MOTIF: Contextualized Images for Complex Words to Improve Human Reading

Xintong Wang[*1] **, Florian Schneider**[*1]**, Özge Alaçam**[2]**, Prateek Chaudhury**[3]**, Chris Biemann**[1]
[1]Universität Hamburg, [2]Universität Bielefeld, [3]Indian Institute of Technology Delhi
{xintong.wang, florian.schneider-1, christian.biemann}@uni-hamburg.de
oezge.alacam@uni-bielefeld.de, prateekchaudhury@gmail.com

Abstract

MOTIF (MultimOdal ConTextualized Images For Language Learners) is a multimodal dataset that consists of 1125 comprehension texts retrieved from Wikipedia Simple Corpus. Allowing multimodal processing or enriching the context with multimodal information has proven imperative for many learning tasks, specifically for second language (L2) learning. In this respect, several traditional NLP approaches can assist L2 readers in text comprehension processes, such as simplifying text or giving dictionary descriptions for complex words. As nicely stated in the well-known proverb, sometimes "a picture is worth a thousand words" and an image can successfully complement the verbal message by enriching the representation, like in Pictionary books. This multimodal support can also assist on-the-fly text reading experience by providing a multimodal tool that chooses and displays the most relevant images for the difficult words, given the text context. This study mainly focuses on one of the key components to achieving this goal; collecting a multimodal dataset enriched with complex word annotation and validated image match.

Keywords: Context-dependent image retrieval, L2 reading material, Complex word identification

1. Introduction

Whether in human cognitive processes or computational systems, multimodal information is crucial for adequate concept formation, accordingly for language acquisition. Babies learn their native language by combining words with visual cues, e.g., the sound of the word "cat", an image of a cat, and a cat sound are all essential for the concept *"cat"*. Over the past two decades, literature has provided convincing evidence on the facilitating role of cross-modal information in (Ecalle et al., 2009; Dalton and Grisham, 2011; Hahn et al., 2014; Gerbier et al., 2018; Xie et al., 2019; Albahiri and Alhaj, 2020).

Although words are powerful symbolic representations, explaining the message (communicative intent) verbally yields unwieldy over specified sentences. Successful communication in daily communication settings usually involves linguistic information accompanied by other modalities like visual representations, gestures, or audio. The advantage of multimodal information holds for second language (L2) acquisition. Modern language learning applications or dictionaries like Babble[1] or Duolingo[2] benefit from multimodality by using audio, visual illustrations, and video to enhance the L2 learning experience.

There are several approaches to assist non-native speakers in their reading activities. Through Lexical Simplification (LS), complex words can be replaced with simpler alternatives while preserving the meaning and syntactic function. It has been shown that LS leads to better text comprehension, improving text recall, especially for L2 learners at lower proficiency levels (Rets and Rogaten, 2021).

Instead of automatically simplifying the text, another approach would be to provide additional information about the complex word/phrase. This actively involves the reader by inviting her to process the supplementary/complementary information. Such a system can provide readers with dictionary definitions of the complex words in a more straightforward form. As we also address in this study, a more holistic approach can utilize multimodal information, e.g., an image, that depicts the information represented in the language modality. This would not only improve the understanding of the text but also facilitate the acquisition of new (complex) words by providing multimodal cues.

2. Application Scenario

Our ultimate goal is to provide language learners with a multimodal tool that chooses and displays the most relevant images for the difficult words, given the context, to support their reading comprehension. To prevent any misunderstanding, a contextualized image should be chosen carefully to be in line with the information given in the rest of the sentence as much as possible. To achieve this, three central components should be addressed; (i) a multimodal dataset enriched with complex word annotation and contextualized images, (ii) complex word identification, and (iii) context-sensitive image retrieval. In this paper, although we touch upon the last two items, we mainly focus on the dataset of comprehension texts for Language learners enriched with images for the complex words. The texts provided in this dataset target English language learners below

[*]These authors contributed equally to this work.
[1]https://www.babbel.com/
[2]https://www.duolingo.com/

B1 proficiency level according to the Common European Framework of Reference for Languages (Council of Europe, 2001).

The following example illustrates what the model is expected to do. The text piece below[3] provides general information about stingray characteristics. Our focus is on their ability to camouflage in a sandy bottom, with the word "camouflage" as a complex word for our L2 readers.

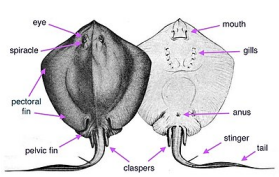
> Stingrays use a wide range of feeding strategies. ... *Stingrays exhibit a wide range of colors and patterns on their dorsal surface to help them* ***camouflage*** *with the sandy bottom.* Some stingrays can even change color over the course of several days to adjust to new habitats. Since their mouths are on the side of their bodies, they catch their prey, then crush and eat with their powerful jaws.

