Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework

Chayanon Phoodai[1], **Richárd Rikk**

[1] European Master in Lexicography (EMLex),
Károli Gáspár University of the Reformed Church in Hungary,
H-1088 Budapest, Reviczky str. 4.
E-mail: chayanon507@gmail.com, rikk.richard@gmail.com

Abstract

Artificial Intelligence (AI) has seen success in many areas of science in the past few years. From computer science to linguistics, deep neural networks have the ability to perform better than the previous state-of- the art solutions. Indeed, generative text-based models like ChatGPT are able to imitate human writing, however its capabilities in lexicography have not been studied thoroughly. This paper compares the lexicographical data provided by ChatGPT and the Oxford Advanced Learner's Dictionary in the scope of microstructure. Two main datasets are created for manual analysis and similarity score tests. The aim is to demonstrate the effectiveness of ChatGPT in providing lexicographical data to English language learners as compared to the Oxford Advanced Learner's Dictionary.

We accomplish this by comparing the provided data related to lexicographical items, using Wiegand's item classes to identify the co-occurring items within the microstructure of both platforms. The framework of item classes provides us with a list of lexicographical items that serve as our criteria. We then examine each lexical entry individually to determine whether each lexicographical item is present in both tools. The results are presented in a comparative table as percentages. Also, using Bilingual Evaluation Understudy (BLEU) and Recall Oriented Understudy for Gisting Evaluation (ROUGE) methods we calculate the similarity between the lexicographical data provided by ChatGPT and the Oxford Advanced Learner's Dictionary. Since ChatGPT has been trained on human data, we investigate how similar its generated answers are to the ground truth.

This study provides valuable insights into the potential of AI-generated dictionary content and its applicability in pedagogical lexicography. Additionally, it highlights the challenges and limitations that need to be addressed in order to inform the development of AI models for lexicography.

Keywords: Artificial Intelligence; Generative Models; ChatGPT; E-lexicography; Microstructure; Oxford Advanced Learner's Dictionary

1. Introduction

Artificial Intelligence (AI) plays a significant role in natural language processing (NLP). Large language models (LLMs) Bahdanau et al. (2014) can provide better solutions than the previous state-of-the-art in areas such as machine translation Brants et al. (2012), code synthesis Poesia et al. (2022), text summarization Pilault et al. (2020), and more Araci (2019); Dathathri et al. (2019); Kant et al. (2018); Yasunaga et al. (2021).

Despite the success of LLMs, their applicability in lexicography remains mostly understudied. In this paper, we evaluate ChatGPT Brown et al. (2020); OpenAI (2023) in a lexicographical context by comparing it to the 10th Edition of Oxford Advanced Learner's Dictionary (OALD) Hornby (2019). We do this by using the Wiegand (1989) item classes and similarity scores.

We use Wiegand (1989)'s item classes to determine how answers of ChatGPT to lexicographical questions and the information provided by OALD align with the structural requirements of a dictionary. Item classes provide a comprehensive method to determine which and how lexicographical items should be presented. This allows us to compare OALD and answers of ChatGPT in an objective manner, and gain useful insights of these tools including what information they do and do not provide. In order to use this method, we compile two main datasets containing information from ChatGPT and OALD regarding the lexicographical items of the most frequently used English words according to the British National Corpus (BNC) from Oxford Text Archive (2007). Iterating over our first dataset, we manually check if the given lexicographical item satisfies the criteria given by the item classes. Our findings then collected into a comparative table. We aim to show how effective ChatGPT is in providing lexicographical data for English language learners compared to a conventionally assembled dictionary.

Also, we calculate similarity scores using Bilingual Evaluation Understudy (BLEU) Papineni et al. (2002) and Recall Oriented Understudy for Gisting Evaluation (ROUGE) Lin (2004) methods which are widely used in NLP for determining the syntactic similarities of texts. The scores are calculated programmatically on one of our datasets. This dataset includes only those lexicographical items that can be compared using BLEU and ROUGE, like item giving pronunciation, spelling, part of speech, definition, and etymology. These items have been chosen for their unambiguity, which not only makes it possible to calculate similarity scores on them, but also allows us to consider the information provided by OALD as ground truth. We apply BLEU and ROUGE method to every lemma for every lexicographical item in the dataset, and visualize our results on separate figures. Last, we calculate the average of BLEU and ROUGE scores by lexicographical items only. Since ChatGPT has been trained on large amounts of human generated data gathered from the internet, we aim to show how much ChatGPT deviates from the ground truth. Large deviations have to be examined further as ChatGPT has some tendency to state incorrect information. In these cases, the expected score should be close to zero. Therefore, manual analysis of BLEU and ROUGE scores can allow us to investigate the reliability of ChatGPT as well.

Using Wiegand's item classes and similarity scores, we provide comparative analyses in a lexicographical context between ChatGPT and OALD. Our research gives insights into the viability of AI-generated dictionary content, and aims to help the adoption of such technologies in language learning and education. Also, it tries to identify some of the limitations and challenges of AI in lexicography to inform the development of models in the field.

In the next sections, we go over our method in detail. First, an overview is provided highlighting all the main parts of our method. Then, the item classes and the comparative table provided by them are discussed in detail. Next, we describe the similarity scores, results of the calculation, and their meaning. After that, we summarize our results from the two different methods and finish with our conclusions.

2. Related work

Previous studies have explored different aspects of monolingual learner's dictionaries (MLDs), such as their interface, software, structure, and user experience. In this section, we review related work that provides valuable insights and guidelines for conducting comparative studies on MLDs.

Herbst (1996) examines the features of four popular English learners' dictionaries: Oxford Advanced Learner's Dictionary (OALD5), Longman Dictionary of Contemporary English (LDOCE3), Collins COBUILD English Dictionary (COBUILD2), and Cambridge International Dictionary of English (CIDE). The study's methodology involves a detailed analysis and comparison of the dictionaries' features, including their target users, corpus basis, definitions, pronunciation, example policies, valency information, collocations and phrases, labelling system, illustrations, access structure etc. The paper employs a qualitative research approach, relying on the author's expert judgement and critical evaluation of the dictionaries' strengths and weaknesses. The study's findings are based on a thorough and systematic comparison of the dictionaries' features. The author provides clear and detailed explanations of the criteria used for evaluation. Overall, the study's methodology is rigorous and comprehensive, and the findings are based on a thorough analysis of the dictionaries' features and feedback from language experts and users. However, the study does not provide statistical analysis or quantitative data, and the evaluation criteria used by the author are subjective to some extent.

Ivančič & Fabijanić (2017) present an approach for analysing the chronological development of the macro- and microstructure of the OALD. Ten editions were investigated to find out the similarities and differences. This study involves methodology of the analytical standpoint of the authors, because it takes us thoroughly through different lexicographical item within the macro- and the microstructure. The findings are shown comparatively between the ten editions in tabular form. The study shows that the both macro- and microstructure have been expanding increasingly over each edition. Variety of new sections in MLDs has been introduced. This is to encourage the EFL learner's language skills. This study is highly relevant to our research as it focuses on the development of OALD specifically and its treatment of lemmas within the dictionary.

While these two studies provide us with comprehensive framework for conducting detailed manual analysis within the microstructure and offer guidelines for comparing different dictionaries, they lack objective criteria as both studies rely solely on the author's opinion. To address this limitation, we propose the use of reliable criteria for analysis, specifically *Wiegand's item classes* described in the methodology section. By adopting these established criteria, we can ensure a more reliable and unbiased approach to our analysis, moving beyond the subjective viewpoint of the authors alone.

3. Methodology

This research paper is a comparative study that aims to show the capabilities of ChatGPT for lexicographical purposes and compare it with the OALD focusing on the microstructural elements. To accomplish this, we provide a detailed explanation of the methods employed for this study in the following section. In addition, this section will provide a comprehensive overview of the entire study process (see Figure 1).

3.1 Corpus and lemma selection

In order to notice the differences of microstructural elements we selected the ten most frequently used words from five different parts of speech (POS) including noun, verb, adjective, adverb, and preposition. According to the frequency counts in Davies & Gardner (2013), our chosen five POS belong to the most commonly used functional word classes in English. We choose lemmas from different POS because the lexicographical items in dictionary entries can vary even within the same category. We selected 50 lemmas from the British National Corpus (BNC) Oxford Text Archive (2007). While various corpora may produce slightly different outcomes, our choice of corpus does not significantly affect our study's purpose of showcasing the likeness of the most frequently utilized English words.

3.2 Wiegand's item classes

According to Wiegand (1989), dictionaries have more than 200 classes of functional text segments that serve as structural indicators within the dictionary microstructure. However, for the purpose of our study, we focused only on the lexicographical items suggested by Wiegand for general and learner's dictionaries. Since the OALD falls into this category and our objective is to assess the capabilities of ChatGPT as a learner's dictionary, we have chosen the suggested item classes and lexicographical items as our criteria for evaluation. Engelberg & Lemnitzer (2009) adapted these lexicographical items and simplified their names. To ensure coherence and ease of understanding, we adapted these names of the item classes for our English language research and assigned them acronyms for easy reference in subsequent sections, namely Lemma Sign (LES), Phonetic-Phonological and Orthographic Information (PPOI), Morphological Information (MOI), Syntactic Information (SYI), Syntactic-Semantic Information (SYSI), Semantic Information (SEMI), Pragmatic Information (PRAI), and Other Items (OTI), that contains lexicographical items which do not belong to the other item classes. We use the lexicographical items belonging to the above mentioned classes to test ChatGPT's competence as a learner's dictionary. ChatGPT is assigned to provide information for each item to complete the task.

3.3 Lexicographical prompts

Prompt engineering is becoming an essential skill for effective communication with large language models (LLMs) like ChatGPT. These prompts serve as instructions that enforce rules, automate processes, and ensure specific qualities and quantities of output. They are also a programming tool that can customize the outputs and interactions with the language model White et al. (2023).

In order to make ChatGPT to complete a task effectively, an appropriate prompt needs to be formulated. Different style of prompt may affect the quality of the outputs. Initially, we begin formulating our own prompt giving ChatGPT to *provide dictionary information* *for a given "lemma"*. With this prompt, the responses generated by ChatGPT can be inconsistent, even with the same prompt used. Each time, ChatGPT produces a different structure. Prompt 1 shows how ChatGPT typically generates the structure. The overall result does not meet our satisfaction due to the lack of data within one dictionary entry. Prompt 1 demonstrates that ChatGPT only provides us data related to five lexicographical items which are the form of the lemma sign, the spelling, part of speech, and polysemy. Only some of the polysemous senses are displayed in Prompt 1, despite the fact that the original response contains approximately 20 senses. Due to the lack of data in an entry we then ask ChatGPT itself for the advice of the possible prompts for our lexicographical tasks. Prompt 2 shows the response of possible prompts for different lexicographical tasks. While the prompts provided by ChatGPT show potential for lexicographical tasks, their effectiveness has not been established in the field of lexicography. As such, we consider them as an initial attempt to assess the response of our lexicographical items within Wiegand's item classes.

