Contextual Definition Generation

Jeffrey T. Yarbro, Andrew M. Olney

[1] University of Memphis, Memphis, TN 38152, USA
{jyarbro2, aolney}@memphis.edu

Abstract. This paper explores the concept of dynamically generating definitions using a deep-learning model. We do this by creating a dataset that contains definition entries and contexts associated with each definition. We then fine-tune a GPT-2 based model on the dataset to allow the model to generate contextual definitions. We evaluate our model with human raters by generating definitions using two context types: short-form (the word used in a sentence) and long-form (the word used in a sentence along with the prior and following sentences). Results indicate that the model performed significantly better when generating definitions using short-form contexts. Additionally, we evaluate our model against human-generated definitions. The results show promise for the model, showing that the model was able to match human-level fluency. However, while it was able to reach human-level accuracy in some instances, it failed in others.

Keywords: GPT-2, contextual definitions, human evaluation, definition generation

1 Introduction

Prior studies have suggested that 95% of words in a text must be within a reader's vocabulary for adequate reading comprehension to occur [10, 15]. This presents a problem for academic texts, which often use low-frequency words to describe key concepts related to their field. For readers to comprehend these concepts, they must first acquire the field-specific vocabulary. One of the classical ways to help with this problem is to have a glossary near the end of the textbook or a list of key terms at the beginning of the chapter, allowing the reader to have a relatively easy way to familiarize themselves with words that they are unfamiliar with and reference these definitions during reading. This classical approach in paper-based textbooks has a few notable issues: (1) the reader is forced to change pages and find the term that they are looking for (2) the number of definitions is limited to a small subset of words within the book (3) the definition might not be appropriate to the context, as happens when a word has multiple definitions.

Intelligent systems have made the process of finding a definition much easier. Rather than turning pages or firin[1]g up a search engine, readers can now click on a piece of text, and a pop-up will appear displaying the definition for the selected term [5, 7]. These definitions are typically acquired by querying a definition database. While this is a good solution, it does come with some issues. (1) The source of definitions must contain the definition for the word in the context displayed. (2) If the word has multiple definitions, one must either display a list of definitions or deploy a word sense disambiguation model to find the most fitting definition. (3) Definitions are not tuned to the precise context of the word.

This paper explores the concept of using a deep-learning model to generate dynamic, contextual definitions by paying attention to the surrounding context of the word. We do this by creating a new dataset consisting of words, a definition for each word, and a list of contexts associated with each definition. We then fine-tune a GPT-2 based model on this dataset, resulting in a model capable of autoregressively generating a definition for any English word with only the word and a context as inputs. We assess the model using human evaluation with three research questions in mind: (1) What type of context provides the best initialization for the model (2) How does the model perform relative to human-generated definitions. (3) Is the model biased towards any particular subject?

We answer these questions by analyzing how human raters rate machine-generated and human-generated definitions for terms from 5 college-level textbooks in terms of accuracy and grammatical fluency.

2 Data Collection and Training

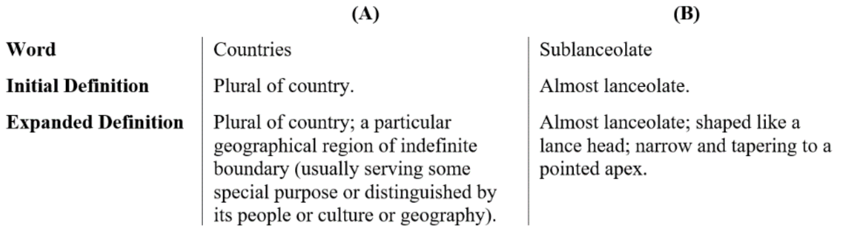
2.1 Data Collection

A key constraint before beginning data collection was to find sources where word entries had both a definition and a context that matched that definition. A matching definition and contextual pair are important since sampling random sentences/paragraphs the term is within and attempting to find the most appropriate definition could lead to errors. We did not impose strict types of contexts. Contexts could be words in example sentences, words in example paragraphs, hypernyms, synonyms, hyponyms, etc. The primary goal was to give the model enough of a contextual clue to internally disambiguate the sense of the word and generate a definition that fit that sense.

