WordNet under Scrutiny: Dictionary Examples in the Era of Large Language Models

Fatemah Almeman[*△], **Steven Schockaert**[*], **Luis Espinosa-Anke**[*◇]
[*]CardiffNLP, School of Computer Science and Informatics, Cardiff University, UK
[△] College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, KSA
[◇]AMPLYFI, UK
{almemanf, schockaerts1, espinosa-ankel}@cardiff.ac.uk

Abstract

Dictionary definitions play a prominent role in a wide range of NLP tasks, for instance by providing additional context about the meaning of rare and emerging terms. Many dictionaries also provide examples to illustrate the prototypical usage of words, which brings further opportunities for training or enriching NLP models. The intrinsic qualities of dictionaries, and related lexical resources such as glossaries and encyclopedias, are however still not well-understood. While there has been significant work on developing best practices, such guidance has been aimed at traditional usages of dictionaries (e.g. supporting language learners), and it is currently unclear how different quality aspects affect the NLP systems that rely on them. To address this issue, we compare WordNet, the most commonly used lexical resource in NLP, with a variety of dictionaries, as well as with examples that were generated by ChatGPT. Our analysis involves human judgments as well as automatic metrics. We furthermore study the quality of word embeddings derived from dictionary examples, as a proxy for downstream performance. We find that WordNet's examples lead to lower-quality embeddings than those from the Oxford dictionary. Surprisingly, however, the ChatGPT generated examples were found to be most effective overall.

Keywords: lexical resources, dictionary examples, semantics

1. Introduction

Lexical resources are a fundamental repository of knowledge. They include information about words, in the form of definitions, as well as other critical information such as examples of usage, morphology, syntax and etymology. Lexical resources play a key role both in traditional (knowledge-rich) and data-driven NLP (Camacho-Collados et al., 2018). For instance, they have been used for decades for improving word-to-word Machine Translation systems (Masterman, 1957; Sparck Jones, 1986; Artetxe et al., 2017), or in low resource settings, e.g., for improving English-Malayam translations by providing additional vocabularies and inflected verbal forms (S and Bhattacharyya, 2016). Other areas such as word sense disambiguation (WSD) (Kwong, 2001; Fellbaum, 2001) or question answering (Pasca and Harabagiu, 2001) have also benefited. More generally, NLP applications rely on lexical resources for modeling semantics, either directly (Budanitsky and Hirst, 2001; Silber and McCoy, 2002), or via the enrichment and refinement of both knowledge bases (Espinosa-Anke et al., 2016b; Xu et al., 2022) and language models (LMs) (Joshi et al., 2020; Chen et al., 2022).

Despite the importance of lexical resources in NLP, there has only been limited work on evaluating their intrinsic quality, fleshing out their specific features (e.g., type and style of definitions, or readability and informativeness of examples) and studying the extent to which such features dictate the performance of NLP systems. For instance, WordNet (WN) is the de-facto lexical database for English (Miller, 1995), and has been embedded in a myriad of applications (cf. Section 2.1), among others, due to its large provision of `<word,definition,example>` triplets. Here, `example` is a sentence showing the usage of `word` (as defined by `definition`) in context. This is helpful, among others, for definition modeling (Noraset et al., 2017; Gadetsky et al., 2018; Giulianelli et al., 2023), which has in turn been shown to benefit existing lexical semantics systems (Bevilacqua et al., 2020). However, despite WN's popularity, Almeman and Espinosa-Anke (2022) concluded that its examples are often shorter and less informative than those from other lexical resources (e.g., the Oxford Dictionary). This makes it hard to learn a good representation of a word by (only) relying on WN examples. This issue was further investigated by Giulianelli et al. (2023), who used Large Language Models (LLMs) such as FlanT5-XXL (Chung et al., 2022) to generate definitions for words in context, finding this harder for WordNet examples than for examples from the Oxford Dictionary.

In this context, and even in today's LLM era, we find that WN remains widely used. For instance, WN was recently used to create a novel abstract/concrete hypernymy dataset which proved to be challenging even for recent LLMs (especially when it comes to abstract terms) (Liao et al., 2023). WN has also been used with success for curating datasets and, in combination with LLMs, for

LREC-COLING 2024, pages 17683–17695

solving highly specialized tasks such as pun detection (Ermakova et al., 2023). More generally, at the intersection of lexicography and NLP, we can now find studies focused on the capabilities of LLMs when compared to more traditional resources, such as English learners' dictionaries. For instance, Rees and Lew (2023) evaluated AI-generated definitions against those provided by the Macmillan English Dictionary to resolve vocabulary uncertainties within a multiple-choice reading task, which was aimed at testing lexical knowledge. However, there was no significant difference between the performance of students who used MED definitions, had no definitions at all, or were provided with AI-generated definitions. In contrast, Phoodai et al. (2023) found that ChatGPT generally outperforms the Oxford Advanced Learner's Dictionary in providing lexicographical data to English language learners, particularly on microstructural elements.

