INNOVATION IN LANGUAGE LEARNING International Conference

Text-Based Glossary Generation by ChatGPT for FL Learners

Iglika Nikolova-Stoupak[1], Gaël Lejeune[2], Eva Schaeffer-Lacroix[3]

Sorbonne University, France

Abstract

Vocabulary acquisition, a key aspect of foreign language (FL) learning, proves challenging to learners for a number of reasons, ranging from a lack of shared vocabulary with their native language to personal constraints relating to time management and memory profiles. Vocabulary is learnt more easily and efficiently when encountered in natural context rather than in isolation [5, 14]. Learning materials such as graded readers, which are proven to have a strong positive effect on a learner's vocabulary skills [5], typically feature glosses, i.e. definitions or translations of the words or expressions that are most complex for the proficiency level at hand and/or most likely to be useful outside the framework of the given text. In an attempt to relieve FL teaching professionals in terms of time, effort and finances, the present study seeks to define and test the best practices for automatic generation of glossaries based on learner-friendly reading materials. In the process, the glosses in a selected corpus of adapted literary texts are investigated. Then, the state-of-the-art ChatGPT large language model (LLM) as per its 4o version is asked to assemble glossaries based on the first chapter of Lewis Carroll's Alice's Adventures in Wonderland *(1865) – in its original version as well as professional translations into Japanese and Bulgarian. The output is optimised through the inclusion of concrete extracts from the corpus as well as the use of quantitative and qualitative observations pertaining to the investigated professional glosses.*

Keywords: *large language models (LLMs), lexicography, multilingual glossaries, vocabulary acquisition*

1. Introduction

Vocabulary acquisition is one of the main aspects of foreign language learning. It is especially challenging due to a variety of factors, such as a lack of shared vocabulary with a learner's first language (L1) and personal constraints related to time management and memory profiles. Different solutions have been suggested in order to at least partially overcome said limitations. It has been established that it is easier for vocabulary items to be acquired when they appear in a natural context [5, 15]. As a result, 'extensive reading' is a commonly suggested practice for the improvement of vocabulary knowledge. Specific book series, known as 'graded readers', have been especially conceived to maximise the process of unintentional vocabulary acquisition. Popular series of the type include Oxford Bookworms, Cambridge English Readers and Heinemann Guided Readers, and they commonly feature adaptations of famous literary texts. Their difficulty, such as in terms of levels as per the Common European Framework of Reference for Languages (CEFR), is typically indicated and defined as a number of 'headwords' (the vocabulary knowledge that a reader needs to possess in order to read the text). Complex words are few in number and density in graded readers.

However, although graded readers are highly motivating for students [4], their simple use does not straightforwardly lead to fast and efficient vocabulary acquisition. Wan-a-rom notes that only around 80% of the word families associated with the initial proficiency level tend to be included in a single book of the level, therefore necessitating the reading of a high number of texts [20]. Thongsan further confirms this hypothesis by demonstrating through an experiment that, on average, only three new words are retained by students for an hour of independent reading [18]. Also, Waring and Takaki note that graded texts are truly efficient when it comes to affirming already known vocabulary, but much less so when new items are concerned [20]. Even more worrisomely, inferences made by readers may easily be wrong if an item is not encountered a certain number of times [8].

The mentioned constraints of extensive reading can be significantly mitigated when a teacher mediates the reading process, ensuring that the learners have acquired the necessary vocabulary for a given proficiency level and providing additional discussions and exercises that help re-encounter and affirm the new words and phrases involved. Hill notes that promising educational practices of the kind have been conducted in Malaysia and several East African countries, offering positive results in terms of the students' interest in the associated stories and the established sense of community [6].