Prompt 1: Provide dictionary information for "time"

Sure, here is the dictionary information for the word "time":

noun

1. as a measurement, "time" refers to a duration or a quantity that is measured in units of seconds, minutes, hours, days, weeks, months, years, or other units depending on the context.
2. the duration of a particular event or situation.

verb

1. to measure or record the speed, duration, or rate of.
2. to adjust (a mechanism) so as to synchronize its operation with a timing device or with a sequence of operations.

Prompt 2: Provide ten concise prompts or templates that can make you do lexicographical tasks

Sure, here are ten concise prompts or templates for lexicographical tasks:

1. Define the word [insert word here].
2. Create a list of synonyms for [insert word here].
3. Provide an antonym for [insert word here].
4. Explain the origin and etymology of [insert word here].
5. Write a sentence using [insert word here] in context.
6. Give an example of a word that is commonly confused with [insert word here].
7. What part of speech does [insert word here] belong to?
8. Identify the syllable stress in [insert word here].
9. Provide a definition for [insert phrase or idiom here].
10. List three words that are derived from [insert word root here].

3.4 Datasets

All the information we gather from the BNC, lexicographical items from Wiegand's item classes, responses from ChatGPT and OALD are stored in two primary tables. The first table lists 50 selected lemmas from the most frequently used English words across five different parts of speech from BNC. Additionally, the table presents 41 lexicographical items belonging to eight distinct item classes. Each column of the lexicographical items is marked with *present* or *absent* indicating whether ChatGPT and OALD can provide information related to the corresponding lexicographical item for each lemma. The second table lists 50 selected lemmas as like in the first table and five chosen lexicographical items LES, AUSA, RA, WAA, ABED[1], and ETYA. The table also contains the actual answers in textual form that we gather from both sources to be calculated for their similarities with BLEU and ROUGE.

3.5 Manual Analysis

We use the collected dataset to analyze the results for eight item classes: LES, PPOI, MOI, SYI, SYSI, SEMI, PRAI, and OTI. This evaluation helps us assess the capabilities of ChatGPT. For each item class, we examine whether ChatGPT and OALD are capable of providing the corresponding lexicographical items within the microstructure. Additionally, we analyze how they present the corresponding data, if available. The tables display lexicographical items in each class, lemma count[2], and three different types of symbols: percentages (%), plus signs (+), and minus signs (–). Percentages represent the availability of data provided by both tools for related lexicographical items, while a minus sign indicates unavailability of the data. A plus sign indicates that the related data is available but beyond the scope of our selected 50 lemmas.

3.6 Similarity Scores

In addition we calculate how similar the provided answers from ChatGPT and OALD are by using BLEU Papineni et al. (2002) and ROUGE Lin (2004). It is important to note that these scores do not indicate the quality of the answers, but rather measure the extent to which they align with the human-edited dictionary entries in a learner's dictionary. Both calculation methods are not simple scoring functions, but robust frameworks aimed at evaluating NLP model outputs using given reference texts. Therefore, we only cover parts of these methods that are relevant for our research purposes.

For clarification, let us describe the most important definitions before we go over our calculations. In the field of NLP, an n-gram is a contiguous sequence of $n \in \mathbb{N}$ tokens from a given sample of text. They are instances of a sequence of characters that are grouped together as a useful semantic unit for processing. Depending on the application in which they are used, tokens can be a simple character, few characters, or even words. This paper considers tokens that represent words. When $n = 1$ the n-gram is called a unigram, $n = 2$ a bigram, and $n = 3$ it is a trigram. In our calculations, we use multiple n values to provide a more complete picture.

[1] This includes polysemous senses of the definition within the entries.

[2] This indicates the number of lemmas that can undergo certain lexicographical items, as some items are only applicable to certain parts of speech. The provided percentages also correspond to this.

Two other useful definitions are reference and candidate text. The former can be considered as ground truth and it is usually compiled by humans, while the latter is generated by a NLP model. In our case, reference text is information gathered from OALD, while candidate text refers to answers collected from ChatGPT. Compering reference and candidate texts yields a similarity score $s \in [0, \ldots, 1]$. If $s = 0$ the texts are completely different, while $s = 1$ means they are the same according to the used method. However, it is important to highlight that BLEU and ROUGE only considers the syntactics and not the semantics of a text.

3.6.1 Method BLEU

Originally, BLEU is designed for machine translation tasks. However, it is widely used in other areas such as code comparison Rikk et al. (2022) for program synthesis. This section gives an overview of the method and introduces all key concepts of it.

This method calculates the n-gram overlaps between the reference and candidate texts. Usually, we have multiple of the former as there can be multiple correct translation for a given text. However in our case, the reference text is obtained from OALD, because we are only interested in the similarities between it and ChatGPT.

Now, we go over how BLEU is calculated. Let us define the count function which given a text T and a n-gram g returns the number of times g is in T.

$$count(g, T) = \sum_{\substack{t \in T \\ t = g}} 1 \tag{1}$$

Next, a clipped count $count_c$ function given a list of reference texts $\mathcal{R}$ and candidate text C calculates the maximum number of times a n-gram occurs in any single reference translation. Then clips the total count of each candidate n-grams by its maximum reference count.

$$count_c(g, \mathcal{R}, C) = \min\left(count(g, C), \max_{R \in \mathcal{R}} count(g, R)\right) \tag{2}$$

With Equations (1) and (2), BLEU is calculated as follows. We first compute the n-gram matches sentence by sentence. Next, we add the clipped n-gram counts for all the candidate sentences and divide by the number of candidate n- grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus.

$$p_n = \frac{\sum_{c \in C} \sum_{g \in c} count_c(g, c)}{\sum_{c' \in C'} \sum_{g' \in c'} count(g', c)} \tag{3}$$

Then, we take the geometric mean of the test corpus' modified precision scores and then multiply the result by an exponential brevity penalty factor. If k is the length of the candidate translation and r is the effective reference corpus length, then the brevity penalty BP:

$$BP = \begin{cases} 1 & \text{if, } k > r \\ e^{\frac{1-r}{k}} & \text{if, } k \leq r \end{cases} \tag{4}$$

Last, $BLEU$ function is calculated as

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5}$$

where $w_n \in \mathbb{R}$ is called weight and $\sum_n w_n = 1$. In our calculations, we use a variety of weights to obtain a more robust evaluation. Depending on the n-grams used in Equation (5), it is also referred to as BLEU-n.

3.6.2 Method ROUGE

ROUGE is a set of metrics, rather than just one method. In this section, we cover the main approaches that are used in our tests, starting with ROUGE-N.

Formally, ROUGE-N is an n-gram recall between a candidate summary C and a set of reference summaries $\mathcal{R}$. ROUGE-N is computed as follows:

$$\text{ROUGE-}N = \frac{\sum_{R \in \mathcal{R}} \sum_{g \in R} count_m(g, C)}{\sum_{R' \in \mathcal{R}} \sum_{g' \in R'} count(g', C)} \tag{6}$$

where g is a n-gram, function $count_m$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries, while function $count$ is defined as Equation (1). With ROUGE-N, N represents the n-gram that we are using. For ROUGE-1, we would be measuring the match-rate of unigrams between our model output and reference.

ROUGE-N can calculate three different values. These are recall, precision, and F1 score. Recall counts the number of overlapping n-grams found in both the model output and reference, then divides this number by the total number of n-grams in the reference (Equation (6)). This ensures that our model is capturing all of the information contained in the reference, but this is not so great at ensuring our model is not just pushing out a huge number of words to game the recall score. To avoid this, we use the precision metric, which is calculated just as the recall except, we divide by the model n-gram count and not with the reference n-gram count. Last, the F1 score is calculated as

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

ROUGE-L measures the longest common subsequence (LCS) between our model output and reference. We can apply our recall, precision, and F1 calculations just like before, but this time we replace $count_m$ with the LCS count.

4. Manual Analysis

This section presents the findings of our manual analysis, which is organized according to the item classes proposed by Wiegand, each containing relevant lexicographical items.

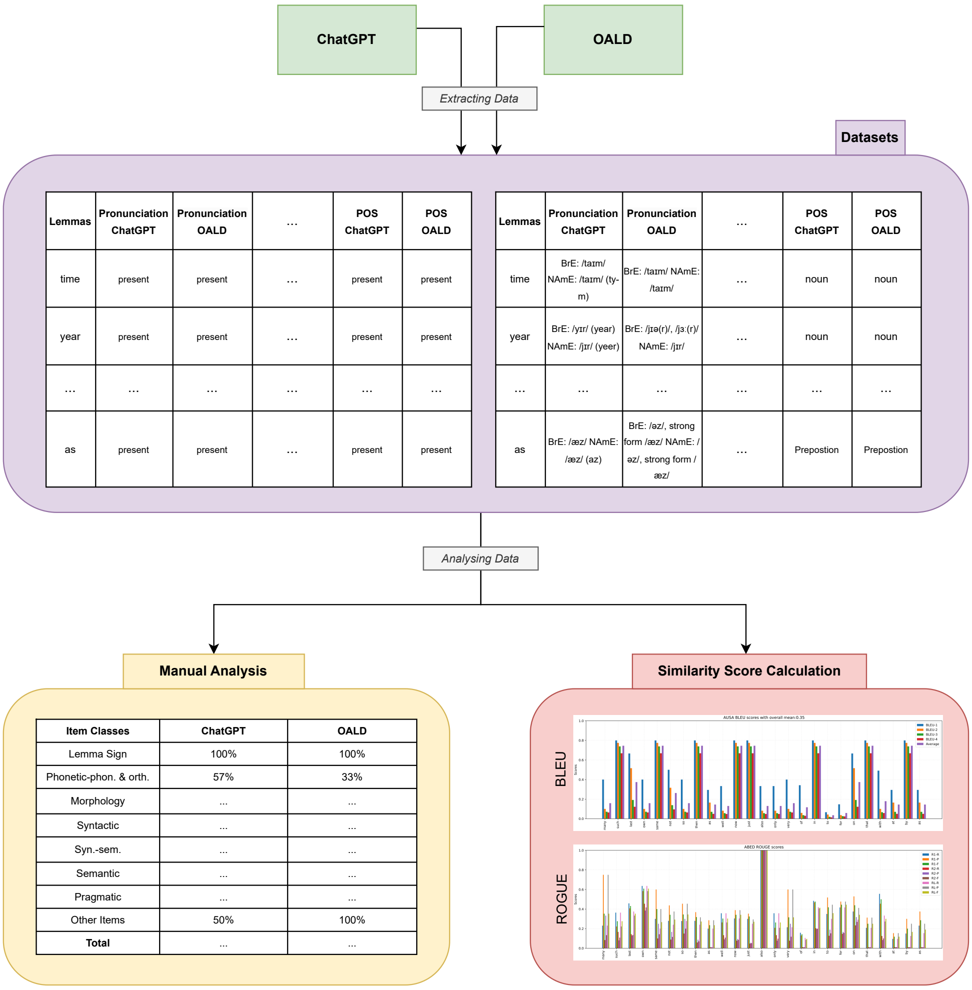

Figure 1: Compering ChatGPT and OALD. First, we extract the information from both platforms manually. This yields two datasets. The first describes the presence or absence of the lexicographical items, while the second contains the actual answers from both tools. Then, we analyse our datasets using Wiegand's item classes and similarity scores. Last, the results are presented as tables for the the former and as figures for the latter.

