Improving the lexicographic accessibility of WN through LLMs

Ágoston Tóth
Department of English Linguistics
University of Debrecen
toth.agoston@arts.unideb.hu

Esra Abdelzaher
Department of English Linguistics
University of Debrecen
esra.abdelzaher@gmail.com

Abstract

This paper reports the results of an ongoing research on the usability of neural language models to improve WordNet (WN) data for pedagogical lexicographic use. We test the efficacy of BERT-based methods for the selection of example sentences from SemCor and the addition of guidewords to WN senses. We probed our method in a series of time-measured classroom experiments that used WN data only and WN data after adding example sentences and guidewords. We compare two methods of the automatic selection of "good" examples for lexicographic use and discuss the value of BERT probability scores to the selection of useful guidewords. The gap between the pedagogical values of the SemCor extracted sentences and the handpicked examples in WN was reflected in the longer time students spent on the decoding tasks after adding examples and guidewords. However, the decoding performance, especially in the synonym selection task, significantly improved. We argue that the use of Large Language Models can help in improving the accessibility of WN information for educational purposes.

1 Introduction

Gouws (2018) proposed that the accessibility, clarity and retrieval of lexicographic information are the three determining factors that predict the success or failure of a dictionary. The outer access to the information is facilitated in WordNet (WN; Fellbaum, 1998) through a search engine that locates lemmas (co-)listed in synonym sets, similar to any online dictionary. However, the inner access to the information at the microstructure level (e.g. senses, glosses, examples) is practically obstructed – for nonprofessional users – in various ways, including the sophisticated hierarchical representation of sense relations, variety of hyperlinks, information overload, limited and sometimes lack of example sentences and absence of guidewords.

This study probes the usability of Large Language Models (LLMs) in facilitating the use of English WN 3.1 senses for pedagogical lexicography. We mainly targeted two lexicographic features that appear to be insufficient in or missing from the WN design (Miller, 1995; Fellbaum, 1998): example sentences and guidewords. We aim at answering the following questions:

1. Can BERT (Devlin et al., 2019) help in selecting useful examples for practical lexicographic use?
2. How typical or good are the sense-tagged sentences in Semantic Concordance (SemCor; Miller et al., 1993) for pedagogical lexicographic use?
3. How far can BERT's most probable words for a target word in a sense-tagged sentence function as a guideword for this sense?
4. Would BERT-based modifications significantly influence the decoding performance or the consultation time shown by English as a Second Language (ESL) learners consulting WN-modified entries?

The rest of this paper will discuss the importance of guidewords and example sentences in pedagogical lexicography (Section 2), overview the usability of LLMs in lexicographic practice in Section 3, describe the methods of utilizing LLMs in example and guideword selection and the collection of data in Section 4, display and discuss the most important findings in Section 5 and draw the conclusion in Section 6.

2 Access and microstructure of WN: missing features

Despite the uniqueness of the access and microstructure features of WN as a lexicographic resource, the database underrepresents two significant features to a lexicographic entry, namely guidewords and example sentences. Whereas the latter is limitedly represented for several senses, the former is totally missing from the database by design.

The importance of examples in a dictionary entry has been stressed since Dr. Johnson's plan for a dictionary (Atkins and Rundell, 2008). Lexicographic examples should tell the user about the standard and idiosyncratic behavior of a word (Kilgarriff, 2005, 2013) and they are particularly important to the explanation of abstract words and disambiguation of related word senses and near-synonyms (Abdelzaher, 2024; Fillmore and Atkins, 1992). Therefore, several proposals have been made to the selection of the best examples for lexicographic resources, e.g. using handcrafted examples by expert lexicographers, corpus-based citations without alternation (Ruppenhofer et al., 2016), automatic algorithmic selection from corpus sentences (Kilgarriff et al., 2008).

Guidewords, signposts or shortcuts are additional words or phrases preceding each sense to help dictionary users locate the required information faster and easily. Guidewords can be hypernyms, hyponyms or even brief glosses that improve the accessibility of information for users (Heuberger, 2016). They are currently present in leading pedagogical dictionaries such as the Oxford Advanced Learner's Dictionary. The importance of guidewords increases in longer dictionary entries as they shorten the consultation time (Abdelzaer, 2022; Lew and Pajkowvska, 2007; Ptasznik and Lew, 2014), increase the accuracy of sense selection (Dziemianko, 2016), their form also affects the accuracy of encoding performance (Dziemianko, 2017).