Furthermore, intentional vocabulary acquisition can be incorporated in the reading process through the use of glosses i.e. supplementary lexical information about vocabulary items, such as definitions, translations or illustrations. For instance, as the vocabulary for a given level in graded readers typically covers 95% of the words in a text [9], the remaining ones are often further explained or translated in glossaries, footnotes or other related formats. Conducting an experiment with participant groups that either simply read extensively or utilise glosses, Teng proves the marked benefits of the latter practice [17]. Glosses are addressed by the field of lexicography, which pertains to dictionary compilation and related theoretical work and is narrowly related to linguistics as well as information science [3]. The question of how many and which vocabulary items to include as glosses in a given text is a major one. Common answers include: the most complex words and expressions, the most important ones for understanding the given text, and the most useful ones outside the framework of the text. Certain parts of speech (POS) may also need to be prioritised (for instance, Nation notes that nouns and adjectives may be easiest to guess from the context alone [12]). Glosses may be monolingual or bilingual. Whilst the former provide an additional source of language practice, the latter tend to be preferred by learners [2, 16]. Concerning the impact of the placement of glosses on the learning process, it appears to be much smaller than the actual information presented [7], although there is evidence that a gloss is more effective when physically close to the respective item it describes [1].

2. Analysed Literary Texts

The corpus that will be analysed and used in this study consists of ten professionally abridged texts in four different languages (English, Spanish, French and Italian), all of which contain glosses. The texts are versions of literary works that are commonly included in educational curricula and appreciated by young readers. Some of the texts are explicitly intended for an audience of language learners, whilst others' audiences are undefined or defined as 'children' (book 2) or 'people with reading difficulties' (books 4, 6, 7 and 9). The most represented series are Oxford Bookworms (by Oxford University Press) and Kalafate (by Almadraba). Book 1 is specifically composed for native speakers of Polish. Only a segment of Book 2 is included (due to accessibility constraints). For detailed information about the texts, please refer to Table 1.

#	title	total words	lang.	author (original)	author (adaptation)	level	publisher	series	year
1	*A Christmas Carol*	6153	EN	Charles Dickens	Patric Lagendijk	-	Mediasat Poland Bis	Czytamyw oryginale	2004
2	*Don Chisciotte della Mancia*	2994	IT	Miguel de Cervantes	Alessandro Massobrio	-	Edisco	I coriandoli	2005
3	*Les Misérables*	9931	FR	Victor Hugo	Brigitte Faucard-Martinez	2/6	CLE International	Lectures CLE en français facile	2016
4	*Madame Bovary*	16529	SP	Gustave Flaubert	Clàudia Sabater Baudet	-	La mar de fàcil	-	2019
5	*Oliver Twist*	27071	EN	Charles Dickens	Richard Rogers	6/6[1]	Oxford UP	Oxford Bookworms	2007
6	*Oliver Twist*	20750	SP	Charles Dickens	N. Martí Constans	-	Almadraba	Kalafate	2016
7	*Romeo y Julieta*	16386	SP	William Sharespeare	Elena O'Callaghan i Duch	-	Almadraba	Kalafate	2009
8	*The Picture of Dorian Gray*	10958	EN	Oscar Wilde	Jill Nevile	3/6[2]	Oxford UP	Oxford Bookworms	2007
9	*Tom Sawyer*	16790	SP	Mark Twain	M. Angels Rubio	-	Almadraba	Kalafate	2005
10	*Tom Sawyer*	6284	EN	Mark Twain	Nick Bullard	2/6[3]	Oxford UP	Oxford Bookworms	2000

[1] 2500 headwords

[2]1000 head words
[3] 400 headwords
Table 1. Investigated abridged texts with glosses: general information

The text that ChatGPT is asked to generate a glossary for is the first chapter of *Alice in Wonderland* (Lewis Carroll, 1865). Professional translations of the text are used for the experiments with Bulgarian and Japanese text (L. Goldman, 1996, Vechnite detski romani; H. Yamagata, 2000, Asahi). The selected textual length is limited to one that the model is comfortable working with as well as is suitable for an in-class educational setting. The level of the first chapter of *Alice in Wonderland* is estimated with the help of available online tools: *Road to Grammar*[1], *Cathoven*[2] and *Poodll*[3]. Due to the last two tools' size limitations, it is approximately 80% and 25% of the text that are, respectively, used in level estimation. In its analysis, *Road to Grammar* utilises word length, sentence length and comparison against a frequency list. *Cathoven* uses a machine learning algorithm that involves sentence structure, tenses and vocabulary as categorised into 10 sublevels in addition to the main CEFR levels. *Poodll*, whose main applicability is the evaluation of student writing, is less specific about its inner workings. The resulting levels for the chapter are A2.7, B1 and B1; an outcome that makes it safe to accept B1 as a reasonable estimation for the text's level. Due to the absence of equivalent tools for the estimation of text in Bulgarian and Japanese, the English text's level is also taken as an approximation of the selected text's level in these languages. In the one-shot generation scenario, the model is also provided with an example of an English-language text from the investigated corpus that belongs to a similar proficiency level. *The Picture of Dorian Gray* is selected as the best candidate. The first two chapters of the text are utilised in order to best approximate the size of *Alice in Wonderland*'s first chapter.

