator preference, we only measure alignment with the annotator who has a computational linguistics background (Section 4.2).

4 Automated rating of example sentences

Beyond baseline performance across different resource levels, we evaluate how well LLMs can assess example quality. This could enable more efficient dictionary creation pipelines, where automated rating systems that align with human judgement could help filter and select the best examples from larger sets of generated candidates, reducing the need for extensive manual review. Furthermore, reliable automated evaluation metrics could facilitate the development of self-improvement systems where LLMs learn from their own assessments to generate increasingly better examples.

4.1 Rating through LLM in-context learning

For each annotator, we study whether in-context learning can successfully teach the annotator's preferences to an LLM, measured through alignment in overall rating (1-5 score).

Lang	Annotator	Rating correl.
fra	A1-fra	0.54
	A2-fra	0.38
ind	A1-ind	0.33
	A2-ind	0.29
tdt	A1-tdt	0.39
	A2-tdt	0.42

Table 6: Correlation between LLM predicted rating and annotator reference rating (both 1-5) with 10 in-context examples of the annotator's ratings. All correlations are statistically significant with $p < 0.02$.

Data preparation and model choice Given 100 annotated example sentence pairs from a specific annotator, we randomly sample 10 pairs as in-context examples and 90 pairs for evaluation. To avoid bias linked to model self-preference (Panickssery et al., 2024), we choose against working with one of the two LLMs used for generating sentences and instead rely on Gemini 1.5 Pro (Gemini Team, 2024) for this task, given that it is the second best ranked model for instruction following on the Chatbot Arena[6] as of September 2024.

Preprocessing through reasoning generation For each sentence pair in the sample of 10 pairs, we first ask the LLM to reason about what led to the annotator's rating, given their comment (if any), their ratings of the GDEX criteria, and the translation correctness. Our prompt for this task is provided in Appendix A.2.

Evaluation We then construct a system prompt that has a list of 10 examples, each with a word and example sentence pair, a reasoning, and final rating from 1 to 5. These examples are injected in the prompt, along with a description of the GDEX criteria (Appendix A.3). We use this prompt to ask for a rating for the evaluation of example pairs.

Results Table 6 demonstrates that in-context learning successfully teaches LLMs annotator preferences across all participants, yielding moderate but significant correlations ranging from 0.29 (A2-ind) to 0.54 (A1-fra). These results span languages of varying resource levels and annotators with diverse backgrounds, highlighting the potential of in-context learning to address challenges related to inter-annotator agreement.

[6]https://lmarena.ai/

4.2 Rating through pre-trained language models

In this section, we aim to determine if computationally derived metrics can effectively approximate human judgements of example sentence quality along GDEX criteria.

Data preparation We work exclusively with ratings from annotators who have a background in computational linguistics. We map each rating to a number between 0 and 1, where No = 0, Somewhat = 0.5, Yes = 1, allowing us to represent the gradations in quality along a continuous scale.

Metrics and hopythesis For each source-side sentence, we compute several metrics using pre-trained language models to test various hypotheses. We examine whether the probability of the entry word (when masked) can serve as a predictor of the "Informative" rating, hypothesising that a lower probability might indicate a more informative context. We also investigate if sentence perplexity can be a good predictor of both the "Intelligible" and "Typical" ratings, with the assumption that lower perplexity could indicate a more intelligible and typical sentence. Additionally, we explore whether context entropy at the position of the entry word could be another predictor of the "Informative" rating, positing that higher entropy might suggest a more informative context.

Choice of models To test the hypotheses, we use pre-trained encoder-only language models: `CamemBERT-large` for French (Martin et al., 2019), `IndoBERT` for Indonesian (Koto et al., 2020). For Tetun, given the absence of existing encoder-only models for the language, we fine-tune `XLM-RoBERTa-large` (Conneau et al., 2019) on MADLAD-400 (Kudugunta et al., 2023) which is the largest Tetun monolingual corpus available, using the hyperparameters in Adelani et al. (2021). We release the weights of this model for future researchers.[7]

Results As Table 7 demonstrates, the probability of the target word serves as a fair predictor of informativeness for French, with a correlation of 0.21, but this relationship does not hold for other languages. High perplexity proves to be a moderately good predictor of low intelligibility for both French and Indonesian, with correlations of -0.57