4.1 Lexicographical Items Regarding LES

Table 1 shows that both ChatGPT and OALD can provide LES to all of our selected lemmas. When providing dictionary information, ChatGPT displays this item or headword in a plain format without any typographical indicators such as font-style, font-size, or colors that make it more invisible than any other information within the entry. Prompt 1 shows that the headword appears within quotation marks ("...") in the answer. In contrast, OALD displays the lemma sign in bold and dark blue color at the top of each entry, making it highly visible and distinct from other elements. The font size is adjusted to ensure effective emphasis. Furthermore, the presentation of the headword is not influenced by the different parts of speech.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Item giving the form of LES	50	100%	100%

Table 1: The percentage result of ChatGPT and OALD's capabilities to provide LES shows that both platforms score 100%. This is essential because the lemma sign is a prerequisite for including an entry in a dictionary.

4.2 Lexicographical Items Regarding PPOI

According to the information presented in Table 2, it is evident that both ChatGPT and OALD provide the pronunciation of all our lemmas using the International Phonetic Alphabet (IPA) transcription system. However, ChatGPT does not include the pronunciation as part of the dictionary entry on the first attempt as shown in Prompt 1. To access the pronunciation, we must specifically ask, but once prompted, ChatGPT delivers just like OALD as demonstrated in Prompt 3. One area where ChatGPT falls short compared to OALD is that OALD provides audio files for users to imitate the sounds. Instead of audio files, ChatGPT compensates by offering detailed instructions on how to pronounce words which related to the lexicographical items giving *accent*, *vowel quantity marker*, *syllable* and *syllable division* which are lacking in OALD. Regarding *item giving the spelling*, the lemma sign serves not only as a means to identify a headword in a dictionary, but it also functions as a reference point for the word's spelling. As a result, both ChatGPT and OALD are equipped to provide this lexicographical item for 100% of our lemmas.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Pronunciation	50	100%	100%
Accent	50	14%	–
Vowel quantity marker	50	100%	–
Syllable	50	14%	–
Spelling	50	100%	100%
Syllable division	50	14%	–

Table 2: The comparison results for PPOI indicate that ChatGPT matches OALD's 100% score for pronunciation and spelling items. Additionally, ChatGPT can provide a 100% score for the vowel quantity marker, which is not a feature possessed by OALD. However, for the items giving accent, syllable, and syllable division, ChatGPT only provides 14%, whereas OALD does not include these features.

Prompt 3: Provide pronunciation in American and British English for "people"

The pronunciation of "people" is generally the same in both American and British English. It is pronounced as \'pi:pl \(PEE-pul) with stress on the first syllable in both dialects.

4.3 Lexicographical Items Regarding MOI

Table 3 demonstrates the capabilities of ChatGPT and OALD for providing data on lexicographical items in this class. *Item giving gender*, not applicable in English, is excluded from our criteria. Regarding *inflexion*, ChatGPT and OALD offer inflection for lemmas within certain parts of speech limitations. ChatGPT covers 77.5% of lemmas across multiple parts of speech, while OALD covers 42.5%, primarily for verbs and nouns. For *degrees of comparison*, adjectives and adverbs in English are considered. Among 50 lemmas, we focus on 20 from POS adjective and adverb. ChatGPT provides information for 11 lemmas (55%), mostly adjectives (see example Prompt 4), while OALD only covers four lemmas (20%), three adjectives, and one adverb. All 50 lemmas can have *compounds* and *word families* identified by ChatGPT (see example Prompt 5). Note that some selected lemmas may not form compounds, which ChatGPT can indicate. OALD provides compounds and word families only for noun lemmas (20%), with certain compounds listed separately. Nouns alone offer *item giving number*. ChatGPT provides plural forms for all ten nouns (100%), while OALD only provides only irregular plural forms for two lemmas: *man* and *child* (20%). Regular plural forms ending in *-s* or *-es* are not covered by OALD.

Lexicographical Item	Lemma count	ChatGPT	OALD
Inflexion	40	77.5%	42.5%
Degrees of comparison	20	55%	20%
Compound	50	100%	20%
Word family	50	100%	20%
Number	10	20%	4%

Table 3: The comparison of MOI reveals that ChatGPT and OALD can provide all the lexicographical items in this class. However, the ability of ChatGPT and OALD to provide information in this class depends on the POS of the lemmata, which affects the percentage of information provided. In terms of MOI, ChatGPT is capable of providing information for a greater number of lemmata than OALD.

Prompt 4: Provide degrees of comparison for "good"

The degrees of comparison for "good" are:

- Positive degree: good
- Comparative degree: better
- Superlative degree: best

Prompt 5: Provide compounds and word family for verb "be"

The verb "be" is an irregular verb and does not typically form compounds. However, here are some related words in its word family:
Word family:

- am (present tense first person singular form)
- is (present tense third person singular form)
- are (present tense plural form and second person singular form)

4.4 Lexicographical Items Regarding SYI

Table 4 indicates that both ChatGPT and OALD can provide all the lexicographical items related to this item class, achieving a 100% coverage. Regarding *part of speech* ChatGPT includes all the relevant POS in a single entry following by the senses as demonstrated in Prompt 1, while OALD tends to present them separately in each individual entry, such as with the lemma ”work” that can function as both a noun and verb. Regarding *item giving valency* ChatGPT displays valency using argument structure[3] and indicates whether a verb is *transitive* or *intransitive* (see example Prompt 6). OALD provides information on valency for the same verb, but only for the latter. For example, for the sense of *do* as a transitive verb, OALD highlights *do something* in bold to indicate that the verb requires an object or thing to receive the action. *Item giving adjective distribution* is limited to POS adjective. Both ChatGPT and OALD present this data through example sentences and collocations (see Prompt 7). Their approaches are similar.

Lexicographical Item	**Lemma count**	**ChatGPT**	**OALD**
Part of speech	50	100%	100%
Valency	10	100%	100%
Adjective distribution	10	100%	100%

Table 4: The percentage results of the capabilities between ChatGPT and OALD in providing information related to SYI show that both platforms are capable of providing 100% of all related lexicographical items.

Prompt 6: Provide valency for verb ”do”

The verb ”do” is a transitive verb and its valency is 2, which means it requires two arguments to form a complete sentence. The first argument is the subject of the sentence, and the second argument is the object that receives the action of the verb.

[Example sentences]

The verb ”do” can also be used as an auxiliary verb to form questions and negative statements, in which case it does not take an object. In these cases, the valency of the verb is considered to be 0.

[Example sentences]

[3] This refers to the number of core arguments or complements required by a given verb to form a grammatically correct sentence.

Prompt 7: Provide adjective distribution for "new"

The adjective "new" can be used to describe a variety of nouns across different categories, including:

1. New objects: new car, new phone, new book
2. New ideas: new concept, new theory, new approach
3. New experiences: new adventure, new opportunity, new relationship
4. New places: new city, new neighborhood, new country
5. New information: new data, new research, new findings

Overall, the distribution of "new" as an adjective is quite wide and versatile.

4.5 Lexicographical Items Regarding SYSI

In terms of the lexicographical items related to this item class, both ChatGPT and OALD demonstrate capability, covering *collocation*, *phraseme*, *proverb*, and *example* (see Table 5). However, there is a distinction between the example sentence types provided. ChatGPT offers *usage examples*, whereas OALD presents *citation texts* without the accompanying *reference*.

ChatGPT provides *collocations* for all 50 selected lemmas (100%) in a list format. Users can choose the lemma's position in the collocations. OALD, however, covers only 18 out of the 50 lemmas (36%), mostly nouns. Its collocations appear in separate information boxes for each definition, sourced from the Oxford Collocations Dictionary.

Regarding *Phrasame* ChatGPT can provide idiomatic expressions for all of the 50 lemmas (100%), although some of the expressions may not include the headword but refer to it by meaning. For POS other than nouns and verbs, ChatGPT may provide some kind of collocations instead of idioms which is not the concept of idiomatic expressions. In contrast, OALD can provide idiomatic expressions for 84% of the lemmas. OALD has a separate section dedicated to idioms located at the end of the dictionary entry. Users can also find a shortcut to this section at the top of the entry below the headword, POS, and pronunciation.

ChatGPT is capable of providing *proverb* for all the lemmas. However, some proverbs may not include the headword, and the accuracy of the provided proverbs is questionable. On the other hand, OALD can only provide this information for 26% of the lemmas, but the proverbs provided are accurate. OALD presents proverbs within the idioms section, indicated by *(saying)*.

ChatGPT and OALD are capable of providing *examples* for all of our lemmas (100%). ChatGPT usually offers ten example sentences for each lemma mixed from all the senses. On the other hand, OALD provides sense-specific examples of varying numbers. ChatGPT generates original example sentences using its own language proficiency derived from its training on large amounts of text, thus we consider the examples provided by ChatGPT to be *usage examples*. OALD offers a different type of examples referred to as *citation text* or *corpus examples*. These examples are usually sourced from the dictionary's corpora and other lexicographical sources. However, OALD does not include *item indicating the reference of the citation* within the entry. Since it's apparent that the examples are extracted from the BNC, this may be the reason why this information is not provided in OALD dictionary entry.

Lexicographical Item	**Lemma Count**	**ChatGPT**	**OALD**
Collocation	50	100%	36%
Phraseme	50	100%	84%
Proverb	50	100%	26%
Example	50	100%	100%
Usage example	50	100%	–
Citation text	50	–	100%
Reference of the citation	50	–	–

Table 5: In the comparison of SYSI capabilities, it was found that ChatGPT can provide collocations, phrasemes, and proverbs for all selected lemmas. In contrast, the percentages of OALD in providing these lexicographical items are consistently lower than those of ChatGPT. While both ChatGPT and OALD can provide example sentences, the approaches used by the two platforms to provide these examples differ.