With this in mind, we collected training data from the following sources: (1) Wiktionary: Extracted definition/context data by cleaning the March 2021 version of the XML dump file [23]. Contexts contained synonyms, example sentences/paragraphs, hypernyms, hyponyms, and sense tags. (2) Lexico: Scraped Lexico for definitions, example sentences, synonyms, and sense tags [11] (3) WordNet: Used the NLTK implementation of WordNet to acquire gloss entries and sense information [8, 13] (4) Wikipedia: Used to expand definitions as discussed in section 2.2.

Data from each source was then combined into a JSON file that totaled approximately 300MB when compressed. All entries from each source were kept should they contain enough information about the word, as further discussed in section 2.2. This includes word entries that had the same (or paraphrased slightly) sense definition in multiple sources. No merging was performed to ensure that all possible senses and contexts were represented in the dataset. However, it may also deter the model discussed in section 2.3 from memorizing definitions due to slight definition variations.

2.2 Modifications to data



	(A)	(B)
Word	Countries	Sublanceolate
Initial Definition	Plural of country.	Almost lanceolate.
Expanded Definition	Plural of country; a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography).	Almost lanceolate; shaped like a lance head; narrow and tapering to a pointed apex.

Fig. 1. Examples of expansion for the word "country" and "sublanceolate."

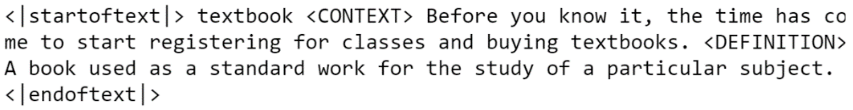
Many definitions within the training dataset did not contain enough information about the word it was describing to be helpful to readers unfamiliar with the word. An example of this can be seen in Figure 1A above. Here the definition of the word countries is "plural of country." While this definition is not wrong, it does not contain enough information for someone who does not have prior knowledge of the root word "country." We attempt to fix this issue by identifying the referenced word and appending the referenced word's definition to the end of the original definition.

The referenced word was found using regular expressions, word frequency, and parts-of-speech tags. Regular expressions were used to find key phrases (e.g., plural of) that regularly pointed to the referenced word (e.g., the original word with grammatical suffixes removed). We additionally found that many definitions contained less than four words and heavily relied upon a key noun/adjective to convey the word's meaning. This was appropriate from our perspective when the referenced word was frequent and well-known to most English speakers. However, there were also cases where the referenced word, such as displayed in Figure 1B, was infrequent and perhaps not well-known enough to strengthen the reader's mental representation of the original word. In cases such as these, we also attempted to expand the definition to something more useful.

To expand the definitions, we identify the referenced word and search the training dataset for all entries of that reference word. We then use SentenceTransformer (RoBERTa-large variant) to embed each found definition into sentence vectors, along with the base term's contexts [16, 18]. We then compute cosine similarity to compare each definition entry for the referenced word with the base term's contexts. The definition with the highest cosine similarity score was then appended to the end of the definition, as displayed in Figure 1. If the referenced word was not within the dataset, we query Wikipedia using wptools API [24]. If the word is found, we append the first sentence to the end of the definition. If not, the word and definition were removed from the dataset.

After removing terms with low information, the total dataset contained 254k unique terms, 512k definitions, and 2.66M contexts.

2.3 Model and Training



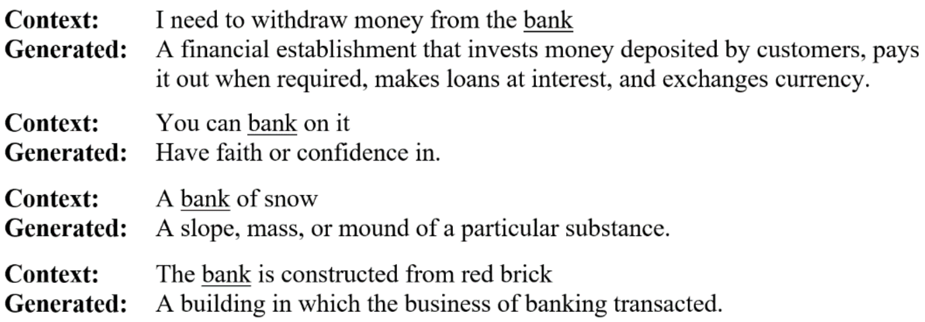
```
<|startoftext|> textbook <CONTEXT> Before you know it, the time has co
me to start registering for classes and buying textbooks. <DEFINITION>
A book used as a standard work for the study of a particular subject.
<|endoftext|>
```

Fig. 2. Example of the formatting used during training for the word "textbook"

Before training, we begin by making the data more machine-readable by placing each word, definition, and context from the dataset into the format shown in Figure 2 for each context entry. This format contains two special tokens: <CONTEXT> and <DEFINITION>. These tokens were used to make it easy for the model to determine where the context ended and where definition generation should begin.