[7]https://huggingface.co/raphaelmerx/xlm-roberta-large-tetun

Lang	Criterion	LM Metric	Correl.
fra	Informative	Word Prob.	0.210*
	Intelligible	Perplexity	−0.566*
	Typical	Perplexity	−0.408*
	Informative	Entropy	0.062
ind	Informative	Word Prob.	0.176
	Intelligible	Perplexity	−0.521*
	Typical	Perplexity	−0.320*
	Informative	Entropy	0.124
tdt	Informative	Word Prob.	0.113
	Intelligible	Perplexity	0.101
	Typical	Perplexity	0.136
	Informative	Entropy	0.068

Table 7: Correlation between GDEX ratings and masked LM metrics. * denotes statistical significant with $p < 0.05$.

and -0.52 respectively. Similarly, high perplexity is a good predictor of low typicality for French (correlation of -0.41) and moderately good for Indonesian (-0.32). Notably, no significant correlations are found for Tetun across these metrics. Contrary to our hypothesis, context entropy at the target word (when masked) does not serve as a good predictor for informativeness across any of the languages studied.

Implications Our results show the potential of sentence perplexity for estimating example sentence typicality and intelligibility, for middle- to high-resource languages. The lack of significant results for Tetun demonstrates that the amount of available corpora in this low-resource language is not sufficient to get a pre-trained language model that captures sentence quality with a high degree of accuracy.

5 Discussion

Our study provides several insights into the capabilities and limitations of LLMs for generating and evaluating bilingual dictionary examples. First, we demonstrate that LLMs are capable of producing reasonably good quality example sentences across multiple language pairs. However, there is a clear degradation in performance as we move from high-resource languages like French to low-resource languages like Tetun. The variability in output quality across languages underscores the need for careful evaluation and potential supplementary techniques when applying LLMs to lexicographic tasks, especially for less-represented languages.

A notable challenge revealed in our study is the high variance in personal preferences for example sentence quality, as evidenced by low inter-annotator agreement rates. This variability poses difficulties in establishing a single, universally accepted metric for evaluating dictionary examples. However, our experiments with in-context learning demonstrate that LLMs can be successfully aligned with individual annotator preferences, even for low-resource languages like Tetun. This finding suggests a promising avenue for tailoring LLM outputs to specific lexicographic standards or individual annotator judgements, potentially facilitating the example generation and evaluation process.

The low inter-annotator agreement observed in our study highlights the need for annotations from multiple annotators before drawing conclusions about the quality (or lack thereof) of example sentences. This multi-annotator approach can help capture a more comprehensive range of perspectives and mitigate individual biases. Additionally, our findings, particularly for French where most GDEX criteria were rated "Yes" due to the high quality of generated sentences, suggest the need for finer measures of criteria to better capture nuanced levels of quality. We recommend developing more granular rating scales or additional sub-criteria, especially for high-resource languages where LLMs perform well. This refinement in evaluation methods could provide more discriminative assessments of LLM-generated example sentences.

6 Conclusion

We contribute a first evaluation of LLM capability to generate bilingual example sentences, across languages of various resource levels. We show that although LLMs are capable of generating good bilingual example sentences on average, their performance degrades with language resource level. We further show that even when using a shared framework for sentence evaluation (GDEX), annotators tend to disagree with each other on sentence quality, but that in-context learning can be leveraged to align LLMs with a specific annotator's ratings.

Our findings highlight the potential of LLMs in lowering the cost of lexicographic work, and their ability in aligning with human judgement in a field where human judgement can be highly variable.

This is of particular value in low-resource lexicographic work, where lack of human resources may prevent the widespread compilation of lexicographic resources.

Limitations

While our study shows LLMs can play a helpful role in the generation and rating of bilingual dictionary examples, our choice of experiment constraints can limit the reach of our results. We work exclusively with languages that use Latin script, and with English on the target side, which raises the question of how our results would hold for languages that use other scripts and with lower-resource target languages. We did not include part of speech information when generating examples, and do not study performance on words that have several definitions; both choices may have skewed the quality of generated example downwards.

The low inter-annotator agreement, while part of the experiment, and expected in this lexicographic context, raises questions about how we could have better aligned annotators, for example by using pre-qualifying questions, or by exclusively relying on linguists for annotation.

We identify several areas for future work. First, LLM rating of example sentences could be integrated in the example generation pipeline, for instance by having an LLM generate a number of candidate examples, and another LLM automatically rank them, similar to the approach by Cai et al. (2024b). Second, the quality of LLM-generated example sentences could be compared against sentences retrieved from a corpus. Last, the incorporation of retrieved sentences in the LLM prompt could guide the LLM to generate more typical or informative sentences.