Therefore, and given the relevance of WN as a go-to resource in NLP, in this paper we extend previous analyses of this resource by conducting a comprehensive evaluation of its examples. In addition to comparing WN's examples against those from other lexical resources, we also include examples that were generated by ChatGPT[1] in our analysis. Specifically, we propose an anlysis that involves human assessments of dictionary examples from several standpoints, namely *naturalness*, *informativeness* and the extent to which examples are *self-contained*. We complement these human judgements with a range of automated metrics. Finally, as a proxy for the usefulness of dictionary examples in downstream tasks, we test the quality of the word embeddings that can be obtained from these examples, using an off-the-shelf model for learning representations of words in context (Liu et al., 2021). Our findings can be summarised as follows:

- WN examples are less informative than those from the Oxford dictionary and those generated by ChatGPT. However, they are generally easy to to understand and are judged to be more natural by human annotators.
- The examples generated by ChatGPT are judged to be considerably more informative than those from WN and the Oxford Dictionary by human annotators. Moreover, word embeddings that are learned from ChatGPT's examples perform considerably better in word similarity benchmarks.
- WN examples tend to be generally (too) short, often include words that are highly ambiguous, and tend to lack fluency.

[1] https://chat.openai.com

The rest of the paper is structured as follows. First, we provide the necessary background for the two main concerns of this paper, namely WN and the principles behind *good dictionary examples* (a.k.a. GDEX criteria). Then, we report on the results of the human evaluation of WN's examples against GDEX criteria. We contrast our findings to an NLP experiment, specifically the evaluation of word similarity, where we use contextualised embeddings of words-in-context to represent words. Specifically, by obtaining these embeddings from dictionary examples, we can use this task to estimate the informativeness of the examples. We then complement our questionnaire and extrinsic evaluation with an intrinsic evaluation of WN examples through automatic metrics (GDEX and readability-based). This allow us to perform a larger-scale comparison with other lexical resources. Finally, we sum up our conclusions and outline directions for future work.

2. Background

In this section we first introduce WordNet (Miller, 1995), as the lexical resource that we focus on in this paper. Second, we recall a set of criteria known as *Good Dictionary Examples* (GDEX) (Kilgarriff et al., 2008), since our evaluation builds on them.

2.1. WordNet

WN is an electronic lexical dictionary for English that organizes words in groups of synonyms called "synsets" (Miller, 1995). Each synset is described by its definition, lemmas (i.e. the set of words or phrases that make up the synset), examples of usage (for some but not all synsets), and its relation to other synsets. Some of the relations that are covered include hypernymy (is-a), meronymy (is-part) and troponymy (manner-of). WN is mainly used in lexicographic and language learning settings (Morato et al., 2004), but it has also proven to be a high-quality knowledge resource for NLP systems. For instance, previous works have shown that base NLP systems can be improved by injecting knowledge from WN in some way, with applications ranging from information retrieval and extraction (Moldovan and Mihalcea, 2000; Banerjee and Pedersen, 2002) to improving word embeddings (Faruqui et al., 2014; Espinosa-Anke et al., 2016a; Mrkšić et al., 2017; Vulić and Mrkšić, 2018), or "simply" serving as the sense inventory for WSD methods of various nature (Agirre and Edmonds, 2007; Zhang et al., 2022; Pu et al., 2023).

2.2. GDEX

The "Good Dictionary EXamples" tool was first implemented as a system that added around 8,000

new example sentences to the Macmillan English Dictionary by automatically finding good examples in corpora using a set of rules of thumb (Kilgarriff et al., 2008; Bejoint, 2014). GDEX criteria are used in different works for extracting examples or concordances from a corpus, and have been studied and discussed, often, more in terms of their usefulness for lexicographers rather than their benefits for NLP. In a nutshell, a *good* dictionary example must be:

- **Typical**, i.e., showing the, as Kilgarriff et al. (2008) put it, "frequent and well-dispersed patterns of usage" of the target word.
- **Informative**, so that it helps with understanding the definition of the word.
- **Intelligible** to the reader by avoiding difficult lexis and structures which cannot be understood without access to a wider context (a.k.a. *readability*) .

Beyond the above core GDEX criteria, Kosem et al. developed further desiderata, including that a dictionary example is:

- **Natural**: the example should appear like a sentence one would expect to see in usual language use.
- **Authentic**: because they are examples from actual corpora.
- **Self-contained**: the content of the example is understandable without requiring additional context.

3. Human Evaluation

Our core motivations are to assess to what extent WN (1) adheres to GDEX criteria; (2) how it compares with another well known resource, namely the Oxford Dictionary (Oxford Dictionary, 1989); and (3) how well it compares with examples automatically generated using LLMs. In order to answer these three questions, two annotators with extensive expertise in computational lexicography and machine learning, good command of English (although not native speakers), and with annotation experience, were provided with a questionnaire specifically designed for evaluating the quality of dictionary examples. In Section 3.1 we first explain the data that was used for constructing this questionnaire. Subsequently, we describe the design of the questionnaire itself in Section 3.2 and we discuss the results in Section 3.3.

3.1. Questionnaire Data

The definitions and examples that were used in the questionnaire were sourced from `3D-Ex` (Almeman et al., 2023), a unified resource containing several dictionaries mapped against common `<word,definition>` and `<word,definition,example>` triplets. The examples in the questionnaire were derived from all `<word,definition>` instances that were identified through an exact match in both WN and a subset of the Oxford Dictionary. We will refer to the latter dictionary as CHA, which is common practice in the literature, as per the first work that introduced this dictionary into NLP applications (Chang and Chen, 2019).