4.6 Lexicographical Items Regarding SEMI

Regarding the semantic class, ChatGPT and OALD are capable of providing most of the items in this category. However, ChatGPT is unable to provide the *illustration* due to its nature as a text-based LLM. However, as Prompt 8 shows, it can provide detailed and descriptive explanations to help the users understand the concepts and ideas of the lemma. OALD occasionally includes pictures with the definitions in its dictionary entries. However, the entries for our 50 chosen lemmas do not have any illustrations. It is worth noting that OALD has the ability to provide illustrations, but not for the lemmas we selected. We indicate this by using plus symbol (+) in Table 6.

Additionally, neither ChatGPT nor OALD can provide the *item giving an equivalent*, as this belongs to one of the features of bilingual and multilingual dictionaries, that provides translations of the lemmas in the target language(s). However, since OALD is a monolingual learner's dictionary for English, this feature is not available in OALD. Similarly, we cannot compare this feature in ChatGPT for the same reason.

ChatGPT and OALD are both capable of providing *Item giving meaning* and *polysemy* for all the lemmas. In the cases of polysemy, both platforms enumerate each sense and present them in a list. However, OALD uses a restricted vocabulary, known as The Oxford 3000 Oxford University Press (n.d.), to ensure that definitions are easily understandable for EFL learners.

Both ChatGPT and OALD can provide *synonyms* and *antonyms*. But not all the dictionary entries in OALD possess this feature. Only 26% of our lemmas can provide synonyms and only 2% percent can provide antonyms. Out of the 50 lemmas we examine, only the lemma *new* provides an antonym. In OALD, synonyms are presented in two different ways: a fold-out box and list of synonymous words. On the other hand, ChatGPT presents both synonyms and antonyms in a list format, as demonstrated in Prompt 9 and can provide 100% of the lemmas.

Lexicographical Item	Lemma Count	ChatGPT	OALD
Meaning	50	100%	100%
Synonyms	50	100%	26%
Antonyms	50	100%	2%
Polysemy	50	100%	100%
Illustration	50	–	+
Equivalent	50	–	–

Table 6: Upon comparing the SEMI capabilities, it was discovered that both ChatGPT and OALD can provide complete information for the essential lexicographical items of meaning and polysemy, with a 100% success rate. Both platforms can also provide synonyms and antonyms, with OALD consistently shows lower percentages in providing this information for our selected lemmas. ChatGPT is unable to provide illustrations, while OALD has the capability to do so, although it is not available for our selected lemmas. There is no equivalent feature to be considered in this context.

Prompt 8: Provide illustration for "man"

I'm sorry, but I cannot provide illustrations as I am a text-based AI language model. An illustration for the word "man" could be a drawing, painting, or photograph that depicts a male human being. The illustration could show a man in a specific context, such as at work, at home, or in a social setting. It could also show a man in different stages of life, such as a young boy, middle-aged man, or elderly man. The purpose of the illustration is to provide a visual representation of the word "man" to help readers understand and visualize its meaning.

Prompt 9: Provide synonyms and antonyms for "day"

Synonyms for "day":

- daytime
- 24 hours
- sunlight
- ...

Antonyms for "day":

- night
- nighttime
- evening
- ...

4.7 Lexicographical Items Regarding PRAI

OALD typically includes pragmatic information within each individual sense, indicated before the definition in brackets in italic and light grey. An example of *diatechnic labelling* is *(computing)* provided within the entry of *at* for sense 15, referring to the symbol @ used in email addresses. This is the only lemma (2%) out of 50 lemmas that OALD can provide diatechnic label. ChatGPT is capable of providing information on diatechnic labels, but for certain lemmas, it responds with *No diatechnic labelling* and notes that the words are commonly used in general language rather than specialized terminology. It is worth mentioning that ChatGPT can provide diatechnic labels in other cases, but it is not applicable for our selected lemmas.

OALD can provide *diastratic label* for 12% of our 50 lemmas. This information can be found in the entries of certain lemmas such as *have*, *do*, *make*, *get*, *know*, and *well*. Labels such as *(slang)* can be found in these entries. ChatGPT, however, is not capable of providing information related to this label, giving the reason that it requires more context related to the headword.

In terms of *diafrequency labeling*, OALD does not provide this information for our selected lemmas within its dictionary entries. However, it should be noted that the lemmas in OALD are already commonly used and therefore do not require frequency labeling. In some cases, the entries may include a label such as *(rare)*, which refers to diafrequency. However, this does not apply to our selected lemmas. In contrast, ChatGPT provides diafrequency information for all lemmas, indicating whether they are *common* or *very common*.

Out of the 50 lemmas we analyzed, OALD provides *diaevaluative labelling* for eight of them, which accounts for 16% of the total. The labels used in OALD for diaevaluative purposes are denoted by phrases such as *(approving)* or *(disapproving)*. ChatGPT is also capable of providing this information, although it uses different labels. For our specific list of lemmas, ChatGPT indicates whether a word is *neutral* or *positive* in connotation, since there are no words with negative connotations in our list. However, it's important to note that due to ChatGPT's lack of contextual awareness, caution should be exercised when interpreting these labels.

OALD is capable of providing *diachronic labels* for 11 lemmas (22%) of the lemmas in our sample. These labels, such as *(old used)* and *(old-fashioned)* appear within the dictionary entries. Although ChatGPT is unable to provide diachronic labels for our selected lemmas, it is important to note that this is because the lemmas are still commonly used today. However, it is worth mentioning that ChatGPT has the capability to provide diachronic labels for other entries. When asked if it is possible to provide diachronic labels, ChatGPT indicated that terms such as *historic* or *archaic* are used for some entries.

OALD is capable of providing *diatopic labelling* for 48% of its lemmas, indicating regional varieties of English such as American English, Australian English, British English, Northern English, etc. This labelling is provided within individual senses rather than just for the headword. In contrast, ChatGPT can also provide diatopic labelling, but for our chosen lemmas, it only offers a *neutral* label since the words are universal and not associated with any particular region or culture. We consider this as ChatGPT is capable of providing diatopic label, but just not for our chosen lemmas.

OALD is typically able to provide *Item giving the diaintegrative labelling* for loanwords and their original language. However, since our chosen lemmas do not fall under this category, OALD entries do not include this label. Nevertheless, if we were to look up a lemma like "croissant," it would be labeled as *(from French)* for diaintegrative purposes. ChatGPT, on the other hand, explains that such labels would fall under etymology and can provide the word's origin instead. We consider ChatGPT unable to provide diaintegrative labelling

Regarding *Dianormative labelling*, OALD can provide this information for 10% of the 50 chosen lemmas from different parts of speech. OALD shows this information by presenting typical mistakes made by EFL learners with a crossed-out sentence alongside the correct version. However, ChatGPT is not able to provide this type of information, giving the reason that dianormative labelling is a complex process that requires knowledge of the social, cultural, and historical context of a language and its users. It involves identifying the norms and values associated with the use of certain words and how they may vary across different social groups or contexts. This is a task that requires human expertise and cultural knowledge.

Out of our 50 lemmas, OALD can provide *Item giving the diatextual labelling* for 10 of them (20%). The labels provided in OALD entries include *literally* or *figurative*, which indicate the intended meaning of larger textual units such as phrases, sentences, and definitions. However, ChatGPT cannot offer this type of information as diatextual labelling is not applicable to individual words. It is a labelling system that is used to analyse and describe larger textual units, such as those found in OALD entries.

OALD provides *diamedial labels* such as *spoken* or *written*, but surprisingly, none of our chosen lemmas are labeled as such in the dictionary entries. It is possible that this is because they are commonly used words. However, ChatGPT can provide diamedial labels for all of our chosen lemmas, using terms like *spoken*, *written*, *news*, and *academic* to indicate this information. However, it is important to note that ChatGPT provides all four above mentioned labels to all of our lemmas, which may lead to inaccurate information. Users of ChatGPT should be aware of this potential issue.

According to Wiegand et al. (2010), OALD includes some additional diasystem labels, such as *diaphasic labelling*, which indicates whether a lexeme is considered *formal* or *informal*, and *diaattitudinal labelling*, which includes indications such as *humorous* and *ironic*. However, these labels are not included in item classes or lexicographical items as defined by Wiegand.

Lexicographical Item	**Lemma Count**	**ChatGPT**	**OALD**
Diatechnic labelling	50	+	2%
Diastratic labelling	50	–	12%
Diafrequency labelling	50	100%	+
Diaevaluative labelling	50	100%	16%
Diachronic labelling	50	+	22%
Diatopic labelling	50	+	48%
Diaintegrative labelling	50	–	+
Dianormative labelling	50	–	10%
Diatextual labelling	50	–	20%
Diamedial labelling	50	100%	+

Table 7: The pragmatic class percentage outcome indicates that OALD excels in providing all types of labelling, whereas ChatGPT falls short in providing information on this class. This is mainly because the expertise and cultural knowledge of human lexicographers are essential for such labelling, and ChatGPT requires more context in order to provide related information.

4.8 Lexicographical Items Regarding OTI

The two lexicographical items, etymology and cross-reference, do not fall under any of the item classes mentioned previously. As a result, they are categorized separately by Wiegand. Although providing the *etymology* of a word is not a typical feature found in monolingual learner's dictionaries, most printed dictionaries do not provide this information for every lemma in their dictionaries due to space constraints. However, both ChatGPT and OALD are capable of providing etymology for all 50 lemmas with a 100% success rate. In OALD, the etymological information is available in a separate fold-out section titled *Word Origin* The manner in which ChatGPT provides etymological information is illustrated in Prompt 10.

Regarding *Item giving a cross-reference*, ChatGPT generates responses one by one and does not have the capability to refer to external sources of information within a single answer. On the other hand, OALD has a cross-reference feature marked in blue for all of the lemmas, and by hovering over it with the cursor, the linked source is immediately displayed. Additionally, users can double-click on any word in the dictionary entry, which will lead them to the entry for that word. This feature is particularly useful when encountering unfamiliar words in a definition.

Lexicographical Item	**Lemma Count**	**ChatGPT**	**OALD**
Etymology	50	100%	100%
Cross-reference	50	–	100%

Table 8: The percentage result of the OTI comparison shows that OALD is capable of providing information for both lexicographical items in this class, achieving a score of 100%. In contrast, ChatGPT can only provide information related to one item, demonstrating a limitation in its capabilities. This is due to the fact that ChatGPT does not possess the feature to refer to external sources.

Prompt 10: Provide word origin for "child"

The word "child" comes from the Old English word "cild", which meant "a fetus, an infant, or a young person". The word is thought to be derived from the Proto-Germanic word "*kiltham", which meant "womb" or "uterus." This word is related to the Old Norse word "kildr" and the German word "Kinder".