3. Methods

3.1 Analysis of Professional Glosses

The glosses in the ten investigated professional adaptations are described both qualitatively and quantitatively. The associated qualitative characteristics include their placement, outlook and content. The quantitative analysis is performed using common Natural Language Processing (NLP) tools as provided in the Python programming languages, such as *spaCy*[4] and *stanza*[5]. It includes the number of glosses and their relation to the total number of words in the respective text, their parts of speech, their density within the text and the frequency ranks of single-word items based on frequency lists in the texts' languages[6]. The results of the analysis serve in the definition of a gold standard that is then used to evaluate the quality of automatically generated glosses.

3.2 Glossary Generation by ChatGPT

The practical experiment associated with the current study consists in the automatic generation of glosses in the form of separate glossaries by OpenAI's model, ChatGPT (in its GPT-4o version), whose text generation abilities are state-of-the-art [13]. The experiment is performed with reference to three languages from discrete language families and with different levels of resourcedness: English, Bulgarian and Japanese. In addition, two generation scenarios are utilised: zero-shot and one-shot. The prompt for the former consists in the text "Please generate a glossary for the following text. Make it suitable for learners of English/Japanese/Bulgarian as a foreign language whose current level is B1"; followed by the sample literary text (the first chapter of *Alice in Wonderland*). In the one-shot scenario, the following text is also added: "This is an example of a text followed by a good glossary:", along with the selected example of a text with an associated professionally-crafted glossary (the first two chapters of the adaptation of *The Picture of Dorian Gray*).

[1] http://www.roadtogrammar.com/textanalysis/
[2] https://nexthub.cathoven.com/cefr
[3] https://poodll.com/cefr-level-checker/
[4] https://pypi.org/project/spacy/
[5] https://pypi.org/project/spacy-stanza/
[6] taken from the *OpenSubtitles* corpus and available at: https://github.com/hermitdave/FrequencyWords

The issuing glossaries are then described using the criteria elaborated in section 3.1. Frequency lists from the same source are utilised for Bulgarian and Japanese text. Japanese text is tokenised into words using the *janome*[7] tool.

3.3 Additional Prompt Engineering

A final experiment that results in a ready-to-use teaching material is conducted. The qualities of the two glossaries output by ChatGPT for each prompting scenario and per each of the three languages are juxtaposed and the better generation scenario is thus defined. Then, the better English-language glossary along with the initial text are once again provided to the model, and it is prompted to improve the glossary based on any discovered shortcomings.

4. Analysis and Experimentation Results

4.1 Professional Glosses

Please see Table 2 for the detailed results of the analysis.
Book 1 is a clear outlier due to the high number of glosses it contains (345). Books 8 and 10 are the only ones that do not feature as glosses vocabulary items outside the 50k-word frequency lists. English texts tend to include glosses for more frequently used items compared to the texts in other languages. However, due to the language's high association with zero derivation (the tendency of words to exist as different parts of speech, unchanged), some limitations of the utilised NLP tools arise; e.g. the noun 'will' being glossed in book 5 but the related auxiliary verb considered in frequency ranking. There are no significant differences in terms of the subdivision of the lists of glossed items by parts of speech when it comes to the different languages. Glosses appear roughly equally often within the text and as separate glossaries, and monolingual definitions make up their most common content.