Generating Examples with ChatGPT In addition to data from WN and CHA, we expanded the questionnaire with examples that were generated using GPT-3 (`text-davinci-003`) and ChatGPT (`gpt-3.5-turbo`). We used two different prompts to obtain the examples: the **simple** format (*"Write a sentence showing the word {word}, defined by {definition} in context"*); and the **GDEX** format, which extended the simple prompt with instructions about how the example should be written following GDEX criteria.

Questionnaire Statistics The questionnaire focused on 87 words. Each of these words has a corresponding definition and usage example in both WN and CHA. For each word, we also included four generated examples (two prompts for each of GPT-3 and ChatGPT), resulting in a total of $87 \times 6 = 522$ examples. Each example was evaluated based on three criteria (see Section 3.2) by two annotators, meaning that this annotation exercise resulted in 3,132 unique annotations. Table 1 illustrates the kind of examples that the participants were asked to annotate, in this case for the word "cage". Note that the sources were hidden from the participants.

3.2. Questions Design

The questions presented in the questionnaire were aimed at capturing participants' perception of GDEX criteria in a granular way. To this end, the questionnaire was split into two sections: **definitions evaluation** and **examples evaluation**. While our main focus was on the examples evaluation, we included an evaluation of the definitions as well, to assess to what extent the definitions alone were successful in clarifying the meaning of the considered terms. The primary objective of this exercise is to determine whether (and how much) examples can help readers understand the meaning of difficult or unfamiliar terms that have unclear or difficult definitions. Specifically, for *definitions evaluation*, participants were presented with a word and definition, and were asked to assign one of the following labels:

Example	Source
The animal was caged.	WordNet
In future should I leave the house I will cage the dogs no matter who else is in the house with them.	CHA
The zookeeper had to cage the wild animals to ensure the safety of the visitors.	ChatGPT-simple
The zookeeper had to cage the wild animal to ensure the safety of visitors.	ChatGPT-GDEX
The zookeeper caged the lion to keep it from escaping.	GPT-3-simple
The animal was confined in a small cage, unable to escape its captivity.	GPT-3-GDEX

Table 1: Examples for the term "cage", defined as "confine in a cage".

> **Unclear**: `upon reviewing the provided definition, the meaning of the term remains unclear or difficult to comprehend`
>
> **Borderline**: `the definition gave me some insight into the term's meaning, but it is still unclear`
>
> **Clear**: `the definition clearly and fully explains the meaning of the term`

The *examples evaluation* section of the questionnaire aimed at evaluating the WN and CHA examples, as well as those generated by GPT-3 and ChatGPT. In this case, annotators were asked to rank each example according to the following criteria:

- **Self containment:** `Was the dictionary example fully understandable to you without the need for wider context or to consult external sources? (1-3 where 1:No, 2: Partially, and 3: Yes)`
- **Informativeness:** `Regardless of your prior knowledge of the term, how well did the example clarify or elaborate on its meaning? (1-5; 1 being the lowest, 5 the highest)`
- **Naturalness:** `Naturalness: How well does this example reflect the style and wording you'd expect to find in everyday language use? (1-5; 1 being the lowest and 5 the highest)`

3.3. Results and Analysis

Annotator Agreement First, Table 2 reports the agreement between the two participants, in terms of Fleiss' kappa, as well as the Pearson (PCC) and Spearman (SCC) correlation coefficients. Overall, we can conclude that there was a fair level of agreement between the two annotators. When interpreting relatively low Fleiss' kappa scores, the fine-grained nature of the annotation scales needs to be taken into account. As the Pearson and Spearman correlation scores show, the annotators largely agreed on the overall trends. In particular, the main conclusions from our analysis below remain valid whether we look at the annotations from either annotator alone, or whether we aggregate their scores.

Informativeness for Challenging Definitions We first delve into which resource (or pseudo-resource) provided the highest number of informative examples. For this analysis, we specifically focus on words whose definition received ratings of being *unclear* or *borderline* by at least one of the annotators. This was the case for 45 of the definitions. We focus specifically on these 45 words, as the main purpose of dictionary examples is to help clarify potentially incomplete definitions. Upon examining the examples associated with these words, as shown in Table 3, it can be seen that `ChatGPT-GDEX` yielded the highest number of informative examples, rated 4 or 5 by one annotator or more. In contrast, WN exhibited the lowest count of informative examples. In terms of uninformative examples (those with a score of 1 or 2 from at least one annotator), WN presented the highest number, while `GPT-simple` displayed the lowest. Despite this, the differences are relatively small, especially in the number of informative examples. However, we do observe a trend that will become evident throughout the rest of this paper, which is the struggle of WN in providing informative examples for challenging definitions, as opposed to `ChatGPT` and `GPT-3`. This is further supported by the examples presented in Table 1.