Let's assume that in our image pool, we have six images that a stingray is detected, as illustrated in Figure 1. While all the images are relevant at the surface level, they depict different concepts related to the animal "stingray," such as their different body parts or skin patterns (Figure 1a-d). Understanding the concept of the message is very crucial to provide a contextualized image that is more in line with what is explained in the sentence or the paragraph. For example, Figure 1e that displays a harmless type of stingray will be unfitting to explain the possible dangerous attacks. Similarly, to improve the acquisition of the complex word "camouflage", the system should be able to process the context and narrow it down the image selection to the image in Figure 1f. Providing any other image in such context may disrupt the reading fluency due to the conflict that it presents, or it may even yield misunderstandings of the text.

This study provides a semi-automatized dataset creation. First of all, existing established frameworks are used to detect complex words in the text. Second, state-of-the-art multimodal transformers are utilized to find a set of contextualized images for those words. Further, the match between the sentence, complex word, and image triple has been validated first by employing a crowdsourced platform and then by expert analyses as described in the upcoming sections. These semi-automatic detection methods incredibly reduce the costly and time-consuming manual annotation work.

This dataset has many other potential application areas in Language Education, Natural Language Processing, and Computer Vision, e.g., image-text alignment, sense-disambiguation. For example, in psycholinguistic and education, combined with psychological techniques like eye tracking, this corpus provides rich materials for researchers to investigate the mechanism and

[3]retrieved from `https://en.wikipedia.org/wiki/Stingray`



(a) Drawing of a stingray anatomy

(b) Stingray belly side

(c) Spotted stingray

(d) Colorful spotted stingray

(e) A woman plays with stingrays

(f) Stingray lying on a sandy bottom

Figure 1: Supplementary Image Samples for the word "Stingray" (taken from `https://en.wikipedia.org/`

principles of multi-modal learning of human beings. This corpus can also be used in building a more vivid learning context for children and L2 learners in some educational apps or courses.

3. Related Work

3.1. Complex Word Identification

Detecting the complex words (CWI) in the texts is the first step towards providing L2 readers with assistance. CWI has received much attention in the past decade owing to SemEval 2016, 2018, and 2021 Shared Tasks that attract the attention of many NLP researchers to this domain (Paetzold and Specia, 2016a; Yimam et al., 2018; Shardlow et al., 2021). Complex words in texts can be identified by through a wide variety of methods ranging from more traditional dictionary-based approaches to state-of-the-art deep learning techniques. Traditional approaches, which usually require domain knowledge and expert annotation, are still among the most common methods despite their costs. Using NLP approaches to detect complex words in a text helps minimize the manual work and mitigate the cost. Although there are end-to-end machine learning approaches for automatic complex word identification (CWI) (Paetzold and Specia, 2016b; Yimam et al., 2018; Finnimore et al., 2019; Gooding and Kochmar, 2019) their success is still limited given the limited amount of data that have been trained on.

Although pre-trained language models can be used out-of-the-box on CWI tasks, fine-tuning on similar data is still very crucial to achieve better results on a specific task. Our multimodal dataset and semi-automatic data collection tools aim to close this gap. Therefore, our complex word identification will be more in line with the official proficiency standards or frameworks, such as the CEFR framework (Common European Framework of Reference for Languages) (Council of Europe, 2001). According to this framework, the proficiency levels range from A1 to C2. A1-level readers should understand very simple sentences and familiar words, while C1-level readers should comprehend a wide range of demanding, longer texts and recognize implicit meaning. Based on this coarse-grain classification, the L2 texts can be categorized into levels based on their vocabulary, such as (Uchida et al., 2018; Gooding and Kochmar, 2018). The details of the approaches will be elaborated on in the upcoming section.

3.2. Text-Image Retrieval

In the literature, text-image retrieval research has two directions; (i) text to image, i.e. image retrieval based on a textual query (ii) image to text, i.e. text retrieval based on an image query. However, in this work, we focus on textual queries and images as targets where the goal is to find the best matching images according to a sentence and a focus word within this sentence. Having an additional focus word in the query is an extension to common text-image retrieval and is described with more detail in Section 5. Nonetheless, in this work, we heavily rely on standard approaches described in the following text. Current state-of-the-art approaches for text-image retrieval are trained on multi-modal data comprising text-image pairs to compute the similarity between a text and an image. To find the best matching image, the models compute the similarity between the query and all images in the pool of images to be searched. Then the image with the maximum similarity to the query is selected as the best matching image. Current models are based on Transformer (Vaswani et al., 2017) architectures, and their inputs are textual tokens of a sentence and visual tokens of an image or, to be precise, their dense vector embeddings. Textual tokens embeddings are usually computed using pretrained transformer language models like BERT (Devlin et al., 2019). Visual token embeddings are either regions-of-interest embeddings computed by pretrained object detection and classification models like Faster-R-CNN (Ren et al., 2016) or image-patch embeddings computed by a Vision Transformer (Dosovitskiy et al., 2021).