#	glosses[1]	highest / lowest/ av. frequency rank	# words outside of frequency list[2]	min/ max /av. density[3]	% n./ v./ adj. / expressions/ cultural items[4]	placement	description
1	345 (5.61%)	125/48221/ 9602.91	13	1/14/2.15	36.52/16.23/19.71/ 22.32/0	glossary at the end of book	translation into Polish; illustrations (15)
2	32 (1.07%)	3690/45786/ 26598.57	13	1/11/2.93	50/6.25/18.75/ 28.13/12.5	footnotes	item followed by definition[5]
3	41 (0.41%)	1199/40875/ 17680.38	10	1/16/2.68	60.98/9.76/9.76/ 14.63/0	footnotes	item followed by definition; separate list of 19th-c. professions and titles
4	16 (0.10%)	3715/47151/ 28192.62	7	1/10/2.54	68.75/18.75/6.25/ 6.25/6.25	footnotes	definition
5	52 (0.19%)	64/38221/ 6403.10	2	1/30/7.67	67.31/19.23/11.54/ 1.92/0	glossary at the end of book	definition; POS given if unclear
6	51 (0.25%)	1073/49547/ 20903.90	13	1/36/4.71	54.90/21.57/7.84/ 15.69/5.88	item blue within text; definition in margin	item (as lemma) incorporated in a sentence; illustrations (1)
7	59 (0.36%)	56/48192/ 16798.67	14	1/55/6.24	62.71/18.64/13.56/ 1.69/0	item blue within text; definition in margin	item (as lemma) incorporated in a s-ce; illustrations (5)
8	32 (0.29%)	150/15123/ 3034.53	0	1/116/12.53	59.38/18.75/9.38/ 6.25/9.38	glossary at the end of book	definition; POS given if unclear
9	59 (0.35%)	2116/46712/ 18848.30	0	1/9/2.82	38/18.64/1.69/ 13.56/5.08	item blue within text; definition in margin	item (as lemma) incorporated in a s-ce
10	20 (0.32%)	293/9083/ 2356.05	0	1/17/5.55	55/30/15/0/0	glossary at the end of book	definition; POS given if unclear

[1] explanations of vocabulary items by means of a definition, translation, illustration or a combination thereof

[7] https://pypi.org/project/Janome/

[2] rankings of the glossed words are based on the corresponding language's frequency list as found in https://github.com/hermitdave/FrequencyWords
[3] the number of times a glossed item appears in the entire text. Counts found to be zero (based on processing errors) are disregarded.
[4] the percentage of nouns, verbs, adjectives, expressions (items consisting of multiple words) and cultural items (e.g. 'Prince Charming', 'river Thames')
[5] in the text's language

Table 2. Information about the glosses of the selected abridged texts

4.2 Automatically Generated Glossaries

Please look at Table 3 for a full analysis of the glossaries generated by ChatGPT.
One may note that the model tends to generate a round number of glosses and to make them overly detailed (including parts of speech, definitions, examples of use, etc.), especially when it comes to Japanese text. The only instance of translations being offered instead of definitions is in the zero-shot scenario for Bulgarian, the lowest-resource language involved.

language/ scenario	total words in text	glosses	highest / lowest/ av. frequency rank	# words outside of frequency list[1]	min/ max/ av. density	% n./ v./ adj. / expressions/ cultural items	description of what a glossary entry includes
EN/ zero-shot	1971	20 (1.01%)	917/33124/ 10122.78	2	1/3/1.47	14/3/3/0/0	the item (lemma); POS; definition; example s-ce
EN/ one-shot	1971	24 (1.22%)	917/41209/ 12098.18	2	1/6/2	18/2/4/0/1	the item (lemma); POS; definition
BG/ zero-shot	1784	50 (2.80%)	2189/48774/ 19467.42	6	1/7/2.40	28/16/3/3/0	the item (lemma); English translation
BG/ one-shot	1784	20 (1.12%)	1232/30644/ 14317.33	3	1/3/1.82	14/5/0/1/0	the item; definition; example s-ce
JP/ zero-shot	3733	16 (0.43%)	283/9584/ 5327.29	9	1/4/1.70	10/2/3/0/0	item (kanji[2]); item (furigana[3]); POS; definition; definition in JP; example s-ce; example s-ce's translation in EN
JP/ one-shot	3733	20 (0.54%)	7265/16509/ 9362.17	14	1/3/1.23	10/3/7/0/0	item (kanji); item (furigana); POS; definition; usage; example s-ce; example s-ce's translation in EN

[1] the Japanese frequency list contains 35k words, whilst the others contain 50k
[2] Japanese logographic characters
[3] Japanese syllabic characters

Table 3. Information about the glosses generated by ChatGPT in English, Bulgarian and Japanese in the zero-shot and one-shot scenarios.