Acknowledgments

We would like to extend our gratitude to Professor Hanna Suominen for her valuable feedback and guidance throughout this study. We also thank Gabriel de Jesus and his team, Isabel Pereira (Catalpa International), Tungga Dewi, and Matilda Merx for their support in data collection and annotation. We also thank the anonymous reviewers for their feedback. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131. Place: Cambridge, MA Publisher: MIT Press.

Fatemah Almeman and Luis Espinosa Anke. 2022. Putting wordnet's dictionary examples in the context of definition modelling: An empirical analysis. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 42–48.

Fatemah Yousef Almeman, Steven Schockaert, and Luis Espinosa Anke. 2024. WordNet under scrutiny: Dictionary examples in the era of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17683–17695.

Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):120.

Enrico Benedetti, Akiko Aizawa, and Florian Boudin. 2024a. Automatically suggesting diverse example sentences for l2 japanese learners using pre-trained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 114–131.

Enrico Benedetti, Akiko Aizawa, and Florian Boudin. 2024b. Automatically Suggesting Diverse Example Sentences for L2 Japanese Learners Using Pre-Trained Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 114–131.

Kenza Benkirane, Laura Gongas, Shahar Pelles, Naomi Fuchs, Joshua Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Eduardo Sánchez. 2024. Machine translation hallucination detection for low and high resource languages using large language models. *Preprint*, arXiv:2407.16470.

Alfonso Rascón Caballero. 2024. *The theory and practice of examples in bilingual dictionaries*, volume 165. Walter de Gruyter GmbH & Co KG.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433.

Bill Cai, Ng Clarence, Daniel Liang, and Shelvia Hotama. 2024a. Low-cost generation and evaluation of dictionary example sentences. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3538–3549.

Bill Cai, Ng Clarence, Daniel Liang, and Shelvia Hotama. 2024b. Low-Cost Generation and Evaluation of Dictionary Example Sentences. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3538–3549.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. `https://github.com/pln-fing-udelar/fast-krippendorff`.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Siew Yeng Chow, Chang-Uk Shin, and Francis Bond. 2024. This word mean what: Constructing a Singlish dictionary with ChatGPT. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 41–50.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Gabriel de Jesus and Sérgio Nunes. 2024. Labadain-30k+: A Monolingual Tetun Document-Level Audited Dataset. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 177–188.

Oxford English Dictionary. 1989. Oxford english dictionary. *Simpson, Ja & Weiner, Esc*, 3.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ana Frankenberg-Garcia. 2014. The use of corpus examples for language comprehension and production. *ReCALL*, 26(2):128–146.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Xingwei He and Siu Ming Yiu. 2022. Controllable Dictionary Example Generation: Generating Example Sentences for Specific Targeted Audiences. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–627.

Breanna Jury, Angela Lorusso, Juho Leinonen, Paul Denny, and Andrew Luxton-Reilly. 2024. Evaluating llm-generated worked examples in an introductory programming course. In *Proceedings of the 26th Australasian Computing Education Conference*, pages 77–86.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1, pages 425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra ....

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. In *Proceedings of the 28th COLING*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *Preprint*, arXiv:2310.07849.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *Preprint*, arXiv:2404.11553.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Sandro Nielsen. 2014. Example sentences in bilingual specialised dictionaries assisting communication in a foreign language. *Lexikos*, 24(1):198–213.

OpenAI team. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Liezl Potgieter. 2012. Example sentences in bilingual school dictionaries. *Lexikos*, 22:261–271.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Joan Santoso, Patrick Sutanto, Billy Cahyadi, and Esther Setiawan. 2024. Pushing the limits of low-resource NER using LLM artificial data generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9652–9667.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2023. The generative ai paradox: "what it can create, it may not understand". *Preprint*, arXiv:2311.00059.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced cross-lingual performance. *arXiv preprint arXiv:2305.14288*.

A Prompts used

In the prompts below, the parts in brackets (e.g. `{SRC_NAME}`) are templated out.