Despite having the same inputs, state-of-the-art models can be subdivided into two groups depending on how and when these two different modality representations are fused: early-fusion and late-fusion models. Early-fusion models like UNITER (Chen et al., 2020) or OSCAR (Li et al., 2020) forward the textual and visual tokens through the same Transfomer-Encoder stacks, where a global text-image similarity score is computed via cross-modal self-attention. Despite their remarkable performance, early-fusion models are not applicable in real-time critical applications with large image pools because computing the similarity between a query and all images requires tremendous computational power. This is different for late-fusion models like TERAN (Messina et al., 2021) or VilBERT (Lu et al., 2019), trained to compute joint representations of texts and images in a common vector space, typically by optimizing contrastive loss functions. To compute the representations, the models forward the input tokens through two separated transformer-stacks – one for the textual and the other for the visual input. Then to compute a global similarity score, the outputs of the two transformer-stacks are fused in a cross-modal manner, individual on the model's implementation. This approach has the significant advantage that the image representations of all images in the pool to be searched can be precomputed so that only the query representation and the fusion of both have to be computed at inference time. In real-time critical applications with a large pool of images, this saves enormous amounts of time and computational power. While former late-fusion models generally perform worse than early-fusion models, the recent late-fusion model CLIP (Radford et al., 2021) achieves state-of-the-art performance. However, CLIP was trained on over 400M text-image pairs, which is significantly more training data than in all other mentioned models. Further, training CLIP requires massive GPU clusters due to the enormous batch sizes necessary during training. Fortunately, CLIP easily fits on a single consumer GPU during inference time. However, to collect the dataset presented by this paper, we cannot use a traditional text-image retrieval approach as is since we extend the query by a contextualized focus word (aka complex word given the language proficiency level), which is part of the query sentence. When retrieving the best matching images, we additionally highlight the region in the image where the focus word is best represented according to the model – see Figure 4 for an example. This requirement originates from the language learner scenario, where we want to provide visual cues for complex words, which are the focus words in our dataset. More details of the context-depended image retrieval are described in Section 5.

4. Dataset Collection

A schematic overview of the data collection pipeline is depicted in Figure 2. More details on single steps are described in the respective sections.

First, the sentences from the Simple Wiki Dataset (c.f. Section 4.1) are tokenized using the NLTK tool[4]. Then, we conduct lemmatization in the pre-processing step to convert the inflected forms of each word. After that,

[4] https://www.nltk.org/

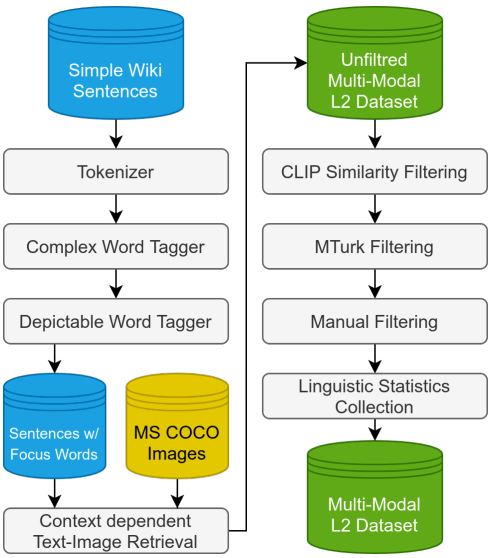


Figure 2: Schematic overview of the dataset collection pipeline.

each token is tagged with respect to its complexity (c.f. Section 4.3) and depictability (c.f. Section 4.4). If a token is both complex and depictable, it is marked as a focus word. The sentences containing less than three focus words are discarded to ensure a level of complexity. The result is a set of samples, where each sample consists of a context sentence and a focus word (a word that will be supported by a contextualized image). Next step is the context-dependent image retrieval. In this stage, the top-5 matching images from MS COCO (Lin et al., 2014) are retrieved, and the focus word region is highlighted in the image with a boundary box (c.f. Section 5).

Since the final dataset should only contain the best image that perfectly matches the context and focus word, additional filtering stages are employed. First image filtering stage has been conducted automatically by using a state-of-the-art multimodal transformer model. A pretrained and publicly available CLIP model (Radford et al., 2021) is used to compute the cosine similarity of each context sentence and the retrieved top-5 images. Samples are discarded, where not all images have a similarity score of at least 0.225. This was inspired by the LAION-400M dataset (Schuhmann et al., 2021) with slightly looser similarity requirements. To further increase the quality of the text image pairs and to make sure that the highlighted image region matches the focus word, we conducted crowdsourcing experiments on Amazon MTurk[5] where human workers are asked to rate how well the respective images match the corresponding focus and context (c.f. Section 6).