4.9 Additional Lexicographical Features

In this section, we have compiled extra lexicographical features from OALD that were not classified by Wiegand but can benefit EFL learners, as illustrated in Table 9. We compare their availability in ChatGPT and note that while both dictionaries offer these features, ChatGPT requires specific prompts for providing certain information.

5. Similarity Scores

In this section, we present our key findings obtained through similarity scores. We first discuss the BLEU results, followed by the ROUGE results. Additionally, we provide

Lexicographical Item	Lemma Count	ChatGPT	OALD
CEFR level	50	100%	100%
Topic	50	+	100%
Abbreviation	50	10%	10%
Cultural Information	50	+	2%
Political statement	50	+	2%
Notes on usage	50	+	28%

Table 9: The percentage comparison results for the compiled additional lexicographical features in OALD demonstrate that ChatGPT is capable of providing information related to those lexicographical items as well.

interpretations for each of the findings. We calculate the similarity scores on the dataset containing the responses from ChatGPT and data from OALD. This contains the following lexicographical items: LES, AUSA, RA, WAA, ABED, and ETYA. Not counting the lemma sign, all lexicographical items define a category. For each category, we collect the answers of ChatGPT and OALD. Since we have 5 categories, this yields $1 + 5 \cdot 2$ many columns (features) with 50 rows for our dataset.

The similarity scores calculated by iterating over all categories row by row. For each row, the calculations return a vector $v \in \mathbb{R}^{1\times l}$, where l is determined by the method used. When using BLEU, the last element of v is the average of the previous elements. The overall mean is calculated by taking the average of the last element in every v in a given category. For each category, our results are described by a matrix $X \in \mathbb{R}^{50\times l}$. These matrices are visualized in the next sections.

5.1 BLEU Scores

This section contains our most important BLEU results and interpretations of these. In the calculations, we have used different n-grams with $n = 1, 2, 3, 4$. Also, the averages of all the n-grams are provided.

The used weights for BLEU-1 to BLEU-4 in order are $w_1 = [1]$, $w_2 = [0.5, 0.5]$, $w_3 = [0.33, 0.33, 0.33]$, and $w_4 = [0.25, 0.25, 0.25, 0.25]$. Additionally, we use a smoothing function Chen & Cherry (2014). This is needed because if there is no n-gram overlap for any order of n-grams, BLEU returns 0. Due to the precision for the order of n-grams without overlap is 0, and the geometric mean in the final BLEU score computation multiplies the 0 with the precision of other n-grams. This results in 0 independently of the precision of the other n-gram orders. Specifically, we use ϵ-smoothing which adds a small ϵ value to the numerator when it is 0 in Equation (3). In our case, $\epsilon = 0.1$.

BLEU scores consistently show that lexicographical items containing more n-grams receive lower scores, indicating that ChatGPT's responses match better with single words (unigrams) than with phrases (multigrams). This trend is observed across all evaluated lemmas and the five chosen lexicographical items. Figure 2 highlights that ABED's complex text elements result in lower scores, compared to single-word representations like RA or WAA. The bar charts clearly demonstrate that shorter candidate texts, like those in AUSA, RA, or WAA, receive higher scores, while longer ones, like ABED and ETYA,

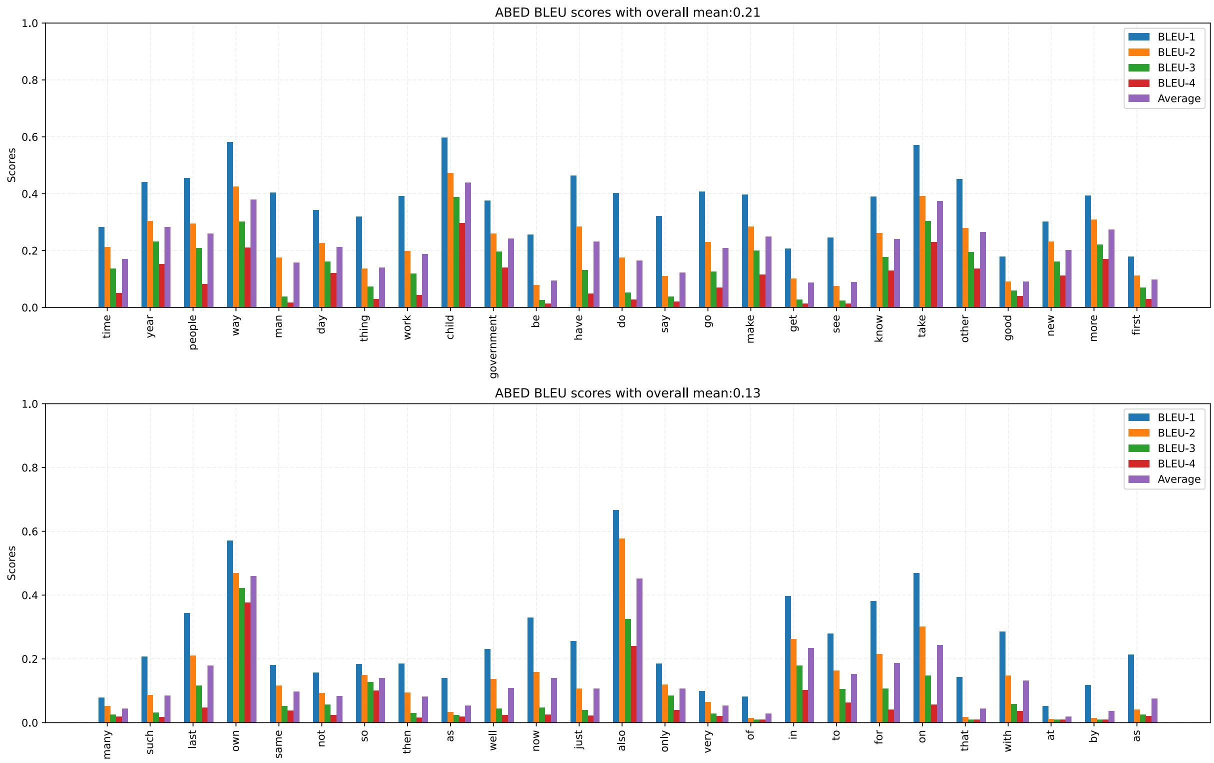

Figure 2: BLEU scores for ABED category. Each bar represents five scores, from left to right: BLEU-1 to BLEU-4, and the average. The top figure shows the first 25 lemmas, while the bottom figure displays the remaining 25. The overall mean for the data is indicated in the title of both figures.

receive lower scores. In fact, ETYA had the longest candidate text and obtained the lowest BLEU score among the five selected lexicographical items.

5.2 ROUGE Scores

This section contains our most important ROUGE results and interpretations of these. We calculate ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) with recall, precision, and F1 scores for each method. On the figures, these values are denoted with their first letter. For example, ROUGE-1 recall is abbreviated to R1-R.

ROUGE scores consistently indicate that bigrams (R2) have the lowest scores compared to unigrams and RL for all of our lemmas, and all five selected lexicographical items. Remarkably, all n-grams scores for the lemma *also* achieve a perfect score of 1.0, as illustrated in Figure 3. This is due to the 100% match between the candidate and reference texts. ChatGPT and OALD provide the same definition, *in addition; too*, with identical punctuation. The trend of R2 scores being the lowest is consistent across all of our ROGUE score charts.

6. Evaluation

In this section, a summary of the results obtained from manual analysis and similarity score tests are presented. The manual analysis included a thorough evaluation of the

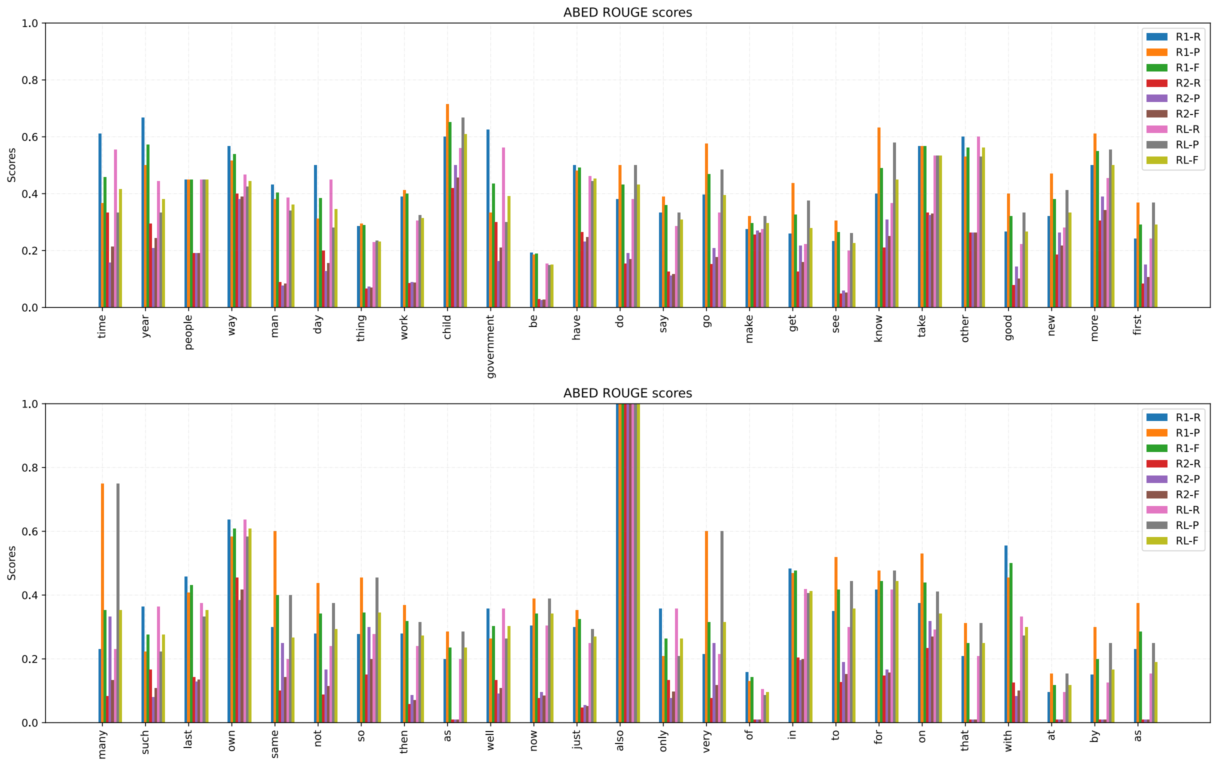

Figure 3: ROUGE scores in the ABED category. We visualize nine values for each lemma. In order from left to right, these are ROUGE-1 (R1), ROUGE-2 (R2), and ROUGE-L (RL) for each we provide recall (R), precision (P), and F1 (F) scores.

capabilities of ChatGPT and OALD. This was done by comparing the percentages of each item class between the two sources. Additionally, the analysis presented the total average score for all item classes.