GDEX Criteria Let us now take a closer look at the assessment of the different GDEX criteria for the 6 considered resources. Figure 1 shows the response means and standard errors, which we can interpret as follows. WN examples appear highly natural but they are somewhat lacking in informativeness, which would suggest they may be easy to understand but not very useful examples. In comparison, the CHA examples are more informative but less natural, as well as being slightly less self-contained. Interestingly, the GPT-3 and ChatGPT generated examples were rated as being considerably more informative, while at the same time also being more natural than those from CHA, albeit less natural than those from WN. Moreover, when comparing `ChatGPT-simple` with `ChatGPT-GDEX` and

Metric	Kappa	PCC	SCC
Definitions Clarity	0.32	0.42	0.39
Examples Naturalness	0.20	0.62	0.62
Examples Informativeness	0.23	0.59	0.50
Examples Self-containment	0.24	0.37	0.36

Table 2: Annotator agreement results

Source	Informative	Uninformative
WN	30	22
CHA	31	14
ChatGPT-simple	34	2
ChatGPT-GDEX	**36**	2
GPT-3-simple	33	**1**
GPT-3-GDEX	33	4

Table 3: Table 3: Number of informative and uninformative examples for unclear definitions in each source (**bold**: best, underlined: worst).[2]

comparing `GPT-3-simple` with `GPT-3-GDEX`, we cannot see any benefits from the GDEX based prompting strategy. This confirms that, while the GDEX prompt leads to longer and more complex examples, in practice, they prove to be just as effective as the zero-shot approach without explicit instructions.

4. Similarity Experiment

This experiment aims to validate the human judgments collected through the questionnaire by employing WN, CHA, and ChatGPT examples to compute similarities between word pairs. Specifically, for this experiment, we use the examples to generate word embeddings, using MirrorWiC (Liu et al., 2021), a state-of-the-art model for learning high-quality representations of words or phrases in context. The idea behind this experiment is that informative examples should lead to higher-quality embeddings. To evaluate the quality of the word embeddings, we rely on a number of standard word similarity benchmarks, namely SimLex-999 (Hill et al., 2015), SimVerb-3500 (Gerz et al., 2016), Stanford's Contextual Word Similarities (SCWS) (Huang et al., 2012), and MEN Test Collection (Bruni et al., 2014). We first extracted the common words between WN and CHA along with their examples, and for each word we generated 5 different examples from ChatGPT using the two different prompts (Section 3.1). Then for each similarity dataset we retrieved the word pairs that can be found in the common words set. For each pair, we computed the cosine similarity between the MirrorWiC embeddings of their associated examples. If a word has multiple examples in WN or CHA, we select the one that leads to the highest similarity score. By comparing the similarity scores with the gold scores provided by the similarity datasets, we found that ChatGPT examples have the best encoding for all datasets while WN-derived embeddings seems less suitable for the task. Table 5 shows the Pearson's Correlation Coefficient (PCC) and the Spearman's Correlation Coefficient (SCC) between the gold similarity scores and the cosine similarity between examples' encodings.

In addition to the word similarity results, we also list a few illustrative examples (Table 4) where, for different word pairs, we show the dictionary examples pair with the highest cosine similarity for each resource. The disparity in the quality of the resources (and GPT generations) becomes apparent. For instance, for *easy* and *tough*, we find that the most similar WN examples are less informative, and most critically, the antonymic relationship between both words is not actually reflected by the given sentence pair. The CHA and GPT generations do not suffer from this issue. A similar situation happens with *dull* and *funny*, where the antonymic relationship is not captured by the WN example pair, and instead we find that both examples elicit health-related senses. CHA, in this case, also falls short (*dull edge* and *funny stomach*), but both of the GPT generated pairs are expressing the sense related to entertainment. Finally, for *rock* and *jazz*, the WN examples pair again shows a conflating of meanings (music, but also exaggerated talk in *"don't give me any of that jazz"*), with CHA and GPT_s both providing accurate music-themed senses. Interestingly, however, GPT_g provides an example pair where a visual arts sense of *jazz* (*"decorated with a jazz theme"*) was found to be most similar to the music sense of *rock*.

5. Automatic Evaluation

In order to complement the insights derived from the questionnaire analysis, we are also interested in comparing the examples from different sources in terms of automatic metrics. We focus on evaluating their readability (i.e., how easy it would be to understand them), as well as a number of other measurable criteria which are sometimes considered as important for good dictionary examples Kilgarriff et al. (2008). For this analysis we look at two different settings. First, we consider all dictionaries with examples included in `3D-ex`. In this case, we randomly sample 1,000 examples from each resource. Note that these examples are necessarily for different words, since the overlap between some of the sources is small. Therefore, in our second

[2]When an example is labeled as informative by one annotator and uninformative by the other, it appears in both counts.