After that, in the manual next filtering stage, the authors hand-selected the best matching image from every sample and dropped samples where none of the images represented the context and focus word well enough. With this filtering step, we ensure that only the highest quality samples are included in the final dataset. Since it might be useful in various use case scenarios and downstream research to have linguistic statistics like the number of tokens, POS tags, or named entities in a sample, we collect those using a spaCy[6] powered pipeline and released with the rest of the dataset.

4.1. L2 Learner Reading Material

It is essential to gather appropriate texts for L2 learners that target L2 word acquisition. In order to be included in this dataset, we define several criteria that the text should meet. Firstly, the topic of a text should be open domain (not created based on pre-defined templates), such as the paragraph about stingrays extracted from Wikipedia (Section 2). Unlike widely used multimodal datasets, whose text pieces are merely daily contents, our text pieces cover a wider variety of topics, such as art, culture, geography, nature, science, technology, etc. Furthermore, the text structure should also display complexity. Specifically, the average tokens per paragraph (the text length) should be more than typically used captions. Meanwhile, the comprehension level for a given text should be aligned with the respective reading level for L2 learners. The last criterion is to have *named entities* in the text. Constrained by their annotation method, which first provides an image then asks annotators to write sentences, caption-based datasets exclude *name entities*. However, *named entities* commonly appear in reading materials, such as biography, geography, etc., and thus they are an essential part of reading materials. For this reason, unlike the existing multimodal datasets, our L2 text should involve *name entities*. Motivated by these criteria, we have preprocessed text from five different sources elaborated in the Future Direction section and Appendix A. In this paper, we choose the Wikipedia Simple Corpus as the textual part for the following procedures considering both its scale and related characters that match our hypotheses mentioned above. A thorough comparison among these five datasets is included in the Appendix.

Wikipedia Simple Corpus is the dataset collected from Simple English Wikipedia [7]. There, editors use simple English words and grammar but contain the same entries and content resulting 201.531 articles, which are suitable for children and L2 learners as compared to Normal English Wikipedia [8]. We have utilized the From the raw Wikipedia Simple Corpus (Benzahra and Yvon, 2019), which is a single file containing 505.974 paragraphs where several consistent paragraphs belong to an identical article. After processing the raw data, 59.769 unique articles are kept in this phase. For each article, the average number of paragraphs is *eight*, while for each paragraph, the average number of tokens is *eighteen*. Besides, we compute the average score for

[5] https://www.mturk.com

[6] https://spacy.io/

[7] https://simple.wikipedia.org/

[8] https://www.wikipedia.org/

the Flesch–Kincaid readability tests (Flesch, 2007) to assess the readability grade level of this corpus. In the Flesch reading ease test, higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. The average Flesch–Kincaid readability for Wikipedia Simple Corpus is 64.2, which means the reading materials are suitable for 13-15 US school students in grades 8-9.

4.2. Supplementary Images

The source of the images (Lin et al., 2014) used for this dataset is retrieved from the 2017 version of MS COCO, a popular dataset for various computer vision tasks on natural, non-iconic images. It comprises about 123K carefully selected, annotated, and captioned images from Flickr[9]. We chose this dataset because the images show a wide range of different objects and scenes and are further under the Creative Commons License, which allows non-commercial use and distribution.

4.3. Complex Word Tagger

CWI is developed as a fundamental prerequisite for lexical simplification (LS). However, as described in section 3.1, providing a simplified version is a different task than providing additional information to enhance the acquisition of the complex word. Therefore, the annotated data in the existing datasets created for LS commonly tend to be rare words or long phrases. Due to this inconsistency, in our current study, we prefer to use an established CEFR framework designed for language proficiency to annotate the complex words in our datasets instead of deep learning approaches.

The CEFR framework describes the skills learners should develop at each of the six proficiency levels of the scale. But, it doesn't provide a word list directly with corresponding levels. To overcome this limitation, we obtain a word list with related CEFR level labels by dealing with a word frequencies list released by EFLLex (Dürlich and François, 2018) in the lexical learning domain. The EFLLex (Dürlich and François, 2018) collected word frequencies based on materials designed for English (as L2) learners as a foreign language, which contain $15,282$ words. In this frequency list, a word with different *part-of-speech* tags will be listed separately. For simplification, we combine the frequencies of a word with different POS tags. Then, this list has $9,396$ unique words with frequencies from A1 to C2. To transform the word's frequency to the corresponding CEFR level, we adopt the strategy that the level containing the most significant frequency is the proficiency level for a particular word. B1 is a threshold to distinguish an L2 learner whether the acquisition language can be used freely in daily, study, and work scenarios. Thus, we label a word as a complex word if its CEFR label is above B1 level. In the end, 1690 words are labeled as easy words, while 7706 words are marked as complex, in proportion to 82% of the words in the list.