3 The usability of LLMs in lexicography

Lexicographers have been arguing for and against the usability of generative and non-generative LLMs, especially recent GPT models, in performing traditional lexicographic tasks which require the processing of large corpora given the large corpora involved in the training of such models. This may facilitate the lexicographic tasks such as updating lists of headwords, writing definitions and selecting examples, but at the same time imposes risks of reproducing linguistic bias and circulating hallucinations (McKean and Fitzgerald, 2024). Prominent lexicographers such as de Schryver participated in a Youtube-registered talk about the possibility of replacing lexicographers with ChatGPT. He explained how this AI-based model can translate strings of words, create dictionary entries for existing words, propose humorous fake entries for words and produce XML-formatted entries (De Schryver and Joffe, 2023). Similarly, Phoodai and Rikk (2023) attempted to compare lexicographic information in ChatGPT-generated entries to lexicographic information in OALD to show the effectiveness of AI models to date.

In contrast, Jakubíček and Rundell (2023) effectively responded to de Schryver's and Joffe's (2023) with a detailed evaluation of 99 entries that were generated using ChatGPT. They highlighted the lexicographic limitations of using such methods despite their outperformance of existing NLP technologies. First, the word sense induction task is limited by (a) generation of false polysemy of the same sense, (b) missing senses despite their frequency based on corpus analysis and (c) suggesting senses which are not evident to the authors without providing citation. Furthermore, they detected some syntactic errors in formulating the definitions and illustrated the lack of diversity in the example sentences suggested by the system which limited their pedagogical value. Moreover, they judged the examples as saliently formulaic and unnatural. However, they praised the system's ability to assign labels to marked uses, such as the "archaic" ones and acknowledged the definitions of general and technical words. Nichols (2023) referred to the model's improper handling of synonymy, frequent generation of syntactically erroneous responses and considerable change in the responses to the same question.

Therefore, the more human-supervised and corpus-centered approach of using the output of non-generative LLMs could be more helpful in lexicography. Tóth and Abdelzaher (2023), for instance, explored the combined use of dimensionality reduction algorithms and the output of neural word embeddings (BERT representations) in finding clusters of word senses. The results showed the usability of visualized clusters in detecting semantic and syntactic

patterns of word uses, although the suggested clusters did not correspond to the sense categories in the studied dictionary.

In the present study, we argue for the use of non-generative LLMs in pedagogical lexicography as they let us use authentic corpus data rather than generated output, and they are directly trained to carry out token unmasking, which we will make use of (see Section 4). These characteristics make non-generative LLMs highly relevant for our purposes despite the fact that they tend to feature less parameters than modern generative models.

4 Methodology

We conducted three experiments focusing on sense and synonym selection tasks. The lexicographic entries in each experiment contained the same words (e.g. *appear*, *tell, development*) but included different lexicographic information cited from WN and Semantic Concordance (SemCor 3.0). We accessed SemCor 3.0 through Sketch Engine. The sense-annotated corpus was first announced by Miller et al. (1993) and it has been continuously updated based on the updated senses in WN. The version we use has been automatically mapped to the WN 3.0 senses by Rada Mihalcea (https://www.sketchengine.eu/semcor-annotated-corpus).

4.1 BERT-based selection of examples and guidewords

We used TPEX scoring (Tóth, forthcoming) to characterize SemCor sentences that contained the selected headwords. TPEX relies on large pretrained neural language models that natively support the word unmasking function to list the most probable tokens and the probability of their appearance in the masked position, the position which is originally occupied by the headwords in our experiments. Tóth tested 4 models (BERT, RoBERTa, ALBERT and BigBird) and discussed the use of two TPEX variants, TPEX-*abs* (which returns the probability with which BERT predicts the headword to appear in the masked position, disregarding other candidates and their probabilities) and TPEX-*rel* (the probability of the appearance of the headword in the masked position divided by the probability of the most probable token predicted for that position). TPEX returns values in the [0,1] interval.

In the present paper, we use TPEX-*abs* to characterize SemCor sentences. We also collect probability scores for other tokens predicted to hide behind the masks, and use these lists to select guidewords (see section 5.1 below). In every case, we use BERT (bert-large-uncased) from the Happy Transformer library available at happytransformer.com to carry out the unmasking procedure.

BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) is directly trained on the task of revealing masked tokens in context, the same task that we use it for. Therefore, we do not rely on transfer learning and other techniques that equip GPT and other generative LLMs with AI (or AI-like) capabilities; instead, we employ a machine-learning system to perform a task it is originally trained on.

Since the networks that we use are known to create contextualized word embeddings, lexical ambiguity is not an issue with TPEX scoring; the tested sentences do not have to be disambiguated or annotated in any way, and we are not restricted to use SemCor sentences in future applications.

A limitation of our research is that we make predictions on BERT *tokens.* The token dictionary is restricted to about 30000 token types in BERT, and we compute the probability of appearance in the masked position for single tokens. It is possible to change the token dictionary and include the headwords that we need to cover, but this process is rather resource-hungry as it requires the training of a large neural network (albeit a much smaller one than those driving generative LLMs). It may not be an option in some lexicographical projects, which restricts them to tokens available in the token vocabulary. Whether this issue has a technical solution (perhaps a fine-tuning procedure with a modified vocabulary) or needs further fundamental research is an open question.

4.2 Designing WN-based lexicographic entries

We represented WN data in the conventional lexicographic entries ESL learners are familiar with. In the first experiment, we kept only the synset and the gloss, in the second experiment we added an example sentence selected using GDEX scoring (Kilgarriff et al., 2008) using the default GDEX configuration in Sketch Engine (https://www.sketchengine.eu), and in the – LLM-assisted – third experiment, we added a guideword and an example using the TPEX scoring explained in Subsection 4.1. We replaced the target word with a pseudo word to avoid the influence of previous

exposure; replacing the target word with a coined or obsolete word to test the influence of a lexicographic variable on the decoding performance of learners is a common practice in lexicography (Chan, 2014; Dziemianko, 2016). We used words from the *Compendium of Lost Words*. Appendix 1 shows a sample of the modified lexicographic entries in the third experiment, which embeds the most modified entries.

4.3 User-based testing

We designed two decoding tasks to test the learners' ability to understand the meaning of the target sense from WN's original and modified data. The first task asked the students to read a lexicographic entry and respond to a grid-form question in which all the senses of the target word are present and four test sentences are provided. Participants are required to match each sentence with its correct sense in a one-to-one correspondence task. The task was scored according to binary values (0 for incorrect answers; 1 for correct answers) regardless of the similarities between the correct sense and the chosen sense by the participants.

In the second task, participants were asked to choose the word from six options that could replace the target word in each of the four sentences where they identified the correct sense. The options included synonyms, hyponyms and hypernyms of the target word in a sense other than the one instantiated in the target sentence We also included distractors in the options (i.e., words that are not semantically similar to the target word but they fit within the context of the test sentence). The test sentences have been cited from the WN database. The task has been graded according to the same grading method used in the sense selection task. Samples of task 1 and task 2 are present in Appendix 2.

The test has been conducted using Psytoolkit (Stoet, 2017) which allows automatic measuring of the time spent on each task, supports various question types (e.g. multiple choice, short and long text responses, voice recording) and allows the insertion of video, audio or graphic files in a question. The participants in the experiments were ESL learners in the 3rd and 4th year of English-major programs at a European higher educational institute. More information about the proficiency levels, frequency of using dictionaries and familiarity with the WN database are available in Appendix 3.

5 Results and discussion

5.1 Selection of examples and guidewords

There were salient differences between the scores of GDEX and the TPEX scores which is reflected in the anticorrelation (Pearson-r = -0.0225), but the differences were not statistically significant (P = 0.529284). On several occasions the highest TPEX scores corresponded to 0 GDEX scores and vice versa. Therefore, our suggested approach of multiplying the TPEX score by the GDEX score led to the discard of the examples which are judged as totally not good or atypical, and kept only the examples which are to some extent good and typical according to both algorithms. TPEX selected the following example as the most typical use of *tell*: *and grandma is n't strong enough to take on something like that, and to tell you the truth neither am I* (TPEX score = 0.9). On the contrary, the GDEX score assigned to the same sentence was 0, which led to its exclusion from the experiments. Similarly, GDEX scores were the highest for the sentence *He felt tired and full and calm* (Target sense = *full*_4, GDEX = 0.9), but – according to TPEX – the probability of the appearance of *full* in this context was 0. The sentence was accordingly excluded from the test. It was evident that GDEX scores reflected the overall readability of the sentence, but they were not sensitive to the typical or canonical occurrences of the headwords.

