de Varda, Andrea Gregor, Daniele Gatti, Marco Marelli, and Fritz Günther. 2024. Meaning Beyond Lexicality Capturing Pseudoword Definitions with Language Models. *Computational Linguistics*, uncorrected proofs.

Meaning Beyond Lexicality Capturing Pseudoword Definitions with Language Models

Andrea Gregor de Varda*
University of Milano-Bicocca
Department of Psychology
a.devarda@campus.unimib.it

Daniele Gatti
University of Pavia
Department of Brain and
Behavioral Sciences
daniele.gatti@unipv.it

Marco Marelli
University of Milano-Bicocca
Department of Psychology
marco.marelli@unimib.it

Fritz Günther
Humboldt-Universität zu Berlin
Department of Psychology
fritz.guenther@hu-berlin.de

Pseudowords such as "knackets" or "spechy"—letter strings that are consistent with the orthotactical rules of a language but do not appear in its lexicon—are traditionally considered to be meaningless, and used as such in empirical studies. However, recent studies that show specific semantic patterns associated with these words as well as semantic effects on human pseudoword processing have cast doubt on this view. While these studies suggest that pseudowords have meanings, they provide only extremely limited insight as to whether humans are able to ascribe explicit and declarative semantic content to unfamiliar word forms. In the present study, we utilized an exploratory-confirmatory study design to examine this question. In a first exploratory study, we started from a pre-existing dataset of words and pseudowords alongside human-generated definitions for these items. Using 18 different language models, we showed that the

* Corresponding author.

Action Editors: Marianna Apidianaki, Abdellah Fourtassi, and Sebastian Padó. Submission received: 16 December 2023; revised version received: 30 April 2024; accepted for publication: 19 June 2024.

https://doi.org/10.1162/coli_a_00527

definitions actually produced for (pseudo)words were closer to their respective (pseudo)words than the definitions for the other items. Based on these initial results, we conducted a second, pre-registered, high-powered confirmatory study collecting a new, controlled set of (pseudo)word interpretations. This second study confirmed the results of the first one. Taken together, these findings support the idea that meaning construction is supported by a flexible form-to-meaning mapping system based on statistical regularities in the language environment that can accommodate novel lexical entries as soon as they are encountered.

1. Introduction

1.1 Pseudowords and Their Meanings

Imagine receiving a text from a friend asking, "Did you feed the cat?". You could easily understand the meaning of what they are asking for. Now imagine this text: "Did you feed the quocky?". You would have more trouble understanding what a *quocky* is, but you could probably still be able to rely on some contextual information; it might be a slang word that your friend uses to name cats or other animals. Now imagine that you are walking and you hear someone saying: "quocky!". Again, you might have no real idea what a *quocky* is, but you could rely on the tone of the voice (and other non-verbal sources of information) to try to grasp its meaning. Finally, imagine reading on a sign by the street with the word "quocky" shown in isolation, without any external sources of information. Would you still be able to intuitively understand its meaning?

This empirical question is particularly intriguing as, on the one hand, it allows us to investigate and understand how speakers deal with novel (verbal) information, and on the other hand, it fits within the fervent debate regarding the relationship between word form and meaning (Dingemanse et al. 2015; Haslett and Cai 2023). Previous studies have investigated this topic by using verbal stimuli labeled "novel words" or "pseudowords". Briefly, these labels are generally used to refer to strings of letters that are consistent with the orthotactical rules of a given language, but do not appear in the language (e.g., "boppies", "knoddled", or "quocky"), and thus can be safely assumed to not be known to a given speaker.[1]

Contrary to the naïve perspective that these out-of-vocabulary stimuli are completely meaningless, recent behavioral studies have shown (by demonstrating that well-established semantic effects in word processing literature are also found for pseudoword targets) that pseudowords can be indicative of meaning (Bonandrini et al. 2023; Chuang et al. 2021; Hendrix and Sun 2021) and that the same mechanisms governing word meaning can also subserve pseudowords (Gatti, Marelli, and Rinaldi 2023). Other studies have further demonstrated that humans are able to reliably assign affective content (Aryani, Isbilen, and Christiansen 2020; Gatti et al. 2023; Sabbatino et al. 2022; Sulpizio, Pennucci, and Job 2021) or lexical categories (Cassani, Chuang, and Baayen 2020) to these stimuli. A recent study by Pugacheva and Günther (2024) even showed that speakers, when instructed to do so, can generate pseudowords to communicate given word meanings to other speakers, who are then occasionally able to infer the original word meanings back from the produced pseudowords.