4.4. Depictability Tagger

Unlike previous work in constructing multi-modal datasets, we also take the depictability of words into account. The word *dog* is more depictable than the word *beautiful*. Learners can comprehend complex words better with visual clues if visually depictable words exist in the context. Besides, we need to conduct a context-dependent text-image retrieval task after hand. Our model can obtain the most relevant images paired for given sentences only if visually depicted elements exist in sentences. To this end, after tagging the complex words, tokens are labeled concerning their depictability by the depictable word tagger (a binary classification; *yes* or *no*).

(Brysbaert et al., 2014), in his psycho-linguistically motivated research, ask native speakers to label 40K words using the Amazon Mechanical Turk platform. In their work, a 5-point rating scale is used to rate a word as abstract or concrete, where 1 means abstract most, whereas 5 means concrete most. The concreteness parameter corresponds to the concept of *depictability* in our research. Finally, 2.3M ratings were collected, and they released the average rating score for each token. Using this rating list for 40K English lemmas[10], we compute the depictability score for each token with the min-max normalization equation below.

$$d_{score} = \frac{r_{token} - r_{min}}{r_{max} - r_{min}}$$

where r_{token} is the average rating point under the 5-point rating scale for a given token, and r_{min} and r_{max} are minimum and maximum rating points in the list respectively. More specifically, Figure 3 shows the depictability score distribution for words after the normalization.

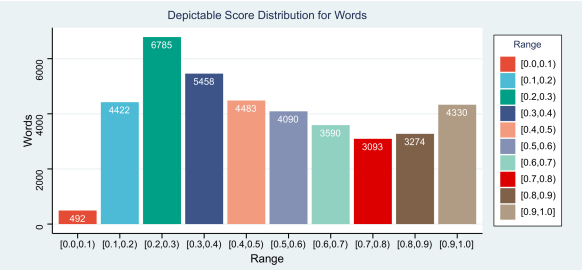


Figure 3: A bar chart of depictable score distribution for 40K English lemmas. We group tokens conditioned on the depictable scores into ten ranges, from $[0.0, 0.1]$ to $[0.9, 1.0]$. This figure shows that around 10.84% of tokens whose depictable scores are over 0.9.

Object and attribute labels of MS COCO dataset obtained from the Faster R-CNN model (Ren et al., 2015)

[9] https://flickr.com

[10] http://crr.ugent.be/archives/1330

Caption: *The track also have a lot of cargo on the train , be one of the big train track to Birmingham Freightliner Terminal (on the site of Birmingham Lawley Street railway station) .*
Focus Word: *cargo*

Figure 4: An example of a context-dependent image retrieval query with the best matching image where the focus word region is highlighted. The query comprises a context sentence (referred to as "Caption" in this figure) and a focus word with the context sentence. Best viewed digitally with zoom and color.

are another source to label a word as depictable, because these labels are mostly noun words whose objects are appeared in images. There are 1625 tokens in the object label list of the MS COCO dataset. In the end, for each word in the sentences, we apply the equation below to get respective depictable labels.

$$label_{depictable} = \begin{cases} 1 & \text{if } d_{score} \geqslant 0.9 \text{ or} \in l_{object} \\ 0 & \text{if } d_{score} < 0.9 \text{ or is } OOV \end{cases}$$

where d_{score} is the depictability score for a given token, l_{object} is the list of object labels, OOV is a token out of these two lists, and 0.9 is the threshold score to decide a word as depictable or not.
By utilizing both complex and depictable word taggers, a word in the textual modality is set as a focus word if its hard label and depictable labels are both positive. At the same time, in the visual modality, a focus word is the object label detected by an object detection model.

5. Context-dependent Image Retrieval

As an extension to the common text-image retrieval task introduced in Section 3.2, where the best matching images for a textual query consisting of a sentence or a word must be found, we introduce context-dependent text-image retrieval in previous work (Schneider, 2021). The difference is that the query is a pair that comprises a sentence, referred to as context, and a focus word contained in the sentence. Further, the goal is to retrieve the best matching images regarding the context with particular attention to the focus word within the context and find the image region where the focus word is represented best (c.f. Figure 4).
To accomplish this goal, we use a pretrained TERAN model for standard text-image retrieval and apply a reranking stage to attend to the focus word specially and to find the region where the focus word is represented best. TERAN is a late-fusion model that computes the global similarity between an image and a textual query – in our case called context – by aggregating a fine-grained word-region-alignment (WRA) matrix $\mathbf{A}$. The cells of $\mathbf{A}$, are the cosine-similarities of the visual regions of the image I and textual tokens of the context sentence C are defined as

$$\mathbf{A}_{i,j} = \frac{\mathbf{v}_i^T \mathbf{t}_j}{|\mathbf{v}_i||\mathbf{t}_j|}$$

where $\mathbf{v}_i \in I$ and $\mathbf{t}_j \in C$.
The global similarity, i.e., the "context-score" $s_{context}$, of an image and a context sentence is defined as

$$s_I^{(c)} = \sum_{j \in |C|} \max_{i \in |I|} \mathbf{A}_{ij}$$

To specially attend to the focus word, we first compute a "focus-score" s_{focus} based on the WRA matrix.