5. Discussion and Final Artefact

The average density of the one-shot ChatGPT output for the English-language text is higher and thus closer to the professional baseline (where 2.15 is the smallest encountered value) and demonstrates higher variance between its minimal and maximal values (5 vs 2). This observation, in addition to the fact that there is a cultural item included and no example sentences (an element present in the zero-shot output but absent from the investigated professional texts) leads to the selection of the one-shot output as the better option. The same scenario is selected for the Bulgarian text due to the fact that there is a significantly smaller number of glossa involved (20 as opposed to 50) and that the average frequency ranking of the glosses is higher. Finally, it is once again the one-shot output that is opted for in relation to Japanese text. The reasons are that there are no translations of the definitions into Japanese and that the frequency ranks of the items are significantly higher. To sum up, the one-shot scenario is concluded to provide better output, independently of the language involved.
As a final experiment, ChatGPT is provided with its initial one-shot English-language output and then instructed to refine it based on observations and values discovered in relation to the investigated professional glosses (the prompt's exact phrasing is based on a short process of trial and error):

Please make the following changes to the glossary:

1. Remove the POS. Please keep them only in the case of items for which there is associated polysemy (e.g. 'empty' can be an adjective or a verb)
2. Add 1-2 expressions (i.e. items that are longer than a single word) to the glossary if you find suitable ones in the text. Exchanging a word from the glossary with an expression or combining words into an expression is also possible.
3. Provide an image to illustrate a suitable item from the glossary.

Please do not change anything else.

Fig. 1 shows the ChatGPT's final output. An expression ('burning with curiosity') is added, as demanded, and a relevant and high-quality image is generated. Parts of speech are indeed removed, unfortunately even for the item 'bank', despite the model being specifically instructed to retain them in cases of polysemy (a phenomenon of which 'bank' is a perfect textbook example).

bank: the ground along the edge of a river or lake
peep: to look quickly or secretly at something
remarkable: something unusual or surprising
waistcoat: a type of sleeveless jacket worn over a shirt, often part of a suit
burning with curiosity: having a very strong desire to know or learn something
hedge: a line of bushes or small trees that form a boundary or barrier
rabbit-hole: an entrance or tunnel leading underground (also used metaphorically for a strange or complicated situation)
cupboard: a piece of furniture with doors where you can store things, like plates or food
marmalade: a type of sweet spread made from citrus fruit, like oranges
latitude: imaginary lines on a map that run east to west, showing how far north or south a place is
longitude: imaginary lines on a map that run from the North to the South Pole, showing how far east or west a place is
Antipathies: a mistaken word used by Alice; it refers to the people living on the opposite side of the Earth (correct word: antipodes)
curtsy: a polite bow made by bending the knees, usually by girls or women
saucer: a small, shallow dish that holds a cup
bat: a small flying animal that comes out at night
whiskers: the long hairs that grow near the mouth of some animals, like cats or rabbits
hall: a large room or passage in a building
lamp: a light, usually electric, that gives light to a room
golden: made of or looking like gold
telescope: a device used to see faraway objects, like stars
poison: a substance that can make you very sick or kill you if you eat or drink it
slippery: smooth and difficult to hold or walk on
currants: small dried fruits used in baking
respectable: something or someone considered to be proper or good

A rabbit-hole

Figure 1. The final English-language glossary as generated by ChatGPT based on the first chapter of *Alice's Adventures in Wonderland* (produces in one-shot scenario plus an additional prompt addressing the output's characteristics)

6. Conclusion and Future Directions

In the current study, a selection of professionally adapted texts in different languages was analysed qualitatively (e.g. the nature of glosses) and quantitatively (e.g. the frequencies of the items involved). The GPT-4o model was instructed to generate glossaries based on a learner-friendly text as present in three distinct languages (English, Bulgarian and Japanese) in both a zero-shot and one-shot scenario. With the professional glosses regarded as gold standard, the one-shot output was concluded to be of better quality. No major limitations or specificities were noted in relation to the different languages involved. As a final experiment, the English-language glossary provided by ChatGPT in the one-shot scenario was further improved through prompt engineering, demonstrating serious potential of the model for glossary generation based on chain-of-thought prompting. Minor shortcomings include the model's inability to define items that may be associated with polysemy.