TPEX scores, in contrast, are primarily assigned according to the probability of the occurrence of the target word in the given sentence without considering other pedagogical factors such as the length of the sentence, the presence of pronouns or advanced (e.g. CEFR C1 and C2-level) vocabulary. Although the highest TPEX scores would recommend the most canonical uses of a target word from a corpus, they would not reflect other pedagogically relevant aspects (i.e. the features observed in the GDEX algorithm). It was not accordingly predictable which entries would be more valuable for the learners when they perform the tasks. The scores of the selected examples in the third experiment ranged from 0.8 for *tell* senses 1 and 2 to 0.1 for *development* according to our new, composite score (i.e., GDEX*TPEX).

The list of most probable words suggested by BERT included examples of multiple sense

relations present in the WN database with all their pedagogical values and challenges. To elaborate, *be* was recorded as a hypernym for *appear* in WN and was also frequently suggested by BERT as the most probable word when *appear* was masked. Including *be* as a guideword may not be of any pedagogical value for the learners especially in our experiments (which already disguise the target word). The list of the most probable words included direct and indirect hypernyms, synonyms and near-synonyms, hyponyms and distributionally similar words. Whereas hypernyms and synonyms were usable as guidewords in several cases, the rest of the words were not suitable for the representation of WN senses. BERT probability scores do not seem to mirror the fine-granularity of WN senses even if sense-tagged sentences are processed. *Growth* and *improvement*, for example, appeared as the most probable words for sentences representing different senses of the target word *development.* While *improvement* is the direct hypernym of the first WN sense of development, *growth* is a synonym of the third sense. They cannot be used interchangeably as guidewords in a WN-based entry disregarding their respective senses. However, it is noteworthy that the senses of *development* were highly overlapping in SemCor, too, which led the annotators to assign two senses the same sentence more than once. In the third experiment, a guideword was successfully added to 54% of the senses collectively in all entries. Whereas the entry of *tell* had the highest number of guidewords (for 5 senses out of 7), the entry of *sound* had the least number of guidewords (for 2 senses out of 8) due to the high overlap of the most probable words for almost all senses.

5.2 Differences in the decoding performance and consultation time

Examining the differences in the time and performance among the three groups showed significant variations. First, the consultation time varied significantly between the three groups. Even though participants in the first experiment consulted the shortest entries which included only the synset and the gloss, they spent longer time on the sense selection task than participants in the second group. The consultation time decreased by 35 seconds on average per task for the second group if compared to the first group and increased by 20 seconds for the third group if compared to the first group. Time differences were statistically significant for the three groups (F = 3.51, P = 0.021). The Post Hoc Tukey test showed the significant differences were between the first and third groups (Q = 3.51, P = 0.036) and between the second and third groups (Q = 3.39, P = 0.045). The length of the entry (i.e., the total number of words in the entry) was negatively correlated with the time of sense selection for the three groups but the anti-correlation was statistically significant for the third group only (r = -0.1914, P = 0.00291).

Second, participants in the first group showed the poorest performance in synonym and sense selection tasks whereas participants in the third experiments showed the best performance but spent the longest time on the consultation process. There was a significant correlation between the time spent on the task and the accuracy of sense selection in the second (r = 0.2103, P = 0.001047) and third (r = 0.452, P = 0.00341) experiments. The differences in the sense selection task among the three groups were significant according to one-way ANOVA test (F = 6.812, P = 0.0011). The Post Hoc Tukey test showed the significant differences were between the first and third groups (Q = 5.19, P = 0.0007). The same applies to the accuracy of synonym selection task (F = 7.8055, P = 0.00454). The differences were significant between the first and third groups. Figure 1 shows the overall accuracy of sense and synonym selection among the three groups.

Third, the performance of the participants in the synonym selection task was better than their responses to the sense selection task in the three experiments. However, the difference was statistically significant in the third experiment only, according to ANOVA test (F = 7.12, P = 0.001). There was also a significant anti-correlation between the time spent on the sense and synonym selection tasks in the third experiment (r = -0.362, P = 0.0217).