The average capabilities of ChatGPT and OALD for each item class are presented in Table 10. ChatGPT has an average score of 68% in providing dictionary information for the 50 chosen lemmas. This is 11% higher than the average score of OALD, which is 57%. Both ChatGPT and OALD can provide 100% of related information for the LES and SYI item class. ChatGPT has a higher average score than OALD for all item classes except OTI, where OALD has a perfect average score of 100%.

The similarity scores of both BLEU and ROGUE suggest that higher scores are attained when candidate texts are evaluated at the unigram level, with those containing only one word unit reaching a perfect score of 1.0. Conversely, longer word units tend to receive lower scores, as demonstrated by the lower scores of lexicographical items ABED and ETYA. Of the five chosen items, AUSA holds the highest similarity scores followed by RA and WAA, while ETYA has the lowest scores indicating the least similarity to the reference text.

7. Conclusions and future work

The paper compares the abilities of ChatGPT and OALD for lexicographical purposes, specifically focusing on microstructural elements. The study finds that ChatGPT performs better on average than OALD in providing information related to lexicographical items,

Item Class	Average Score	
	ChatGPT	**OALD**
LES	100%	100%
PPOI	57%	33%
MOI	71%	21%
SYI	100%	100%
SYSI	71%	49%
SEMI	67%	38%
PRAI	30%	13%
OTI	50%	100%
Total	68%	57%

Table 10: The comparison of ChatGPT and OALD in providing information related to each item class yielded average scores. Both platforms achieved 100% for LES and SYI. ChatGPT had overall higher average scores than OALD in all item classes, except for OTI.

indicating its potential as a learner's dictionary. However, ChatGPT has limitations such as the absence of contextual information and limited interactivity, which are important aspects of learner's dictionaries. The paper also measures the similarity between the data generated by ChatGPT and OALD using BLEU and ROUGE metrics. While single words show high similarity between the two tools, responses consisting of multiple words differ significantly, suggesting variations in phrase construction and data presentation. The study acknowledges the need for further research on ChatGPT as a learner's dictionary, including potential prompts for lexicographical tasks, the development of evaluation criteria, comparisons with other learner's dictionaries, and assessment of response accuracy for different lexicographical items. Despite the limitations, the paper concludes that ChatGPT shows promise as a language learning tool and an efficient lexicographic aid for EFL learners.

8. References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Brants, T., Popat, A.C., Xu, P., Och, F.J. & Dean, J. (2012). Large language models in machine translation. US Patent 8,332,207.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp. 1877–1901.

Chen, B. & Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the ninth workshop on statistical machine translation*. pp. 362–367.

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J. & Liu, R. (2019). Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Davies, M. & Gardner, D. (2013). *A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists.* Routledge.

Engelberg, S. & Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung: Stauffenburg-Einführungen.* Stauffenburg (Tübingen).

Herbst, T. (1996). On the way to the perfect learners' dictionary: a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE. *International Journal of Lexicography*, 9(4), pp. 321–357.

Hornby, A.S. (2019). Oxford advanced learner's dictionary. URL https://www.oxfordlearnersdictionaries.com/definition/english/.

Ivančič, I. & Fabijanić, I. (2017). Structural Development of Oxford Advanced Learners' Dictionary. *Journal of Literature and Art Studies*, 7(5), pp. 588–607.

Kant, N., Puri, R., Yakovenko, N. & Catanzaro, B. (2018). Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*.

Lin, C.Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81. URL https://aclanthology.org/W04-1013.

OpenAI (2023). ChatGPT. https://openai.com/blog/chatgpt.

Oxford Text Archive (2007). British National Corpus. http://ota.ox.ac.uk/desc/2554. [dataset].

Oxford University Press (n.d.). The Oxford 3000™ by CEFR level. Online. URL https://www.oxfordlearnersdictionaries.com/wordlist/english/oxford3000-cefr/. Accessed on March 30, 2023.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. pp. 311–318.

Pilault, J., Li, R., Subramanian, S. & Pal, C. (2020). On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 9308–9319.

Poesia, G., Polozov, O., Le, V., Tiwari, A., Soares, G., Meek, C. & Gulwani, S. (2022). Synchromesh: Reliable code generation from pre-trained language models. *arXiv preprint arXiv:2201.11227*.

Rikk, R., Várkonyi, T.A., Borsi, Z.R., Pintér, B. & Gregorics, T. (2022). Generating Algorithmic Patterns from Semi-structured Input Using a Transition-Based Neural Network. In *Intelligent Systems and Applications: Proceedings of the 2022 Intelligent Systems Conference (IntelliSys) Volume 2*. Springer, pp. 819–834.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. & Schmidt, D.C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wiegand, H.E. (1989). Der Begriff der Mikrostruktur: Geschichte, Probleme, Perspektiven. In *Handbücher zur Sprach- und Kommunikationswissenschaft*, volume 1. Berlin: Walter de Gruyter, pp. 409–462.

Wiegand, H.E., Beißwenger, M., Gouws, R.H., Kammerer, M., Storrer, A. & Wolski, W. (2010). Wörterbuch zur Lexikographie und Wörterbuchforschung. *Berlin/New York*.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P. & Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

A. Similarity Scores Results

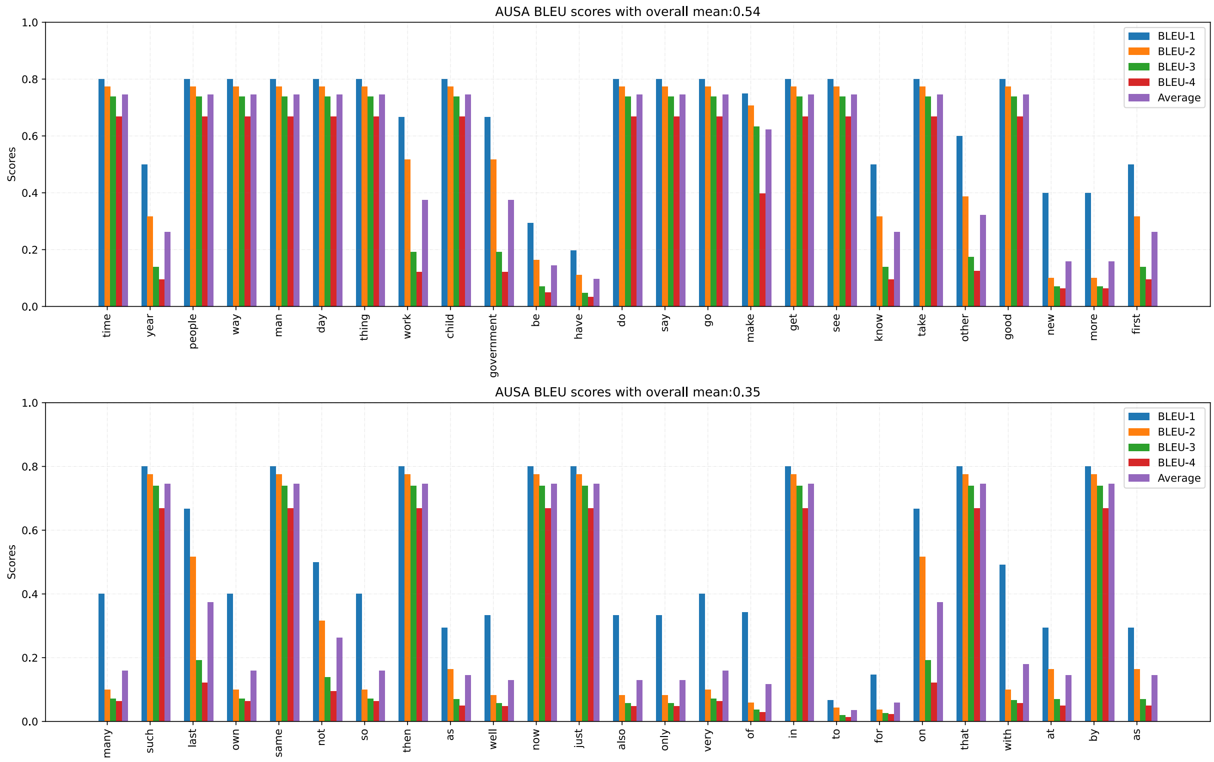

Figure 4: BLEU scores for AUSA category. The score results indicate that the more n-grams present in both candidate and reference texts, the lower the score. Furthermore, the data for AUSA contains a comparable amount of word units. As a consequence, The bars from BLEU 1 to 4 for the majority of the chosen lemmas on the graph show quite similar scores.

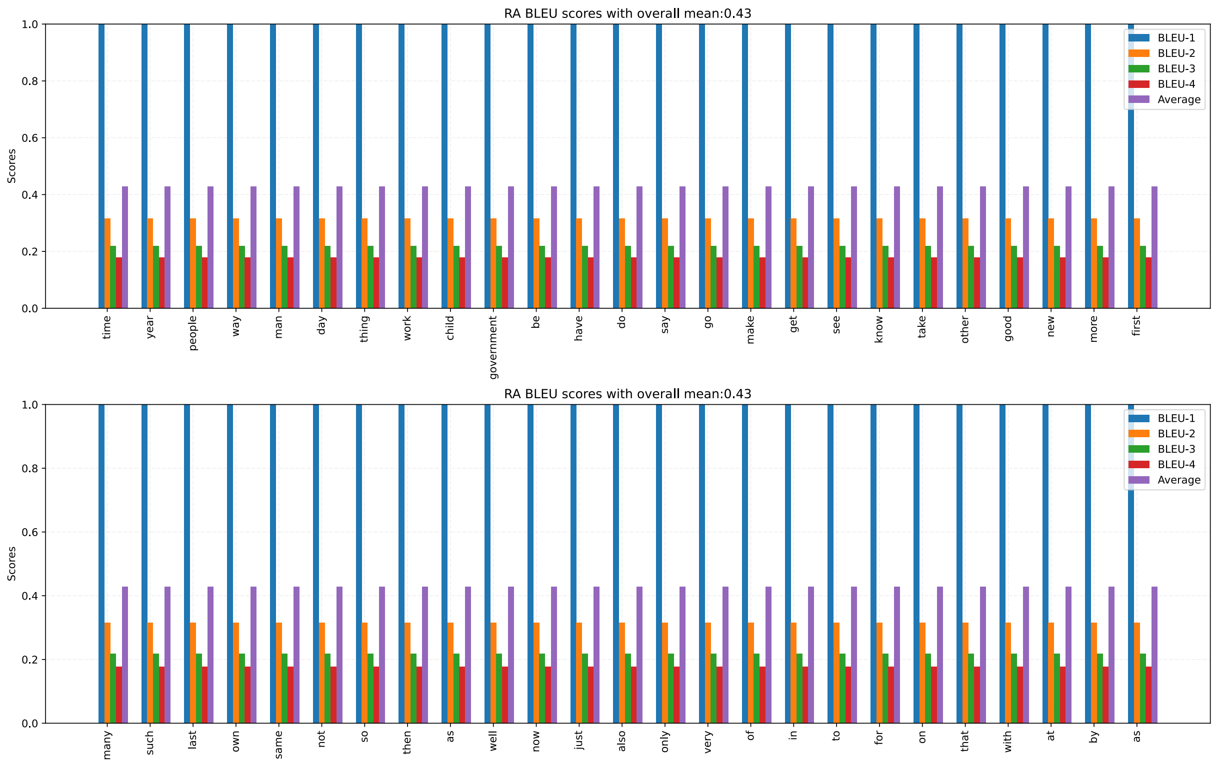

Figure 5: BLEU scores for RA category. Since the data in RA comprises a single word unit that also functions as a lemma sign, the BLEU 1 score is perfect at 1.0, signifying a complete match between the candidate and reference texts. Moreover, all lemmas attain equivalent scores across all BLEU scores from 1 to 4.