A.1 Generating examples

```
You are assisting in the creation of a
bilingual {SRC_NAME}-{TGT_NAME}
dictionary. Your task is to generate
example sentences for dictionary entries
 to help users understand the usage of
words in context.

You will be provided with a {SRC_NAME}
word and its {TGT_NAME} equivalent.
<{SRC_NAME} entry>
{{src_word}}
</{SRC_NAME} entry>

<{TGT_NAME} entry>
{{tgt_word}}
</{TGT_NAME} entry>

Please create a pair of example
sentences for each entry. The sentences
should be:
1. Typical: Show typical usage of the
word
2. Informative: Add value by providing
context or additional information
3. Intelligible: Be clear, concise, and
appropriate for a general audience
4. Using the entries provided above (the
 {SRC_NAME} and {TGT_NAME} words)

Format your response as follows:

<example_sentence_pair>
{SRC_NAME}: [Your {SRC_NAME} sentence
here]
{TGT_NAME}: [Your {TGT_NAME} sentence
here]
</example_sentence_pair>

Please provide your example sentences
based on the given {SRC_NAME} and {
TGT_NAME} entries.
```

A.2 Reasoning about a specific annotator's rating

```
<example>
Src Entry: {src_entry}
Tgt Entry: {tgt_entry}
Src Example: {src_example}
Tgt Example: {tgt_example}

Comment: {comment}
Typical: {typical}
Informative: {informative}
Intelligible: {intelligible}
Translation correct: {
translation_correct}
</example>

Reasoning: what is the reasoning for the
 above ratings? Give your response in
one paragraph.
```

A.3 In-context learning for aligning an LLM with an annotator

A.3.1 Prompt construction

```
TEMPLATE_EXAMPLE = """<example>
<data>
Src Entry: {src_entry}
Tgt Entry: {tgt_entry}
Src Example: {src_example}
Tgt Example: {tgt_example}
</data>
<reasoning>{reasoning}</reasoning>
<rating>{rating}</rating>
</example>"""

def get_templated_example(row):
    return TEMPLATE_EXAMPLE.format(
        src_entry=row[SRC_LANG],
        tgt_entry=row[TGT_LANG],
        src_example=row['src_example'],
        tgt_example=row['tgt_example'],
        reasoning=row['reasoning'],
        rating=row['Overall␣rating']
    )

AUGMENTED_SYSTEM_PROMPT = SYSTEM
for row in sample:
    AUGMENTED_SYSTEM_PROMPT +=
get_templated_example(row)
    AUGMENTED_SYSTEM_PROMPT += '\n\n'
```

A.3.2 Prompt example

An example constructed prompt with two examples. Note that our experiments used 10 examples.

```
You are assisting in the creation of a
bilingual English-Indonesian dictionary.
Your task is to rate a candidate
sentence pair that illustrates
dictionary entries to help linguists
select an appropriate example pair.

Example sentences should should be:
1. Typical: Show typical usage of the
word
2. Informative: Add value by providing
context or additional information
3. Intelligible: Be clear, concise, and
appropriate for a general audience
4. Translation correct: Are sentences a
good translation of each other, with
fluent grammar and correct usage of
words in both languages

You are rating the example sentences,
not the dictionary entries.

<example>
<data>
Src Entry: meriam
Tgt Entry: cannon
Src Example: Meriam itu ditempatkan di
atas bukit untuk melindungi kota dari
serangan musuh.
Tgt Example: The cannon was placed on
the hill to protect the city from enemy
attacks.
</data>
<reasoning>The example sentences are
typical because they demonstrate a
standard use of the word "cannon" in a
military context.  However, they are
only somewhat informative because the
statement about cannons being used for
defense, while not entirely inaccurate,
might not be the most common
understanding. The sentences are
intelligible due to their clear and
concise language, and the translation is
 accurate, reflecting the meaning and
grammar of both the source and target
languages.
</reasoning>
<rating>4 Good</rating>
</example>

<example>
<data>
Src Entry: menanyai
Tgt Entry: question
Src Example: Polisi menanyai saksi mata
untuk memperoleh informasi lebih lanjut
tentang kejadian itu.
Tgt Example: The police questioned the
eyewitness to obtain more information
about the incident.
</data>
<reasoning>The ratings are justified
because the sentences demonstrate
typical usage of the words "menanyai"
and "questioned" in the context of a
police investigation. They are
informative by providing context about
the purpose of the questioning. Both
sentences are clear and concise, making
them intelligible. However, the
translation is slightly off because "
keterangan" would be a more natural
choice than "informasi" in Indonesian,
making the translation somewhat less
accurate.
</reasoning>
<rating>4 Good</rating>
</example>

...

<data>
Src Entry: sehari-hari
Tgt Entry: everyday
Src Example: Saya menggunakan sepeda
sebagai alat transportasi sehari-hari
karena lebih ramah lingkungan.
Tgt Example: I use a bicycle as my
everyday mode of transportation because
it's more environmentally friendly.
</data>
```