We then trained a fine-tuned version of GPT-2 called WikiMorph on the dataset [9, 22]. WikiMorph uses the large variant of GPT-2 made available by Hugging Face to break down words into morphological compounds and definitions associated with each compound [20]. It was chosen for this work because it contains some ability to generate definitions for words and sub-words. Therefore, likely giving the model a good initialization point. We then fine-tune this model on the dataset for 1 epoch. We stopped at only a single epoch because there were no improvements in validation loss or ROUGE scores on the validation data [17].

2.4 Model Usage and Examples



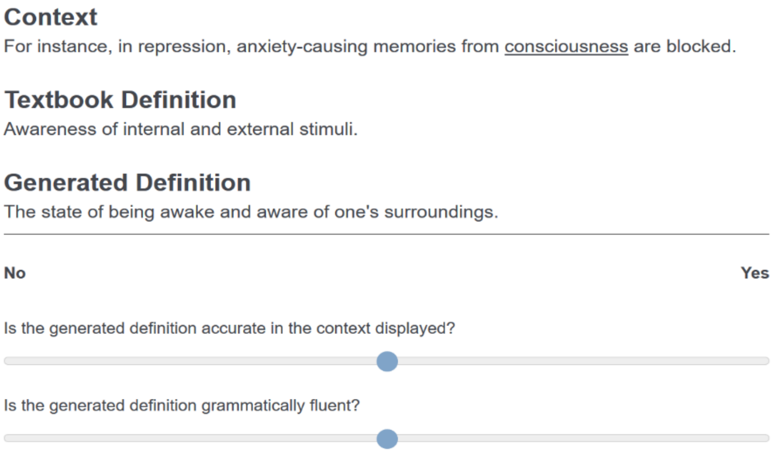
Context:	I need to withdraw money from the bank
Generated:	A financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.
Context:	You can bank on it
Generated:	Have faith or confidence in.
Context:	A bank of snow
Generated:	A slope, mass, or mound of a particular substance.
Context:	The bank is constructed from red brick
Generated:	A building in which the business of banking transacted.

Fig. 3. Sample of generated definitions for the word "bank" used in different contexts.

The model is used by placing the word and a context into the format displayed in Figure 2 and omitting everything following the special <DEFINITION> token. We then feed this text into the model, which uses its tokenizer to encode the text into tensors. The model then begins to autoregressively generate a definition by referencing the given word, the given context, and all prior output tokens until it reaches the end token designated as "<|endoftext|>". We then use regular expressions to find all text between the <DEFINITION> and <|endoftext|> token to display to the user.

3 Evaluation

3.1 Evaluation Setup



Context

For instance, in repression, anxiety-causing memories from consciousness are blocked.

Textbook Definition

Awareness of internal and external stimuli.

Generated Definition

The state of being awake and aware of one's surroundings.

Is the generated definition accurate in the context displayed?

Is the generated definition grammatically fluent?

Fig. 4. Example of survey question for the word "consciousness"

To evaluate the model, we began by collecting sample texts and definitions from 5 different university-level textbooks. These were from the following subjects: Anatomy and Physiology [6], American Government [1], Astronomy [2], Chemistry [4], and Psychology [14]. We then searched these textbooks for paragraphs containing each of the key terms within its glossary. These paragraphs were then divided into sentences using Spacy's sentence segmentation model [19]. We then randomly chose 50 terms from each textbook and generated definitions using two different context types. (1) Short context: The sentence containing the term. (2) Long Context: The sentence containing the word and the sentence before and the sentence after if available. We then fed the model each of these context types to create two different Qualtrics surveys.