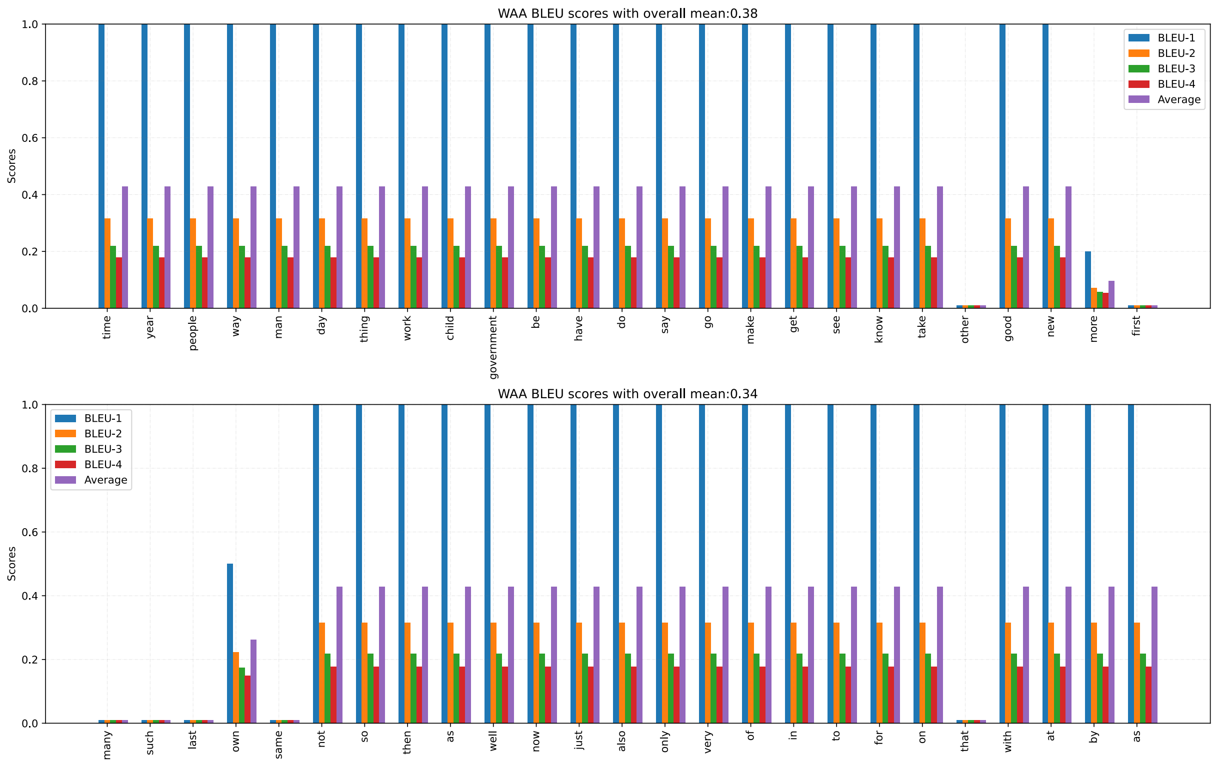

Figure 6: BLEU scores for WAA category. WAA's data comprises a single word unit, leading most lemmas to achieve a perfect unigram score of 1.0. However, certain lemmas nearly attain a score of 0.0 for the same BLEU 1 score, as ChatGPT and OALD assign them different parts of speech.

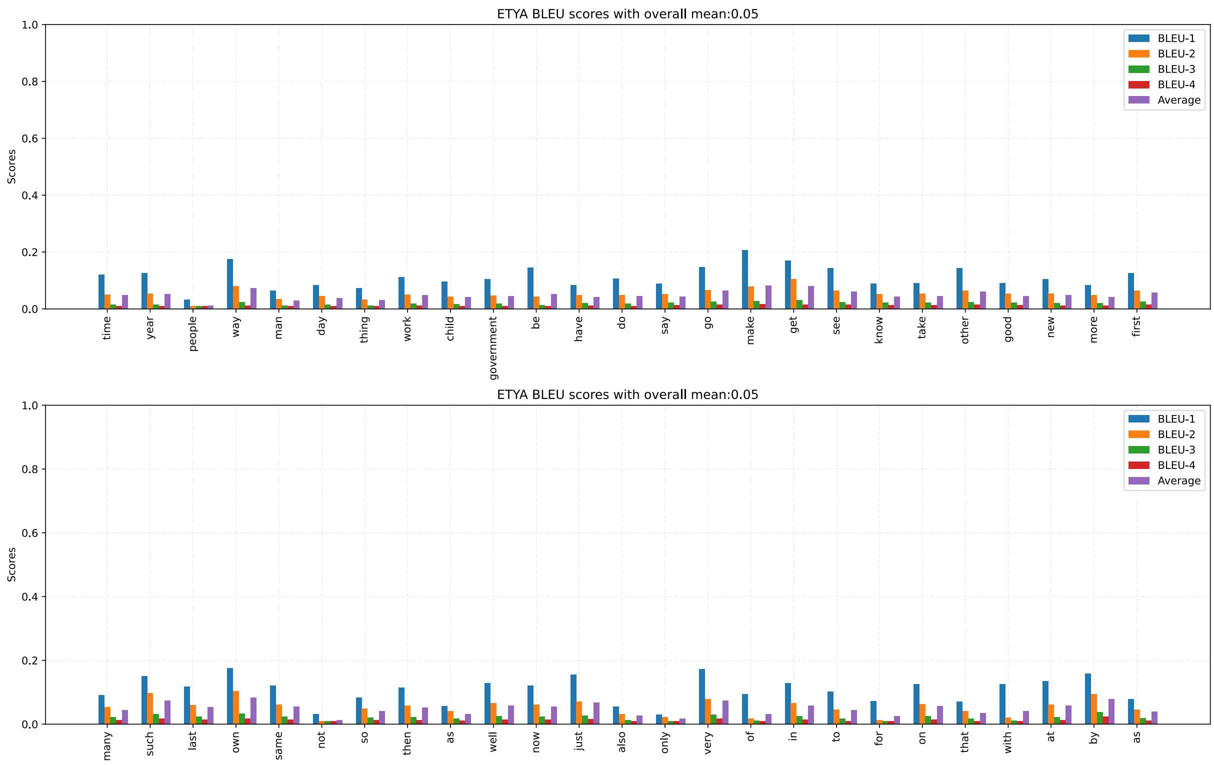

Figure 7: BLEU scores for ETYA category. ETYA contains the most word units among all the selected lexicographical items. The data related to ETYA from both ChatGPT and OALD refer to common origins of the lemmas. However, the formulation of data differs significantly, leading to a considerably lower overall score in this category. When users look up etymological information using ChatGPT, they will still receive the same information pertaining to the lemma.

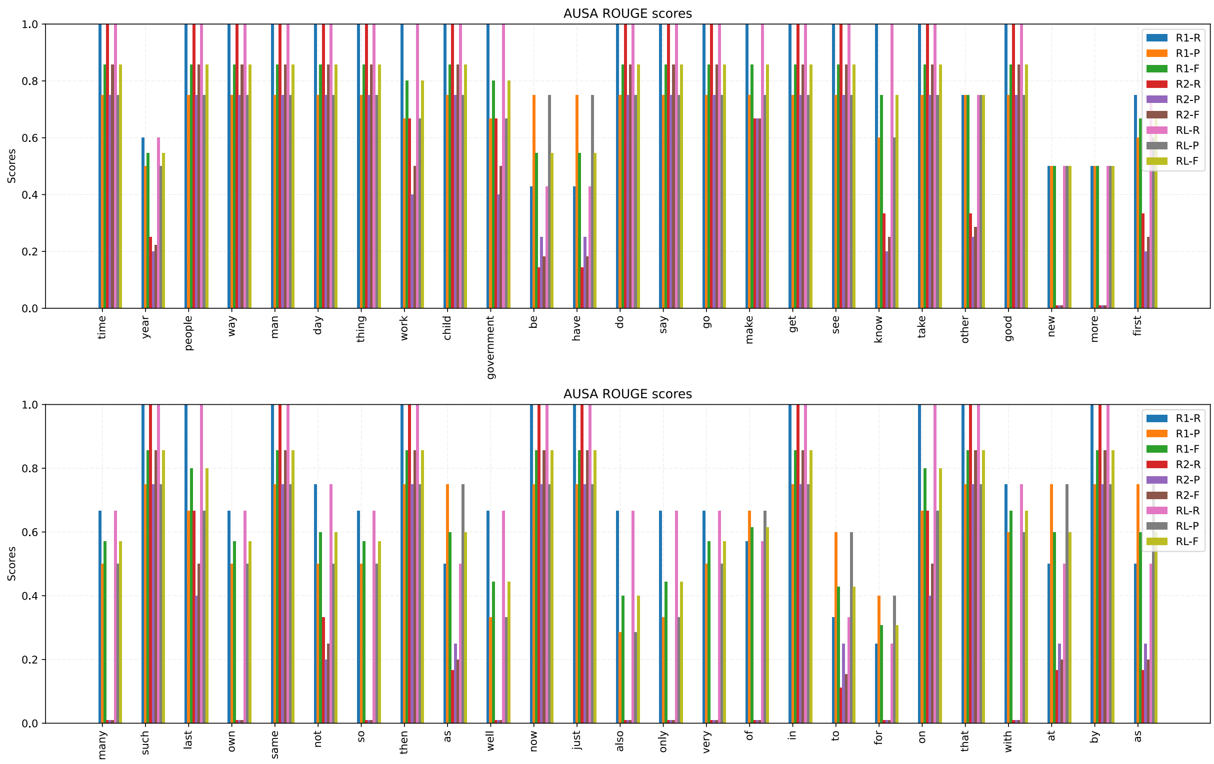

Figure 8: ROUGE scores in the AUSA category. Upon examining the charts, it is apparent that there is a consistent trend in the Recall (R) bars for R1, R2, and RL, with almost all bars reaching a perfect score of 1.0. This trend is particularly notable in the context of our analysis of AUSA data, where we observe high overall similarity scores for all the lemmas.