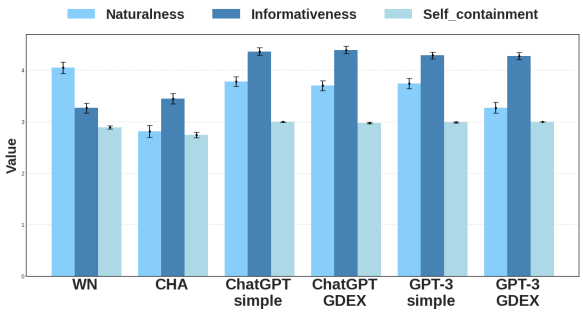

Figure 1: Questionnaire results per source

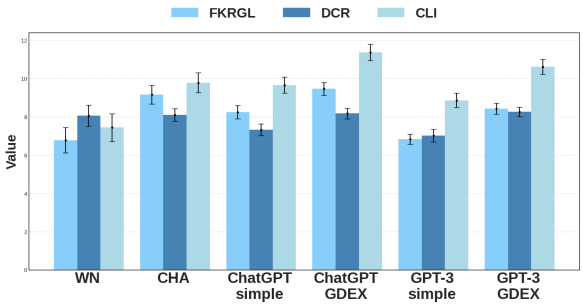

Figure 2: Readability results per source

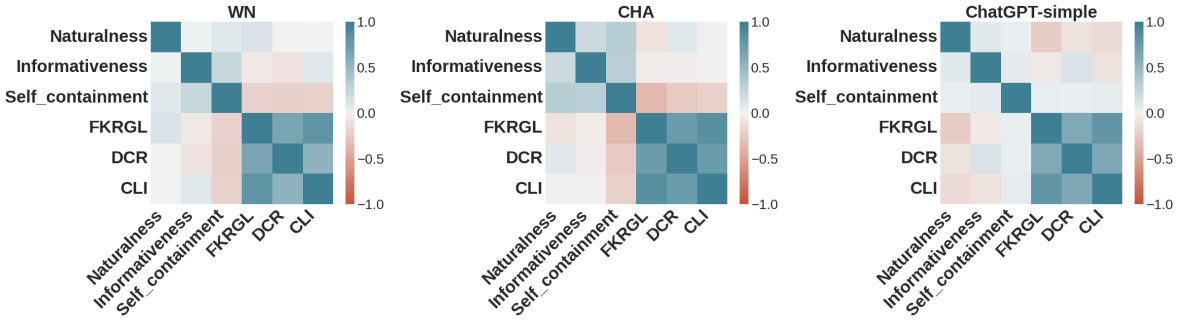

Figure 3: Pearson correlation between questionnaire criteria and readability metrics in WN, CHA, and ChatGPT-simple.

setting, we focus on the words in the intersection of WN and CHA. For this analysis, we will be able to directly compare the examples from these two dictionaries with each other, as well as with the examples generated from GPT-3 and ChatGPT. We first recall the nature and main features of all the resources considered:

- **Wikipedia**: a collaborative online encyclopedia generated by a community of online contributors (Yano and Kang, 2016).
- **Wiktionary**: A web-based dictionary provides different information about terms, including definitions, examples, pronunciation, and more (Bajčetić and Declerck, 2022).
- **Urban** (Urban Dictionary): a crowd-sourced resource focusing on terms that are usually not included in conventional dictionaries (Wilson et al., 2020).
- **CODWOE**: The English version of the dataset used for the CODWOE (Comparing Dictionaries and Word embeddings) SemEval 2022 shared task (Mickus et al., 2022).
- **Sci-definition**: a dataset constructed to generate definitions for scientific terms with controllable complexity (August et al., 2022).

5.1. Metrics

We measured the quality of dictionary examples using two groups of metrics. First, we used readability metrics, which determine the U.S. grade level needed to comprehend a sentence: Flesch–Kincaid Reading Grade Level (FKRGL) (Kincaid et al., 1975), which uses word length and sentence length, Dale-Chall Readability (DCR) (Dale and Chall, 1948), which is based on sentence length and the number of 'hard' words, and Coleman-Liau Index (CLI) (Coleman and Liau, 1975), which is calculated using the average number of letters per 100 words and the average sentence length. Second, we looked at proxies for determining adherence to the following **GDEX criteria**:

- **Sentence fluency**: we use the **GPT-2** [3] language model for measuring fluency of dictionary examples based on their log-likelihood score.
- **Sentence length** (**len-pen**): a good example should be between 10 and 25 words long based on Kilgarriff et al. (2008). Accordingly, the length penalty was calculated as 1 minus

[3]We chose GPT-2 instead of GPT-3 as it provides sufficient performance while being more user-friendly and cost-effective.

W1	W2	Encoded Example 1	Encoded Example 2	Dict.
easy	tough	An easy victim.	A tough character.	WN
		With so many people to choose from, booking several dates in a short amount of time is easy.	Getting published and earning a bit of critical acclaim to spur on further creative efforts is tough for those starting out.	CHA
		She has a natural talent for learning languages, so picking up Spanish was easy for her.	He had a tough time adjusting to his new school.	ChatGPT$_s$
		The hike up the mountain was challenging, but the descent was easy and enjoyable.	The hiking trail was tough, with steep inclines and rocky terrain.	ChatGPT$_g$
dull	funny	Dull pain.	Told the doctor about the funny sensations in her chest.	WN
		Most cooks use the point because the edge is dull.	Suddenly my stomach felt funny.	CHA
		His sense of humor was quite dull, and his jokes rarely elicited laughter.	My friend has a funny way of telling stories; he always adds humorous details.	ChatGPT$_s$
		The lecture was so dull that I struggled to stay awake.	The comedian's jokes were so funny that the entire audience couldn't stop laughing.	ChatGPT$_g$
rock	jazz	That mountain is solid rock.	Don't give me any of that jazz.	WN
		The movie is a disappointment and could have been a lot better if only he had gone out on a few more limbs than just the inclusion of a few rock tunes.	They're playing a kind of light jazz, something lively to listen to without having to know the words.	CHA
		My favorite genre of music is classic rock.	I love listening to jazz music on a lazy Sunday afternoon.	ChatGPT$_s$
		The concert was held in an open-air amphitheater, and the crowd swayed and danced to the rhythm of the rock music.	The interior of the restaurant was decorated with a jazz theme.	ChatGPT$_g$

Table 4: Examples from the word similarity experiment, showing the pair of examples with the maximum cosine similarity between their MirroWiC embeddings for several resources: WordNet (WN), CHA (CHA), ChatGPT-simple with a simple prompt (ChatGPT$_s$) and ChatGPT-GDEX prompted with instructions on writing a good dictionary example following GDEX (ChatGPT$_g$).