$$s_{focus} = \frac{1}{N * (f_e - f_s + 1)} \sum_{i=0}^{N} \sum_{j=f_s}^{f_e} \mathbf{A}_{ij}$$

where N is the number of regions per image; f_s and f_e are the starting and ending indices of the focus in the context, respectively; and $\mathbf{A}$ is the WRA matrix of an image I and the context C.
After that, we first normalize and then combine the global similarity (interpreted as the "context-score") with the "focus-score" with a weighted average to obtain the image score $s_{combined}$ for the context-dependent text-image retrieval.

$$s_{combined} = \alpha \cdot s'_{context} + (1 - \alpha) \cdot s'_{focus}$$

where $\alpha \in [0, 1]$ is the weight for the weighted average; $s'_{context}$ and s'_{focus} are the normalized "context-score" and the "focus-score", respectively. For the dataset presented in this paper, we set $\alpha = 0.9$.
The image with the highest score is the best matching image according to the context and focus word. To highlight the region where the focus word is represented best, we select the region with the maximum "focus-score".

6. Crowd Source Experiments

Since the context-dependent image retrieval stage does not always retrieve images representing the context and the focus flawlessly, a crowdsourcing experiment was conducted to filter out these samples with the aim of increasing the dataset quality. In this experiment held on Amazon MTurk, workers were given the task to rate how well the context and the focus word are represented in the corresponding image and highlighted image region, respectively, on a 5-star scale. Because the default questionnaires available on MTurk do not efficiently support this task, a tool including a custom

web application was developed. Using the "External Question" of MTurk, access to the web application was provided within the MTurk Marketplace environment. The data for the study comprised 3125 samples, each consisting of a context sentence, a focus word, five images, and two 5-star scales for the context and focus word, respectively, for each of the images. Like in an image slideshow, the workers can switch between the images so that only one image and the corresponding 5-star scales are shown at a time. An example of the application UI as it is presented to the workers is shown in Figure 5.

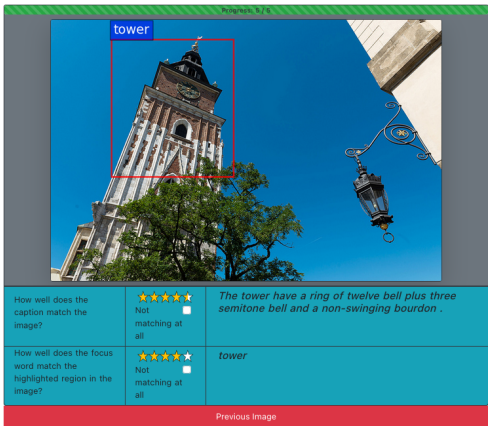


Figure 5: An example of the crowdsourcing experiment UI for MTurk workers. Best viewed digitally with zoom and color.

Using our tool, the samples were published on the MTurk marketplace as HITs. To ensure high-quality results, which are not biased by the opinion of single users, we require three assignments of three different workers per HIT. Further, to accept a HIT, a worker needs at least 1000 approved assignments and an approval rate of 90%. Considering ethical fairness, we set the reward per assignment to 0.2€. With this, an estimated duration of one minute per assignment results in an hourly salary of 12€.
After receiving all assignments for all HITs, the results were filtered as described in the following. First, a score for each image j of the top-5 images per sample i is computed based on the focus rating f_w and the context rating c_w of the three workers

$$\mathbf{score}_i^j = \sum_{w=1}^{3} \max(f_w - 4.0 + c_w - 4.0, 1.0)$$

Then, images are dropped if their score is below or equal to a threshold $T = 2.0$ In other words, an image j of a sample i is kept if at least two workers rated with at least 4.0 stars that the focus and the context is represented well in the respective image. After that, only samples with at least one well-rated image are kept for the manual selection stage.
The authors further filtered down the selected sentence-focus word-image triplets in the last manual stage to ensure the quality of the dataset (*manual expert annotation*). In this stage we discarded 531 samples.