Each survey began by displaying a multiple-choice question asking participants to choose which of the above subjects they felt most knowledgeable in. Their selection would determine what set of generated definitions they would assess throughout the rest of the survey. (e.g., if they chose psychology, they would evaluate the 50 generated definitions from the psychology textbook). For each generated definition, we asked two questions: (1) Is the generated definition accurate in the context displayed? (2) Is the generated definition grammatically fluent? Both questions used sliders that recorded results on a 0-100 scale, with the starting point set at 50 for each. To see an example question from the survey, please refer to Figure 4 above.

We additionally created a third survey containing human-generated definitions to compare our model against. Human-generated definitions were collected from the training dataset described in Section 2. We searched the training dataset for all definition entries for each term used in the prior two surveys. We then selected the most appropriate definition using SBERT sentence vectors. Each definition for the term was embedded into a sentence vector. We then computed cosine similarity between each of these definitions from the training dataset with the definition from the textbook. The definition from the training dataset with the highest similarity value was displayed to the user in the same format shown in Figure 4. If a term happened to not be within the training dataset, it was dropped from the survey entirely. This lowered the number of questions for the human-generated definitions from 50 to 40-45, depending on the subject.

For each survey, we implemented 3-4 control questions. These control questions replaced the generated definition displayed in Figure 4 with a random definition for a different term from the subject's textbook. Participants were required to assign an accuracy value below 50 on the slider question for over half of the control questions to be included in the results. (i.e., to be considered reliable raters). Below 50 was the threshold selected due to 50 being the starting position for the slider. (i.e., the participants had to actively move the slider towards "no" and indicate inaccuracy).

3.2 Participants

Table 1. Inter-rater reliability for each survey with included total numbers.

		American Government		Anatomy & Physiology		Astronomy		Psychology	
		n	α	n	α	n	α	n	α
Short	Accuracy	18	0.891	4	0.710	3	0.930	14	0.795
	Fluency		0.896		0.925		0.735		0.820
Long	Accuracy	16	0.871	8	0.890	5	0.945	10	0.944
	Fluency		0.944		0.961		0.957		0.940
Real	Accuracy	16	0.872	5	0.923	4	0.986	14	0.874
	Fluency		0.954		0.961		0.982		0.934

Participants were sourced using CloudResearch in May 2021 [21]. CloudResearch is a platform that screens the Mechanical Turk population for higher quality participants. The only imposed requirement was that participants were required to be from English-speaking countries. This led to a total of 194 participants. Of which, 53 were excluded from the results for two reasons: (1) They failed to pass control checks (2) They selected Chemistry as their preferred subject. Chemistry was removed from the results due to having too low number of participants to calculate inter-rater reliability. Final numbers for each group, along with inter-rater reliability, are shown in Table 1 above. Reliability was generally high, with Cronbach's alpha mostly in the .80-.95 range.

4 Results and Discussion

Table 2. Mean ratings for each subject for both accuracy and fluency scores.

		American Government	**Anatomy & Physiology**	**Astronomy**	**Psychology**	**All**
Short	**Accuracy Mean**	57.99	74.82	52.28	67.40	62.60
	Fluency Mean	80.89	84.91	89.99	84.60	82.82
Long	**Accuracy Mean**	47.87	60.44	48.81	59.16	53.70
	Fluency Mean	80.82	83.36	79.98	88.03	83.20

Real	**Accuracy Mean**	79.54	74.91	72.65	80.52	78.60
	Fluency Mean	84.39	75.37	82.89	85.42	83.42