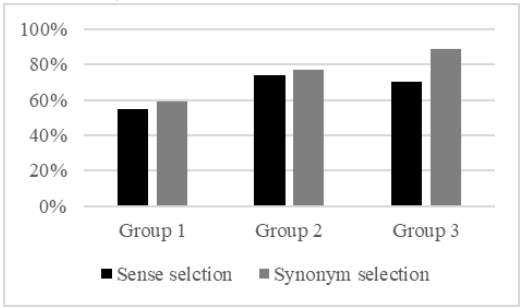


Figure 1: Accuracy of sense and synonym selection tasks

5.3 The influence of the new lexicographic features

It is evident that consulting the entries that included only the synset and the gloss was not effective in helping ESL learners decode the meaning of the target words in either of the tasks, despite spending the shortest time on the task. The addition of a single example sentence for each sense in the entry improved the performance of the participants in the two tasks and surprisingly shortened the consultation time. The examples added to the entries in the second experiment were generally short (average of 6 words) and had relatively low GDEX scores (0.450). They did increase the length of the entry but they did not prolong the consultation period.

The considerable improvement in the decoding performance was noticeable in the third experiment but also there was a considerable delay in the time of the responses to the two tasks. The example sentences added based on the TPEX*GDEX scores for SemCor citations were longer (average of 12 words) than the examples recommended by the GDEX scores for the sentences provided in the WN database in experiment 2. The differences in the length of the example sentences were statistically significant for all test words ($F = 23.278$, $P = 0.00001$). We argue that the presence of the guidewords had a positive effect on shortening the consultation time for long word entries as the shortest consultation time in the third group was associated with the words that had the most number of guidewords and vice versa. Participants spent an average of 4 minutes on the two tasks for the word *tell* (which was the shortest consultation time) and spent 7 minutes on the same tasks when they consulted the entries of *sound* and *development*.

6 Conclusion

This study explored the efficiency of LLMs in improving WN information for pedagogical lexicography. The selection of examples from the WN database or the SemCor corpus has been challenging for the limited number of examples and frequent use of incomplete sentences in the former and the run-on sentences, advanced words and overlapping senses in the latter. It should be noted that WN examples had not been added to the database for lexicographic or teaching purposes. They were, however, added to help in disambiguating one sense from another (Baker and Fellbaum, 2009). Therefore, in many cases they are phrases showing strong associations between the target word sense and another word. They can be beneficial for teaching collocations, but they are not as useful when it comes to explaining word senses for ESL learners (especially if the target word is replaced with a pseudo word). This imposes a challenge on the comprehensive use of WN's sense-tagged sentences in pedagogical lexicography. A combination of two example selection methods (i.e. GDEX and TPEX) appears to be an effective solution for choosing examples of a reasonable length, with accessible words and high probability of the occurrence of the target word in the specified sense. That is to say, the answer to the first question in the present study is yes, the proposed BERT-based method of the selection of examples is helpful in finding typical examples of word use.

Furthermore, enhancing TPEX scores with GDEX score facilitates the selection of more pedagogically valuable examples, which addresses the second question of the study. The SemCor corpus contains many examples that are lexicographically valuable according to the two scoring methods but, unfortunately, around 50% of the SemCor citations for the words tested in our experiments were assigned a score of 0 according to either or both of the scoring methods. Given the challenge of creating a sense-tagged corpus similar to SemCor, future research may consider (a) simplifying the TPEX-selected sentences through shortening their length, (b) replacing C2 words with B1–B2 synonyms or near-synonyms, or both (a) and (b). As annotations are not necessary for TPEX or GDEX scoring, any source of text can be used. Selecting or upscoring examples that only use lemmas that are listed in a controlled (defining) vocabulary, such as the Oxford 3000 list, may also be an option; several learner's dictionaries have already introduced the process of writing the *definitions* based on controlled defining vocabularies, too.

The results of the third experiment indicated the importance of the guidewords to improving the decoding performance of ESL learners, but the challenge of finding appropriate guidewords for the fine-grained senses in WN was also salient. Despite the richness of the sense relations in the database, none of these relations could be systematically used to find a guideword. Synonymy, which is the only relation that is consistently present across all POS

is not available for all word senses, i.e. for single-member synsets. For instance, three of the eight senses of *appear* do not have any synonyms in WN. Moreover, sometimes WN records synonyms that are infrequently used in this sense and would, accordingly, perplex users for either the synonym's unfamiliarity or familiarity in another sense (e.g. *euphony* as a synonym of *music*). The same applies to the use of WN's hypernyms as guidewords. In many cases, the hypernym is too general to provide helpful information to learners (e.g. *process* as a hypernym of *development*). Moreover, the hypernym-hyponym relation is not applicable to the adjective net. In this regard, BERT's most probable words could partially address this challenge by suggesting high-frequency and most probable replacements. However, this does not solve the problem of overgeneralizing a sense by suggesting its direct or indirect hypernym (e.g. *be* for *appear* in several sentences) or further specifying it by suggesting a hyponym (e.g. *ring* or *jingle* for *sound*) or troponym. This shows the complexity of the issue, and we argue that human lexicographical expertise is still key to the success of the guideword-selection process.