The investigated practice of glossary generation does not relieve teaching professionals from selecting suitable learning materials for their students. It would be especially efficient if combined with automatic

textual adaptation or automatic estimation of textual levels; which we are planning to address in further research.

7. Limitations

A larger textual corpus would make the analysis of professional glosses more globally applicable. In the established one-shot scenario, an extract in English was utilised for all languages for which the model was expected to generate output. Whilst this approach is robust and leads to satisfying results, experiments with additional texts in different languages may be included and the outcomes compared. Also, as LLMs develop fast, additional and more recent models should be added to the discussion. Finally, the current study focuses on printed learning materials such as traditional graded readers and does not investigate the preferred and most efficient digital formats that may also be relevant to many readers' practice.

REFERENCES

[1] AbuSeileek, A. F. (2011). Hypermedia annotation presentation: The effect of location and type on the EFL learners' achievement in reading comprehension and vocabulary acquisition. Computers & Education, 57(1), 1281-1291.
[2] Atkins, B. T. S. (1985). Monolingual and bilingual learners' dictionaries: A comparison. In R. Ilson (Ed.), Dictionaries, Lexicography and Language Learning (pp. 15-24). Oxford: Pergamon Press and the British Council.
[3] Bergenholtz, H., & Gouws, R. (2012). What is lexicography? Lexikos, 22, 10.5788/22-1-996.
[4] Dikinson, P. (2017). Effects of extensive reading on EFL learner reading attitudes. In The 21st Conference of the Pan Pacific Association of Applied Linguistics (pp. 28-35). Tamkang University, Taiwan.
[5] Godwin-Jones, R. (2018). Contextualized Vocabulary Learning. Language Learning & Technology, 22(3), 1–19. https://doi.org/10125/44651
[6] Hill, D. R. (2013). Graded readers. ELT Journal, 67(1), 85-125. https://doi.org/10.1093/elt/ccs067
[7] Lee, H., & Lee, J. H. (2015). The effects of electronic glossing types on foreign language vocabulary learning: Different types of format and glossary information. Asia-Pacific Education Researcher, 24(4), 591-601. https://doi.org/10.1007/s40299-014-0204-3
[8] Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. TESOL Quarterly, 37(4), 645-670.
[9] Nation, P. & Wang, K. (1999). Graded Readers and Vocabulary. Reading in a Foreign Language, 12(2), 355-379.
[10] Nation, P. (2013). Learning vocabulary in another language. Cambridge University Press.
[11] Nation, P., & Waring, R. (2020). Teaching Extensive Reading in Another Language. Routledge.
[12] Nation, P. (2007). The four strands. Innovation in Language Learning and Teaching, 1(1), 1-12.
[13] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., … & Zoph, B. (2024). *GPT-4 Technical Report.* arXiv. https://arxiv.org/abs/2303.08774
[14] Piotrowsky, T. (1989). Monolingual and bilingual dictionaries: Fundamental differences. In M. L. Tickoo (Ed.), Learners' Dictionaries: State of the Art. SEAMEO RELC, Singapore.
[15] Restrepo Ramos, F. D. (2015). Incidental Vocabulary Learning in Second Language Acquisition: A Literature Review. PROFILE Issues in Teachers' Professional Development, 17(1), 157-166. http://dx.doi.org/10.15446/profile.v17n1.43957
[16] Takase, A. (2003). The effects of extensive reading on the motivation of Japanese high school students. Unpublished doctoral dissertation, Temple University, Japan.
[17] Teng, F. (2020). Retention of new words learned incidentally from reading: Word exposure frequency, L1 marginal glosses, and their combination. Language Teaching Research, 24(6), 785-812. https://doi.org/10.1177/1362168819829026
[18] Thongsan, N. C. (2023). Vocabulary uptake and retention from reading a graded reader. LEARN Journal: Language Education and Acquisition Research Network, 16(2), 154-167.
[19] Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? Reading in a Foreign Language, 15, 130-163.
[20] Wan-a-rom, U. (2008). Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language*, 20(1), 43-69.