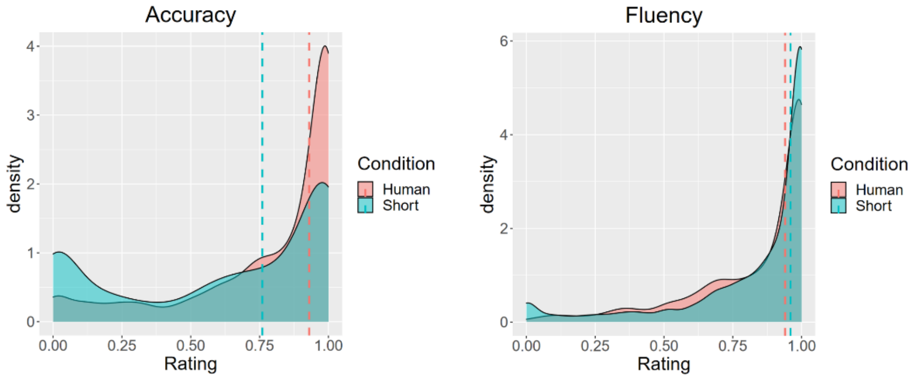
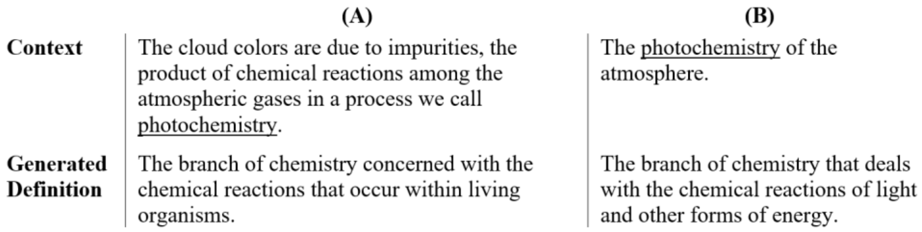


Fig. 5. Density plot comparing human-generated definitions with short-context generated definitions for both accuracy and fluency ratings. The dashed line indicates medians for each.

To answer our research questions, we ran mixed-effects beta regressions for accuracy and fluency using all three survey conditions and random intercepts for definitions and participants. Each model was initially fit with the interaction between context and topic, but if the interaction was not significant, it was dropped, and the model refit. For accuracy, we ran an ANOVA and found a significant main effect of context, $\chi2(2) = 42.70$, $p < .001$. We probed this main effect to answer our first research question, which is what context type gave the model the best initialization for definition generation. Post hoc comparison using Tukey's HSD revealed that definitions generated from shorter contexts (M = 62.60, SE = 0.875) performed significantly better than those generated from longer contexts ($M = 53.70, SE = 0.886$), t(5316)=2.38, $p = 0.045$. We suspect two possible reasons as to why shorter contexts performed better in terms of accuracy: (1) The training data had far more entries with sentence contexts (56.84% of the total contexts) relative to longer contexts with two or more sentences (1.66% of the total contexts), indicating the model may need more examples of longer contexts to learn what to filter out and what to pay attention to. (2) While there could be instances where longer contexts provide additional information, all the evaluation data came from college-level textbooks, which may contain enough information-rich keywords within a single sentence for the model to determine the correct sense and generate an appropriate definition.

To answer our second research question to see how the model performs relative to human-generated definitions, we conducted an additional post hoc comparison using Tukey’s HSD, which revealed that human-generated definitions ($M = 78.57$, $SE = 0.70$) performed significantly better than definitions generated from shorter contexts ($M = 62.60$, $SE = 0.875$), $t(5316) = 4.11$, $p < .001$, as well as those generated from longer contexts ($M = 53.70, SE = 0.886$), $t(5316)=2.38$, $p < .001$. Further examination of the accuracy density plot seen in Figure 5 sheds some light on this result. It shows that



	(A)	(B)
Context	The cloud colors are due to impurities, the product of chemical reactions among the atmospheric gases in a process we call photochemistry.	The photochemistry of the atmosphere.
Generated Definition	The branch of chemistry concerned with the chemical reactions that occur within living organisms.	The branch of chemistry that deals with the chemical reactions of light and other forms of energy.

Fig. 6. Examples of two generated definitions for a word outside the training dataset called "photochemistry." (A) Displays an error case in which the model generated an incorrect definition. (B) Displays a more accurate definition.

the model was able to perform admirably in many situations, with a median rating equal to 75. However, it also had an abundance of error cases where ratings were below 25. The exact reasons for these error cases require further investigation. Two possible culprits include: (1) The model had difficulty reading the context. As demonstrated by Figure 6 and the fact that the model performed significantly better on short contexts, it is safe to conclude that the model is sensitive to contexts. (2) A poor representation of the input word due to it not being in the training data or having too many conflating definitions without enough contextual examples to properly learn all senses.

To answer our third research question to see if the model performed better on some topics relative to others, we did not find significant differences between textbook subjects. Though, a non-significant trend suggests that, while human-generated definitions did equally well across topics, the model might perform better on some topics relative to others. In particular, the average meaning ratings for definitions generated from shorter contexts were approximately equal to those for human-generated definitions on the topic of Anatomy and Physiology. This trend is potentially worthy of further analysis in subsequent studies examining a wider variety of textbook sources and topics.