Finally, the remarkable advancement in the participants' responses after consulting WN entries in the third experiment (with example sentences and guidewords) shows how the database can be successfully integrated in pedagogical lexicography. WN has already been included in the new types of dictionaries such as aggregators (e.g. *The fine dictionary*) and portals (e.g. *Onelook*) probably due to the accessibility of its structure which is less complicated if compared to other resources such as FrameNet.

Acknowledgments

This publication was supported by the Institute of English and American Studies at the University of Debrecen.

References

Esra M. Abdelzaher. 2022. A classroom-based study on the effectiveness of lexicographic resources. *Lexicography: AsiaLex*, 9(2):139–174.

Esra M. Abdelzaher. 2024. *Approaches of cognitive linguistics and ontologies in lexicographic sense delineation*. Ph.D. thesis, Debrecen University.

Sue Atkins and Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford University Press, Oxford, UK.

Collin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*. Association for Computational Linguistics, Suntec, Singapore, pages 125–129.

Alice Y. W. Chan. 2014. How can ESL students make the best use of learners' dictionaries? *English Today*, 30(3).

Gilles-Maurice de Schryver and David Joffe. 2023. The end of lexicography, welcome to the machine: on how ChatGPT can already take over all of the dictionary maker's tasks. In *The 20th CODH Seminar, Center for Open Data in the Humanities*, Research Organization of Information and Systems, National Institute of Informatics, Tokyo, Japan.

Jacob Devlin, Ming Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, volume 1*.

Anna Dziemianko. 2016. An insight into the visual presentation of signposts in English learners' dictionaries online. *International Journal of Lexicography,* 39(*4*):490–524.

Anna Dziemianko. 2017. Dictionary entries and bathtubs: Does it make sense? *International Journal of Lexicography*, 30(3):263–284.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA, MIT Press.

Charles Fillmore and Sue Atkins. 1992. Towards a frame-based organization of the lexicon: the semantics of RISK and its neighbors. In A. Lehrer and E. Kittay, editors, *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pages 75–102.

Rufus H. Gouws. 2018. Dictionaries and access. In Pedro A. Fuertes-Olivera, editor, *The Routledge Handbook of Lexicography*, pages 43–58.

Reinhard Heuberger. 2016. 'Learners' Dictionaries: History and Development; Current Issues'. In Philip Durkin, editor, *The Oxford Handbook of Lexicography*.

Miloš Jakubíček and Michael Rundell. 2023. The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček and S. Krek, editors, *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex*

2023 conference. Brno, 27–29 June 2023. Lexical Computing CZ s.r.o., Brno, pages 518–533.

Adam Kilgarriff. 2005. Putting the Corpus Into the Dictionary. *Proceedings of the Second MEANING Workshop*, Trento, Italy, 3-4 February 2005.

Adam Kilgarriff. 2013. Using Corpora as Data Sources for Dictionaries. In Howard Jackson, editor, *The Bloomsbury Companion to Lexicography*. Bloomsbury, London, pages 77–96.

Adam Kilgarriff, Katy Mcadam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In Elisenda Bernal and Janet DeCesaris, editors, *Proceedings of the 13th EURALEX International Congress*. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, Barcelona, Spain, pages 425–432.

Robert Lew and Julita Pajkowska. 2007. The Effect Of Signposts On Access Speed And Lookup Task Success in Long And Short Entries. *Revista Horizontes de Linguistica Aplicada*, 6(2), 235-252.

Erin McKean and Will Fitzgerald. 2024. The ROI of AI in lexicography. *Lexicography: Journal of AsiaLex* 11 (1): 7–27.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11): 39–41.