	SimLex		SimVerb		SCWS		Men	
	PCC	SCC	PCC	SCC	PCC	SCC	PCC	SCC
WordNet	0.18	0.16	0.21	0.21	0.59	0.54	0.51	0.52
CHA	0.25	0.25	0.28	0.26	0.62	0.58	0.60	0.60
ChatGPT-simple	0.44	**0.43**	0.37	0.36	**0.68**	**0.66**	**0.71**	**0.72**
ChatGPT-GDEX	**0.46**	**0.43**	**0.42**	**0.40**	**0.68**	**0.66**	0.69	0.70

Table 5: Correlation between the gold similarity scores and the cosine similarity between examples' encodings, in terms of Pearson (PCC) and Spearman (SCC) Correlation Coefficient (**bold**: best, underlined: worst).

the reciprocal of the absolute difference from the desired range.

- **Word frequency** (**freq-pen**): a sentence was penalized for each non-frequent word, defined as a word which is not among the top 20,000 most common words on the English language, as derived from the Google Web Trillion Word Corpus (Brants and Franz, 2006). This penalty score is derived by computing the ratio of non-frequent words to the total number of words in the sentence.
- **Anaphoric references** (**ana-pen**): this

penalty score was calculated by dividing the number of pronouns in the dictionary example by its total number of words.

- **Ambiguity**: we penalized the presence of ambiguous words in a sentence by summing up the number of senses for each word (using WordNet senses) and then dividing this by the total number of words in the sentence.

- **The main clause** (**m-clause**): we penalized examples where the target word does not appear in the main clause. Specifically, examples where the target word is in the main clause are scored 1, with other examples being scored 0. To identify the main clause, we used a transition-based dependency parser[4].

5.2. Assessing 3D-EX Sources

Table 6 shows the automatic evaluation results, where we report the average for each metric. Our analysis reveals that WN examples exhibit the lowest sentence fluency, and they also tend to have a higher penalty for using ambiguous or multi-sense words. In contrast, WN does well at ensuring that the target word is included in the main clause of the sentence, and it provides easy to read examples as shown by its scores in the readability metrics. In addition, its penalties for non-frequent words and anaphoric references are low compared to the other resources.

Conversely, Sci-definition examples show the highest log-likelihood scores, suggesting they are more coherent and fluent. Nevertheless, Sci-definition examples demonstrate higher grade levels in all readability metrics, implying a greater level of complexity, which is unsurprising given that they were sourced from scientific journal abstracts. Moreover, Wikipedia has the lowest penalty for sentence length, anaphoric references, and ambiguity. This suggests that Wikipedia's examples are closer to the ideal length, use fewer pronouns for clearer communication, and have fewer words with multiple meanings for easier understanding. However, Wikipedia, along with Urban, received a higher penalty for non-frequent words, while CHA has the lowest penalty for this metric, indicating a preference for commonly used language in its examples. Urban demonstrates high readability, as indicated by its lower grade levels in FKGRL and CLI[5], but has higher penalties for the use of non-frequent words, which is unsurprising given the very nature of Urban Dictionary.

5.3. WN vs CHA vs GPT

In this section, we look at the data used for the dictionary example evaluation via the GDEX-motivated questionnaire (Section 3). Recall that, in addition to including actual lexicographic resources (WN and CHA), we also include two instances of GPT as described in Section 3.1. In terms of results, the most immediate conclusion is that, upon comparing `ChatGPT-simple` with `ChatGPT-GDEX`, and `GPT-3-simple` with `GPT-3-GDEX` in Figure 2, it becomes evident that the GDEX-prompted examples exhibit higher readability grades, indicating that they might be more challenging to read. However, when compared with the questionnaire results, they yield very similar outcomes, especially in terms of informativeness. Figure 3 shows the Pearson correlations between readability metrics and questionnaire criteria for some of the datasets. In both WN and CHA, there is a negative correlation between self-containment and readability, which means that annotators frequently labeled easy to read examples with high self-containment scores. Additionally, A negative correlation between naturalness and readability is observed in ChatGPT-simple and GPT-3-simple. This implies that their easy to read examples tend to also be annotated as natural.

6. Conclusion and Future Work

In this work we have evaluated WN examples in comparison with existing lexicographic resources and similar content automatically generated by GPT. Our findings highlight that although WN is a valuable resource that excels at providing a certain type of dictionary example, it does not seem to be the optimal resource when informative contexts are required. We also found that the gains by using GDEX criteria in a prompt to ChatGPT are negligible, which could point to the fact that ChatGPT *already has a deep understanding of what a good dictionary example should look like*. Finally, in our downstream analysis, using word similarity as a proxy, we found that indeed examples from ChatGPT yielded better embeddings *in all datasets*, leaving little doubt about what to prefer when it comes to using dictionary examples for downstream applications relying on word or phrase representations.

