Grokking Wug Vectors

Jacob A. Matthews and Marten van Schijndel (Cornell University)

We investigate the relationship between two commonplace psycholinguistic tools: *pseudowords* (or *nonce words*) and *word vectors*. Pseudowords have long been used to construct stimuli targeting a wide range of psycholinguistic phenomena, from morphology in language acquisition [1] to lexical decision tasks more generally [2, *inter alia*]. Likewise, computational psycholinguists often model lexical processing using model-generated word vectors and cosine distance as a continuous value of semantic dissimilarity, as in [3]. We study the appropriateness of using pre-trained word vectors with pseudoword (nonce) paradigms using three kinds of vectors, both context-free (FastText [4]) and contextual (BERT [5] and GPT-2 [6]).

Vocabulary selection We studied a set of 25 canonical English nonce "nouns" [1]. To compare these to vocabulary typical of language acquisition or lexical decision stimuli, we used a standard list of ~3,000 simple, common English words (henceforth called VOCAB) [7].

Representations of canonical pseudowords in semantically-reduced contexts We place each word within a semantically reduced context phrase during generation as described in [Fig. 1A]. We then use a representational analysis tool, Minicons [8], to extract the relevant contextualized word vector from the context sequence. We use standardized cosine distance (SCD) as suggested in [9] to measure vector dissimilarity across word pairs. **Results** We report examples of the nearest lexical neighbor in VOCAB to each pseudoword and summary statistics of pairwise similarities [Fig. 2, 3]. We do not observe significantly different SCDs between pseudo- and lexical words across models [Fig. 2], though orthography appears to inconsistently drive similarity across models [Fig 3].

Representations of generated pseudowords in authentic contexts To further investigate the effect of context and orthography across models, we use a pseudoword creation tool, Wuggy [10], to generate 22,113 phonotactically-legal pseudowords from our VOCAB words. We use sample sentences from the Brown corpus to generate contextual vectors for each word, replacing each seed word with an associated nonce and processing the result with BERT and GPT-2 [Fig. 1B]. We then compare SCDs between each input word and its corresponding generated pseudowords, as well as average orthographic Levenshtein distance of the 20 nearest lexical neighbors to each generated word (ORTH). **Results** We report correlations between SCD and ORTH [Fig. 4]. As a baseline, we also include SCD values using representations from semantically-reduced ('bleached') contexts. We find that ORTH is very strongly correlated with our non-contextual model distances, but not to contextual models, regardless of the context's semantic value.

Discussion Using vectors of pseudowords introduces possible confounds in psycholinguistic work, and researchers should be aware of the biases introduced by using different kinds of nonce vectors. Contextual pseudoword vectors taken from authentic contexts are nearly identical to real word vectors taken from the same context, as context "bleeds" into individual word vectors. On the other hand, representational differences between non-contextual word vectors (e.g., FastText) are mostly explained by orthographic differences between real words and their derived pseudowords. To minimize the orthographic confounds between nonce and real words, we found that contextualized models with semantically reduced contexts produce the most relevant vector representations, even though contextual models do not exhibit consistent behavior across architectures and context conditions.

Figure 1: Contextual word vector generation

A. *Semantically-reduced context*

Model('{word} is a word') $\longrightarrow [v_{\{word\}}, v_{is}, v_a, v_{word}]$

$\text{Minicons}_{\text{Model}}([v_{\{word\}}, v_{is}, v_a, v_{word}]) \longrightarrow v_{\{word\}}$

B. *Authentic context*

Model('$\{c_0\}$ … {word} … $\{c_n\}$') $\longrightarrow [v_{\{c\}}, \ldots, v_{\{word\}}, \ldots, v_{\{c\}}]$

$\text{Minicons}_{\text{Model}}([v_{\{c\}}, \ldots, v_{\{word\}}, \ldots, v_{\{c\}}]) \longrightarrow v_{\{word\}}$

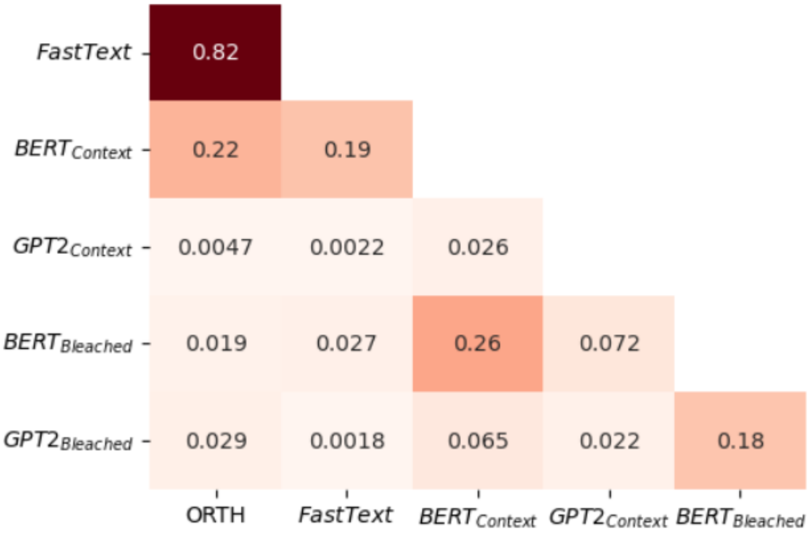
Figure 2: Mean (σ) SCD across models (lower is closer)

Word pairing	FastText	BERT	GPT-2	N = |A||B|
SCD(pseudo, lexical)	0.9988 (0.0658)	0.9999 (0.0741)	1.0003 (0.0644)	37,704
SCD(lexical, lexical)	0.9949 (0.0951)	0.9996 (0.1059)	0.9845 (0.1598)	4.93M
SCD(pseudo, pseudo)	0.9911 (0.2554)	0.9989 (0.2364)	0.9982 (0.2341)	300

Figure 3: Examples of nearest VOCAB word to pseudoword (SCD)

	FastText	BERT	GPT-2
wug	demonstration (0.7417)	sing (0.7633)	wicked (0.7913)
heaf	any (0.7612)	surely (0.7454)	deaf (0.4553)
glack	lack (0.7012)	bake (0.7624)	lack (0.3456)
stup	stupid (0.5821)	commission (0.7778)	pupil (0.7115)
plad	shirt (0.5977)	previous (0.6768)	tradition (0.4171)

Figure 4: abs(Spearman's ρ) of ORTH and SCD(Seed, Pseudoword) across models



References: [1] Berko 1958 [2] Ratcliff et al. 2014 [3] Mikolov et al. 2013 [4] Bojanowski et al. 2016 [5] Devlin et al. 2019 [6] Radford et al. 2018 [7] Modified Longman 3000 [8] Misra 2022 [9] Timkey and van Schijndel 2021 [10] Keuleers and Brysbaert 2010