George A. Miller, Claudia Leacock, Randee Tengi and Ross T. Bunker. 1993. A Semantic Concordance. In *Human Language Technology*: *Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993.*

Wendalyn Nichols. 2023. Invisible lexicographers, AI, and the Future of the Dictionary. Youtube, uploaded by eLex conference, 26 July 2023.

Chayanon Phoodai and Richárd Rikk. 2023. Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček and S. Krek, editors, *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023.* Lexical Computing CZ s.r.o., Brno, pages 345–375.

Bartosz Ptasznik and Robert Lew. 2014. Do menus provide added value to signposts in print monolingual dictionary entries? An application of linear mixed-effects modelling in dictionary user research. *International Journal of Lexicography*, 27(3).

Joseph Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher Johnson and Jan Scheffczyk. 2016. *FrameNet II: Extended theory and practice*. International Computer Science Institute, Berkeley, CA.

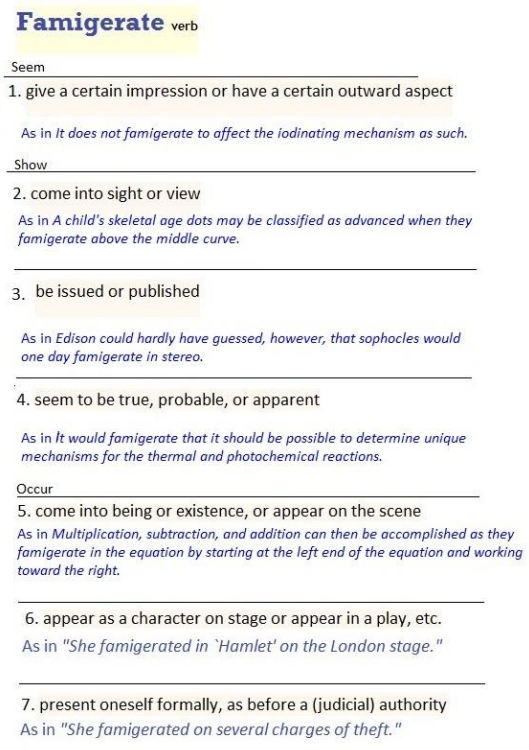
Gijsbert Stoet. 2017. PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, 44(1):24–31.

Ágoston Tóth and Esra Abdelzaher. 2023. Probing visualizations of neural word embeddings for lexicographic use. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček and S. Krek, editors, *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023.* Lexical Computing CZ s.r.o., Brno, pages 545–566.

Ágoston Tóth. Forthcoming. TPEX: Neurális nyelvi modellek alkalmazása példamondatok kiválasztásában ['TPEX: The application of neural language models in selecting example sentences']

A Appendices

Appendix 1. The modified entry for *appear* in experiment 3.



Appendix 2. Samples of sense and synonym selection tasks

Sense selection task for the word *appear*

Choose the meaning of "famigerate" in the following sentences.

Item	1. give a certain impression or have a certain outward aspect	2. come into sight or view	3. be issued or published	4. seem to be true, probable, or apparent	5. come into being or existence, or appear on the scene	6. appear as a character on stage or appear in a play, etc.	7. present oneself formally, as before a (judicial) authority
They famigerate like people who had not eaten or slept for a long time.	○	○	○	○	○	○	○
The new Woody Allen film hasn	○	○	○	○	○	○	○
It famigerates that the weather in California is very bad.	○	○	○	○	○	○	○
A new star famigerated on the horizon.	○	○	○	○	○	○	○

Synonym selection task for the first sense of *appear*

Choose the most suitable word that can replace "famigerate" in the following sentences without making significant changes in the meaning

They famigerate like people who had not eaten or slept for a long time.

- ○ talk
- ○ look
- ○ perform
- ○ speak
- ○ show up
- ○ act

Appendix 3. Description of the participants in the three groups (till the date of submission)

	Ex 1	Ex 2	Ex 3
3rd year students	70%	0	100%
4th year students	30%	65%	0
5th year students	0	35%	0
Proficiency > B2	75%	100%	100%
Daily use of monolingual dictionaries	12%	37%	55%
Daily use of bilingual dictionaries	12%	80%	95%
Familiarity with WN data at the time of the test	0	10%	0

B Supplementary Material

TPEX scores for the sentences cited from SemCor and reported in this study are available through: https://github.com/WNTPEX/TPEX_WN/blob/main/supplementary%20data_TPEX%20scores.xlsx