For the future, we would like to extend the questionnaire to other resources and LLMs, and leverage the scores we obtained for training *dictionary scoring* systems, which we believe would be helpful tools both for lexicographers and NLP practitioners. Additionally, in similar spirit to other works that extended GDEX (or, more generally, studied sen-

[4]Implemented with SpaCy: https://spacy.io/.

[5]This counter-intuitive result might be explained by the large proportion of slang and colloquial lingo. Further analysis could shed light into how to measure readability in Urban Dictionary, considering its obvious idiosyncrasies.

Dataset	fluency	len-pen	freq-pen	ana-pen	ambiguity	m-clause ↑	FKRGL	DCR	CLI
WordNet	-6.08	0.64	0.18	0.08	5.91	**0.97**	6.97	**8.48**	7.93
CHA	-5.06	0.14	**0.09**	0.09	5.89	0.83	10.24	8.80	10.57
Wikipedia	-5.14	**0.11**	0.23	**0.02**	**4.19**	0.96	11.11	11.41	10.77
Wiktionary	-4.99	0.42	0.22	0.08	4.89	0.65	11.60	9.23	10.73
Urban	-5.75	0.28	0.23	0.11	5.74	0.79	**4.17**	8.60	**4.52**
CODWOE	-5.08	0.46	0.21	0.09	5.13	0.80	9.26	8.72	8.87
Sci-definition	**-4.53**	0.31	0.19	0.05	4.46	0.82	16.47	12.16	15.93

Table 6: Examples automatic evaluation results (**bold**: best, underlined: worst, in all metrics, a lower value is better, with the exception of "fluency" and "m-clause".)

tences for assisted language learning), we would like to extend this work to other languages. Finally, considering the intended users of the dictionary is important when comparing dictionaries, as this can affect the complexity of the examples.

7. Acknowledgments

This work was supported by EPSRC grant EP/V025961/1.

8. Bibliographical References

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1).

Eneko Agirre and Philip Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

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 Almeman, Hadi Sheikhi, and Luis Espinosa Anke. 2023. 3D-EX: A unified dataset of definitions and dictionary examples.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Lenka Bajčetić and Thierry Declerck. 2022. Using Wiktionary to create specialized lexical resources and datasets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3457–3460, Marseille, France. European Language Resources Association.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. volume 2276, pages 136–145.

Ahmet Basal. 2019. Learning collocations: Effects of online tools on teaching english adjective-noun collocations. *British Journal of Educational Technology*, 50(1):342–356.

Henri Bejoint. 2014. The bloomsbury companion to lexicography edited by howard jackson. *Dictionaries: Journal of the Dictionary Society of North America*, 35:374–381.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1 (2006). *Linguistic Data Consortium, Philadelphia*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49:1–47.

Meru Brunn, Yllias Chali, and Christopher Pinchak. 2001. Text summarization using lexical chains.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Alexander Budanitsky and Graeme Hirst. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources*.

Jose Camacho-Collados, Luis Espinosa Anke, and Mohammad Taher Pilehvar. 2018. The interplay between lexical resources and natural language processing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 17–23, New Orleans, Louisiana. Association for Computational Linguistics.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Qianglong Chen, Feng-Lin Li, Guohai Xu, Ming Yan, Ji Zhang, and Yin Zhang. 2022. Dictbert: Dictionary description knowledge enhanced language model pre-training via contrastive learning. In *International Joint Conference on Artificial Intelligence*.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Claudio Delli Bovi, Luca Telesca, and Roberto Navigli. 2015. Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543.

Umut Ufuk Demirhan. 2016. A frequency dictionary of turkish: Core vocabulary for learners.

Jörg Didakowski, Lothar Lemnitzer, and Alexander Geyken. 2012. Automatic example sentence extraction for a contemporary german dictionary. In *Proceedings EURALEX*, pages 343–349.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023. Overview of joker–clef-2023 track on automatic wordplay analysis. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 397–415. Springer.

Luis Espinosa-Anke, Jose Camacho-Collados, Sara Rodríguez Fernández, Horacio Saggion, and Leo Wanner. 2016a. Extending wordnet with fine-grained collocational information via supervised distributional learning. In *Proceedings of COLING 2016: Technical Papers. The 26th International Conference on Computational Linguistics; 2016 Dec. 11-16; Osaka (Japan).[place unknown]: COLING; 2016. p. 900-10.* COLING.

Luis Espinosa-Anke, Horacio Saggion, Francesco Ronzano, and Roberto Navigli. 2016b. Extasem! extending, taxonomizing and semantifying domain terminologies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Manaal Faruqui, Jesse Dodge, Sujay Jauhar, Chris Dyer, Eduard Hovy, and Noah Smith. 2014. Retrofitting word vectors to semantic lexicons.

Christiane Fellbaum. 2001. Manual and automatic semantic annotation with wordnet. In *Proceedings of the NAACL WordNet and Other Lexical Resources workshop*, page 3 – 10, Pittsburgh.

Christiane Fellbaum. 2013. Wordnet. In Carol Chapelle, editor, *The encyclopedia of applied linguistics*, pages 6739–6746. Blackwell Publishing Ltd.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Ana Frankenberg-Garcia, Robert Lew, Jonathan C Roberts, Geraint Paul Rees, and Nirwan Sharma. 2019. Developing a writing assistant to help eap writers with collocations in real time. *ReCALL*, 31(1):23–39.