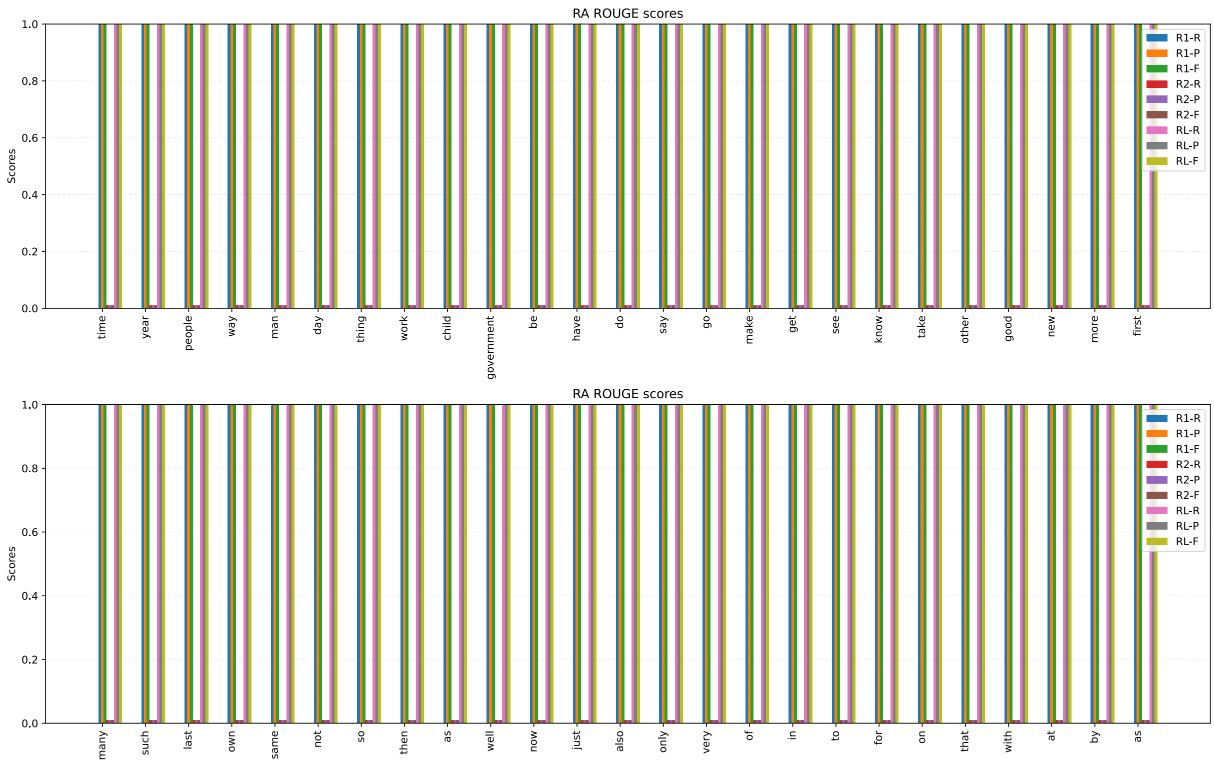

Figure 9: ROUGE scores in the RA category. It is evident that R2 score is not applicable for the data belonging to this category since it consists of only unigrams and not bigrams. Therefore, since the longest word units (RL) are also unigrams, all the lemmas achieve a perfect match score of 1.0 for R1 and RL.

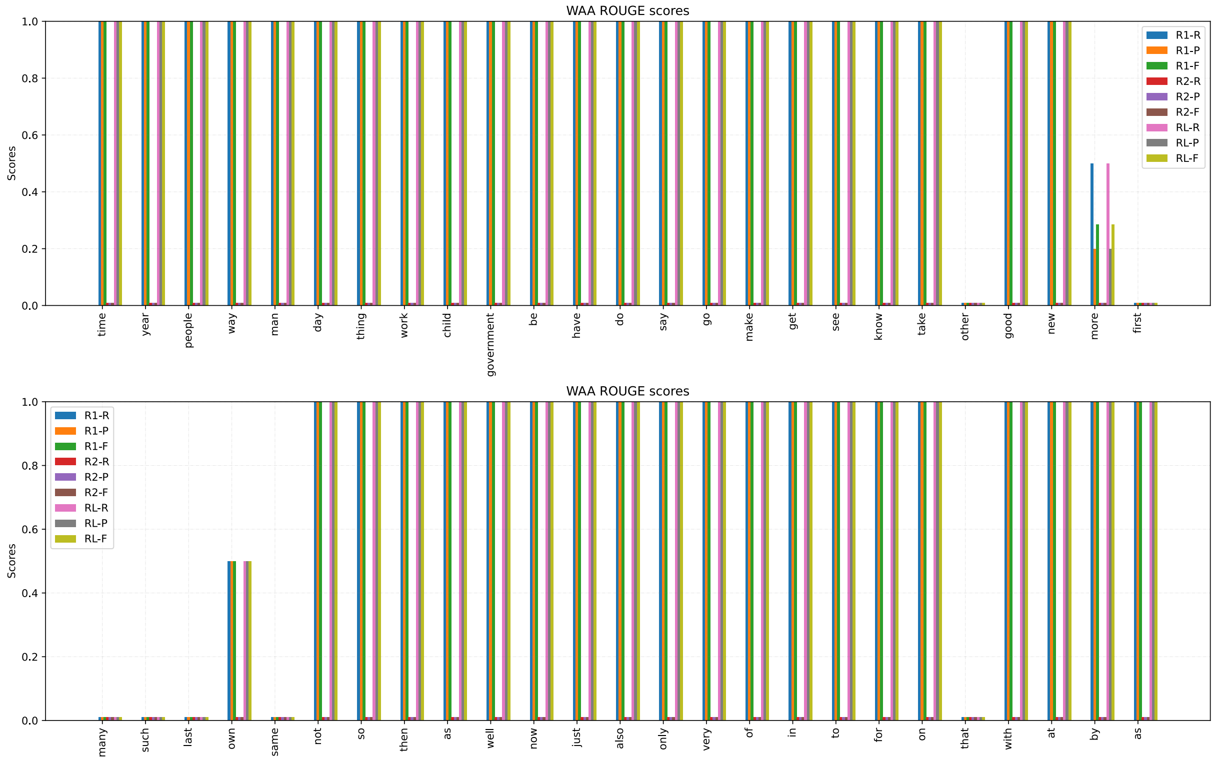

Figure 10: ROUGE scores in the WAA category. The majority of the data in this category comprises of one-word units. As a result, the majority of our lemmas reach a perfect score of 1.0 for R1 and RL. R2 scores are not applicable. However, some of our lemmas receive a score of 0.0 in R1 and RL. This is due to the fact that ChatGPT and OALD provide different part-of-speech information for these lemmas.

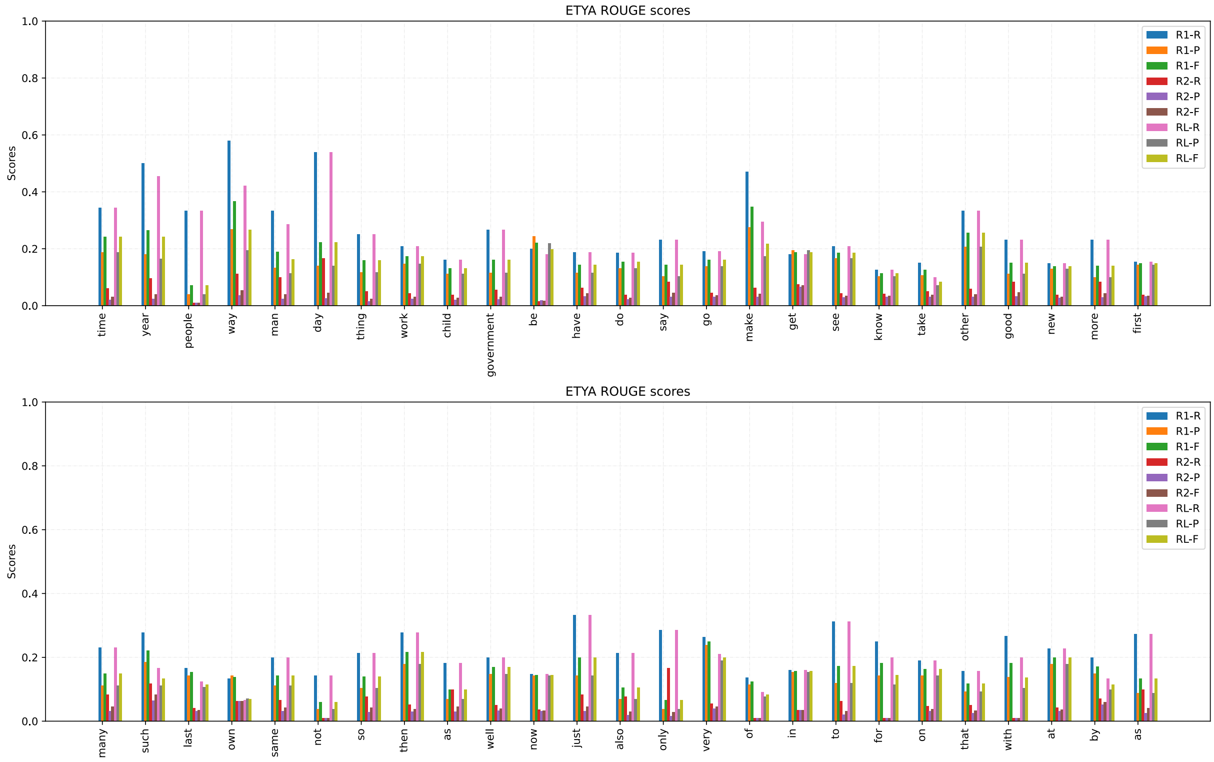

Figure 11: ROUGE scores in the ETYA category. The data in this category contains the highest number of word units, but the bar charts for all lemmas show scores of no more than 0.6, with the majority scoring less than 0.2. This suggests a significant difference between the etymological data in the reference and candidate texts.