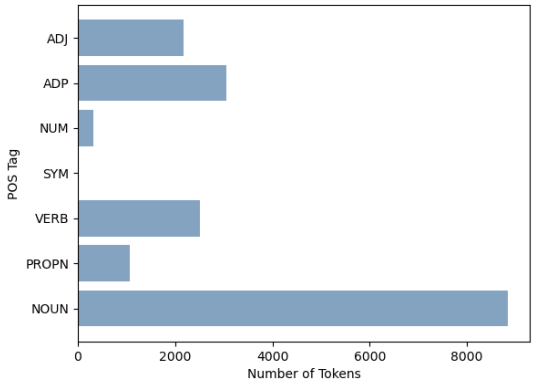


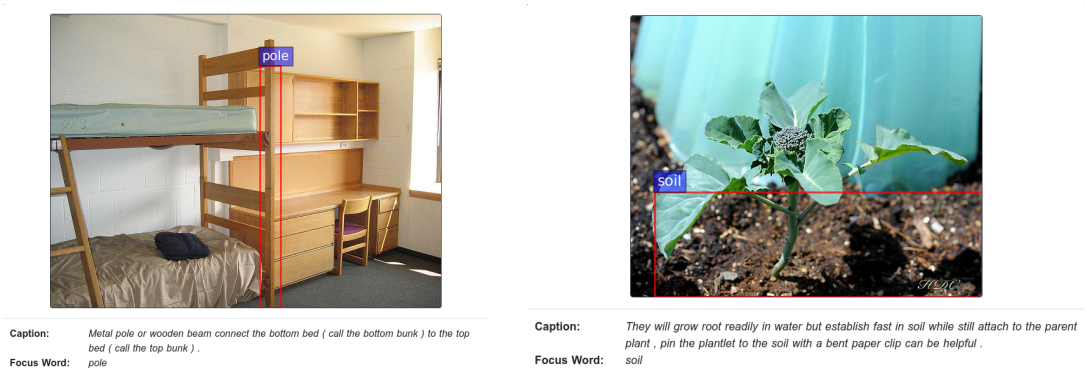
Figure 6: Number of tokens in MOTIF with different POS Tags. NOUN, PRON, VERB, SYM, NUM, ADP, and ADJ mean noun, pronoun, verb, symbol. numeral, adposition, and adjective words respectively.

7. Dataset Structure and Statistics

The original Wikipedia Simple Corpus contains $506K$ sentences. To improve the accuracy of the following retrieval task, we set the threshold of the focus word numbers in each sentence as three. After the complex word tagger and depictable word tagger mentioned in Section 4, $31K$ samples with at least three focus words are kept to be used to conduct the context-dependent image retrieval task discussed in Section 5. At last, after crowdsourcing experiments and manual expert annotation steps, we got 1125 text samples paired with the best matching images for the complex words.
Each sample in the final dataset is contains:

- a sentence referred to as context
- a word within the context referred to as focus word that is both hard and depictable
- an image that globally represents the context and represents the focus word in a highlighted bounding box
- linguistic statistics about the context such as the number of tokens and their respective POS tags

To sum up, there are 1125 samples, where we have 695 unique context sentences and 277 unique focus words. The average tokens per paragraph are 28, while the minimum and maximum tokens per paragraph are 8 and 94, respectively. Meanwhile, we compute the average ratio of tokens associated with name entities vs. all tokens as 3.84%. Besides, the average Flesch–Kincaid readability score is 72.39, interpreted as students' ability in $7th$ Grade in US school. Last, numbers of tokens with different POS tags, as shown in Figure 6. Two visual samples of the final MOTIF dataset are shown in Figure 7. The dataset will be made available on an open-access repository upon acceptance.



(a) (b)

Figure 7: Visual examples included in the MOTIF dataset. Best viewed digitally with zoom and color.

8. Future Directions

There are various ways to extend the dataset in future work. We plan to use additional text-only L2 learner reading material and forward it through the dataset collection pipeline (c.f. Figure 2) to increase the number of samples in the final dataset. Possible resources for this are, e.g., Wikipedia Normal (Benzahra and Yvon, 2019), InScript (Modi et al., 2017), Weebit (Chen and Meurers, 2016), or OneStopEnglish (Vajjala and Lučić, 2018). These datasets vary in size, topics, and length of the sentences but are all specially designed to be understood by 6th to 12th-grade students from US schools.

Further, several components in the dataset collection pipeline can be enhanced to improve the efficiency of the dataset collection and the quality of the final dataset. To begin with, since CEFR word-complexity classification is conducted at the single token level, the use of a simple tokenizer was sufficient. However, this, unfortunately, rips apart multi-word expressions (MWEs) like compound nouns. The complex word tagger can be improved by adapting state-of-the-art complex word identification (CWI) approaches and resources (Kochmar et al., 2020) which pay special attention to multi-word expressions (MWE). Moreover, this dataset can be used to fine-tune SOTA CWI models to improve automatic complex word detection.

Further, the employed object and attribute vocabulary, which comprises about 1600 different terms, can be significantly extended by the vocabulary of the Visual Genome dataset, which contains about 75K unique object types and about 40K attribute types. By improving the pipeline as described, the quality of the output of the context-dependent image-retrieval stage will automatically increase. However, the stage itself can be further improved by various methods briefly summarized in the following.

The currently employed context-dependent image-retrieval model is a TERAN model, where the visual inputs are region-of-interest (ROI) feature vectors computed by a pre-trained Faster-R-CNN model. The advantage of this approach is that we can compute the focus score (c.f. Equation 5) of a focus word and an image (region) from the WRA matrix, which holds fine-grained cosine similarities between words and image regions. However, the bounding box of the image region often does not perfectly fit the underlying object representing the focus word.