We performed an identical ANOVA analysis for fluency and found no significant main effects of context, topic, or interaction. As shown in the fluency density plot in Figure 5, the model's performance was excellent for fluency relative to human-generated definitions. Some of this could be due to the data found within our dataset being slightly less fluent than a typical human definition due to some entries coming from dirty sources or the definition expansion method discussed in Section 2.2. However, even with this considered, the model appeared to perform exceptionally well with mean values effectively greater than or equal to 80 for both short and long-form context surveys.

5 Conclusion

This work presents a deep-learning model capable of dynamically generating definitions based solely on the surrounding context. We examined the model's ability to generate definitions using two context types: short and long-form. Short-form contexts significantly outperformed long-form contexts in human-rated accuracy but fell short of human-generated definitions on this metric. In contrast, short-form and long-form conditions were indistinguishable from human-generated definitions in terms of fluency, displaying some promise for the model.

Acknowledgment

This material is based upon work supported by the National Science Foundation (1918751, 1934745) and the Institute of Education Sciences (R305A190448).

References

1. Krutz G, Waskiewicz S.: American Government. OpenStax (2018). https://doi.org/10.4324/9781351239226.
2. Andrew Franknoi, David Morrison, S.C.W.: Astronomy. OpenStax (2016).
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. 2017-Decem, 5999–6009 (2017).
4. Flowers Paul, Theopold Klaus, Langley Richard, Robinson William: Chemistry 2002. (2002).
5. MeetDeveloper: Dictionary-Anywhere, https://github.com/meetDeveloper/Dictionary-Anywhere, (2020).
6. David Shier, Jackie Butler, R.L.: Hole's Human Anatomy & Physiology. McGraw-Hill Education (2019).
7. Chaudhri, V.K., Cheng, B.H., Overholtzer, A., Roschelle, J., Spaulding, A., Clark, P., Greaves, M., Gunning, D.: Inquire biology: A textbook that answers questions. AI Mag. 34, 55–72 (2013). https://doi.org/10.1609/aimag.v34i3.2486.
8. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: An On-line Lexical Database.
9. Alec, R., Jeffrey, W., Rewon, C., David, L., Dario, A., Ilya, S.: Language Models are Unsupervised Multitask Learners. (2018).
10. Laufer Batia: Lexical Thresholds for Reading Comprehension: What They Are and How They Can Be Used for Teaching Purposes. 47, 867–872 (2018). https://doi.org/10.1 002/tesq. 1 40.
11. Lexico. (2021).
12. Robyn Speer, And, J.C., And, A.L., And, S.J., Nathan, L.: LuminosoInsight/wordfreq, https://github.com/LuminosoInsight/wordfreq, (2018). https://doi.org/10.5281/zenodo.1443582.
13. Bird, S., Loper, E.: NLTK: The Natural Language Toolkit.
14. Spielman Rose, Jenkins William, Lovett Marilyn: Psychology-2e - Tâm lý học. OpenStax (2013).
15. Treptow, M.A., Burns, M.K., McComas, J.J.: Reading at the frustration, instructional, and independent levels: The effects on students' reading comprehension and time on task. School Psych. Rev. 36, 159–166 (2007). https://doi.org/10.1080/02796015.2007.12087958.
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv. (2019).
17. Ganesan, K.: ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks. arXiv. 1–8 (2018).

18. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf. 3982–3992 (2020). https://doi.org/10.18653/v1/d19-1410.
19. Honnibal, Matthew and Montani, Ines and Van Landeghem, Sofie and Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python, https://spacy.io/, (2021). https://doi.org/10.5281/zenodo.1212303.
20. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Transformers: State-of-the-art natural language processing. arXiv. (2019). https://doi.org/10.18653/v1/2020.emnlp-demos.6.
21. Litman, L., Robinson, J., Abberbock, T.: TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behav. Res. Methods. 49, 433–442 (2017). https://doi.org/10.3758/s13428-016-0727-z.
22. Yarbro, J.T., Olney, A.M.: WikiMorph: Learning to Decompose Words into Morphological Structures. 1–6 (2021).
23. WikiMedia: Wiktionary, https://en.wiktionary.org.
24. siznax: wptools, https://github.com/siznax/wptools.