Ana Frankenberg-Garcia, Geraint Paul Rees, and Robert Lew. 2020. Slipping Through the Cracks in e-Lexicography. *International Journal of Lexicography*, 34(2):206–234.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge.

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. Universitat Pompeu Fabra Barcelona.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of the 11th EURALEX International Congress*, pages 105–115, Lorient, France.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Iztok Kosem, Milos Husák, and Diana McCarthy. 2011. Gdex for slovene. *Proceedings of eLex*, pages 151–159.

Iztok Kosem, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit, and Carole Tiberius. 2019. Identification and automatic extraction of good dictionary examples: the case (s) of gdex. *International Journal of Lexicography*, 32(2):119–137.

Oi Yee Kwong. 2001. Word sense disambiguation with an integrated lexical resource. In *Proceedings of the NAACL WordNet and Other Lexical Resources workshop*, page 11 – 16, Pittsburgh.

Lothar Lemnitzer, Christian Pölitz, Jörg Didakowski, and Alexander Geyken. 2015. Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard german. In *Proceedings of the eLex 2015 conference*, pages 11–13.

Jiayi Liao, Xu Chen, and Lun Du. 2023. Concept understanding in large language models: An empirical study.

Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021. Mirrorwic: On eliciting word-in-context representations from pretrained language models. In *Proceedings of the 25rd Conference on Computational Natural Language Learning (CoNLL)*.

Nikola Ljubešić and Mario Peronja. 2015. Predicting corpus example quality via supervised machine learning. In *Proc. Electronic Lexicography in the 21st Century Conference (eLex)*, pages 477–485.

Margaret Masterman. 1957. The thesaurus in syntax and semantics. *Mech. Transl. Comput. Linguistics*, 4:35–43.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

D.I. Moldovan and Rada Mihalcea. 2000. Using wordnet and lexical operators to improve internet searches. *Internet Computing, IEEE*, 4:34 – 43.

Jorge Morato, Miguel Marzal, Juan Llorens, and Jos Moreiro. 2004. Wordnet applications. *Proceedings of the 2nd Global Wordnet Conference*, 2004.

Nikola Mrkšić, Ivan Vulić, Diarmuid O Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the association for Computational Linguistics*, 5:309–324.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

M. Pasca and S. Harabagiu. 2001. The informative role of wordnet in open-domain question answering. *Workshop on WordNet and Other Lexical Resources at NAACL*.

M Phoodai, Richárd Rikk, M Medveď, M Měchura, I Kosem, J Kallas, C Tiberius, and M Jakubíček. 2023. Exploring the capabilities of chatgpt for lexicographical purposes: A comparison with oxford advanced learner's dictionary within the microstructural framework. In *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*, pages 335–365.

Ildikó Pilán, Elena Volodina, and Lars Borin. 2016. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues*, 57(3):67–91.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013. Automatic selection of suitable sentences for language learning exercises. In *20 Years of EUROCALL: Learning from the Past, Looking to the Future: 2013 EUROCALL Conference Proceedings*, pages 218–225. Research-publishing. net Dublin.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pages 174–184.

Xiao Pu, Lin Yuan, Jiaxu Leng, Tao Wu, and Xinbo Gao. 2023. Lexical knowledge enhanced text matching via distilled word sense disambiguation. *Knowledge-Based Systems*, page 110282.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Geraint Paul Rees and Robert Lew. 2023. The effectiveness of openai gpt-generated definitions versus definitions from an english learners' dictionary in a lexically orientated reading task. *International Journal of Lexicography*, page ecad030.

Sreelekha S and Pushpak Bhattacharyya. 2016. Lexical resources to enrich english malayalam machine translation.

H. Gregory Silber and Kathleen F. McCoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4):487–496.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Edgar A Smith and J Peter Kincaid. 1970. Derivation and validation of the automated readability index for use with technical materials. *Human factors*, 12(5):457–564.

Karen Sparck Jones. 1986. *Synonymy and semantic classification*. Edinburgh University Press.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. pages 1134–1145.

Noah Webster. 1900. *Webster's unabridged dictionary of the English language*. Kikwansha.

Steven R. Wilson, Walid Magdy, Barbara McGillivray, Venkata Rama Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang nlp applications. In *International Conference on Language Resources and Evaluation*.

Hongyuan Xu, Yunong Chen, Zichen Liu, Yanlong Wen, and Xiaojie Yuan. 2022. Taxoprompt: A prompt-based generation method with taxonomic context for self-supervised taxonomy expansion. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4432–4438. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Tae Yano and Moonyoung Kang. 2016. Taking advantage of wikipedia in natural language processing.

Guobiao Zhang, Wenpeng Lu, Xueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. 2022. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4061–4070.

Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model.

Ruimin Zhu, Thanapon Noraset, Alisa Liu, Wenxin Jiang, and Doug Downey. 2019. Multi-sense definition modeling using word sense decompositions.

9. Language Resource References

Miller, George. 1995. *WordNet: A Lexical Database for English*. [link].

Oxford Dictionary, English. 1989. *Oxford english dictionary*.