To resolve these issues, we plan to leverage a pre-trained CLIP model in the context-dependent image-retrieval stage. We will utilize class activation mapping techniques introduced (Zhou et al., 2016; Selvaraju et al., 2017) to compute the focus score and more accurate bounding boxes.

9. Conclusion

In this study, we present a semi-automatized pipeline to create a high-quality multimodal dataset containing text pieces for L2 speakers, annotated complex words, and contextualized images that ease the comprehension of the complex word given the context. Our pipeline starts with selecting L2 text, conducting text analysis (number of *named entities*, readability scores etc). It further detects complex words using well-established CEFR levels and employs SOTA NLP approaches for finding contextualized images. However, these automated processes are followed by a careful validation method using the Amazon MTURK crowdsourcing platform and expert analysis. The resulting dataset consists of 1125 text samples annotated with complex words and context-dependent images for these words.

This multimodal support approach and the dataset are not only for the L2 domain, but they can also be used in developing assistive systems for people with low literacy and reading difficulties. Further, these enriched annotations can be instrumental in fine-tuning or testing automatic CWI and contextualized image-retrieval models.

10. Bibliographical References

Albahiri, M. H. and Alhaj, A. A. M. (2020). Role of visual element in spoken English discourse: impli-

cations for YouTube technology in EFL classrooms. *The Electronic Library*, 38(3):531–544.

Benzahra, M. and Yvon, F. (2019). Measuring text readability with machine comprehension: a pilot study. In *Workshop on Building Educational Applications Using NLP*, pages 412–422.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

Chen, X. and Meurers, D. (2016). Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision (ECCV)*, pages 104–120, Online.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge. U.K: Press Syndicate of the University of Cambridge.

Dalton, B. and Grisham, D. L. (2011). eVoc Strategies: 10 Ways to Use Technology to Build Vocabulary. *Reading Teacher*, 64(5):306–317.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, MN, USA.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dürlich, L. and François, T. (2018). EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ecalle, J., Magnan, A., Bouchafa, H., and Gombert, J. E. (2009). Computer-Based Training with Ortho-Phonological Units in Dyslexic Children: New Investigations. *Dyslexia*, 15(3):218–238.

Finnimore, P., Fritzsch, E., King, D., Sneyd, A., Rehman, A. U., Alva-Manchego, F., and Vlachos, A. (2019). Strong baselines for complex word identification across multiple languages. *arXiv preprint arXiv:1904.05953*.

Flesch, R. (2007). Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Gerbier, E., Bailly, G., and Bosse, M. L. (2018). Audio–visual synchronization in reading while listening to texts: Effects on visual behavior and verbal learning. *Computer Speech & Language*, 47:74–92.

Gooding, S. and Kochmar, E. (2018). Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana, USA.

Gooding, S. and Kochmar, E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153.

Hahn, N., Foxe, J. J., and Molholm, S. (2014). Impairments of multisensory integration and cross-sensory learning as pathways to dyslexia. *Neuroscience & Biobehavioral Reviews*, 47:384–392.

Kochmar, E., Gooding, S., and Shardlow, M. (2020). Detecting Multiword Expression Type Helps Lexical Complexity Assessment. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4426–4435, Marseille, France, May. European Language Resources Association.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). Oscar: Object-Semantics Aligned Pre-training for Vision-and-Language Tasks. In *European Conference on Computer Vision (ECCV)*, pages 121–137, Online.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, Switzerland.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 32, pages 13–23, Vancouver, Canada.

Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., and Marchand-Maillet, S. (2021). Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23.

Modi, A., Anikina, T., Ostermann, S., and Pinkal, M. (2017). Inscript: Narrative texts annotated with script information. *arXiv preprint arXiv:1703.05260*.

Paetzold, G. and Specia, L. (2016a). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California.

Paetzold, G. and Specia, L. (2016b). Unsupervised lexical simplification for non-native speakers. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, pages 3761–3767, Phoenix, Arizona USA.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149.

Rets, I. and Rogaten, J. (2021). To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717.

Schneider, F. (2021). Self-Supervised Multi-Modal Text-Image Retrieval Methods to Improve Human Reading. Master's thesis, University of Hamburg.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021). LAION-400M: Open Dataset of CLIP-filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Venice, Italy.

Shardlow, M., Evans, R., Paetzold, G. H., and Zampieri, M. (2021). Semeval-2021 Task 1: Lexical complexity prediction. In *In Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2021*, pages 1–16, Bangkok, Thailand (online).

Uchida, S., Takada, S., and Arase, Y. (2018). Cefr-based lexical simplification dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Vajjala, S. and Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems (NIPS)*, 30:5998–6008.

Xie, H., Mayer, R. E., Wang, F., and Zhou, Z. (2019). Coordinating Visual and Auditory Cueing in Multimedia Learning. *Journal of Educational Psychology*, 111(2):235–255.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

11. Language Resource References