1 In the present article, we use the term **pseudoword** as described here, **word** for letter strings that are part of the lexicon and do appear in a language, and **item** or **(pseudo)word** as umbrella terms for both.

These findings speak in favor of humans' ability to detect systematic and statistical regularities in the (language) environment (Romberg and Saffran 2010; Vidal et al. 2021) to better adapt to it. However, one main limitation of these studies is that, while they suggest that humans are able to exploit certain systematic regularities to make sense of these out-of-vocabulary stimuli, their investigation was mainly limited to behavioral effects attributed to semantics in word processing studies (Bonandrini et al. 2023; Gatti, Marelli, and Rinaldi 2023; Hendrix and Sun 2021) and did not consider the participants' interpretation of the stimuli. That is, these studies suggested that pseudowords do elicit meaning-related responses and can be associated with quantifiable semantic dimensions such as valence and arousal (Aryani, Isbilen, and Christiansen 2020; Gatti et al. 2023), but it is not very clear if this also means that humans are able to produce non-random detailed, free interpretations of their meaning. In the present study, to fill this gap, we analyze speakers' free definitions of pseudowords in two studies by taking advantage of language models that are able to provide quantitative representations for stimuli not included in their underlying training corpora.

1.2 (Pseudo)Word Meanings in Language Models

Defining meaning has proven to be a challenging endeavor across all disciplines that deal with semantics. Philosophical and linguistic approaches define meaning in relation to states of the world (Chierchia and McConnell-Ginet 2000), whereas in psychology, semantics is usually defined by some form of relationship between words and conceptual structure (Murphy 2004; see Lake and Murphy 2023 for an overview). When dealing with pseudowords, adopting the definition of meaning rooted in formal linguistics leads to obvious complications: If one posits that the meaning of the word *cat* is the set of all the cats (or, equivalently, a function that, given an entity, returns a truth value according to whether the entity is a cat or not; see Delfitto, Zamparelli et al. 2009), then, the meaning of "knackets" would be essentially identical to the meaning of "boppies" (an empty set, or a function that returns False for every entity it is applied to). While adjudicating among different approaches to semantics is far beyond the scope of this article, we adopt a conceptual approach to (pseudo)word meaning, whereby (pseudo)words are considered as "pointers" to some components of the conceptual system (Lake and Murphy 2023). Then, we choose to use language models (LMs) as practical approximations for the conceptual semantic content to which such (pseudo)words are pointing.

Language models are neural networks, generally based on the Transformer architecture (Vaswani et al. 2017), that are trained on large collections of texts to predict the next token following a given sequence (a task that is referred to as **causal** or **auto-regressive** language modeling) or some masked tokens within the sequence (*masked* language modeling). Unlike their predecessors (Recurrent Neural Networks) that process data sequentially, transformer-based architectures can process the input data in parallel by relying on *self-attention* to weigh the importance of the different tokens in the input. The self-attention mechanism enables the model to prioritize different segments of the input data by assessing the relevance of each piece of information to the word being processed. This allows Transformers to capture contextual relationships within the data. Self-attention computes a weighted sum of all the values in the sequence, where the weights are based on the compatibility between keys and queries. Essentially, self-attention allows the models to dynamically focus on different parts of the input text to better capture complex dependencies between words and phrases. This process

involves computing a weighted sum of the token representations, where the weights are determined by the model's assessment of how each input element relates to the rest. Contemporary LMs usually have a large capacity in terms of number of parameters, often scaling into billions (Brown et al. 2020), and it has been shown that the neural capacity of LMs is a crucial determinant of their performance (Kaplan et al. 2020). While learning to predict masked or upcoming words, LMs get tuned to the statistical patterns and regularities of language. This process causes their inner representations to encode linguistic properties that serve as high-level conditioning features with predictive capacity for language modeling. Such properties have been shown to extend to several levels of linguistic analysis, spanning from sublexical properties to discourse, encompassing syntax and semantics (Jawahar, Sagot, and Seddah 2019).

From a cognitive perspective, LMs have been proposed to develop some degree of isomorphism between their internal representations and the representations that humans produce when processing language, as assessed across several metrics and tasks. This proposal has been corroborated by the observation that contextualized word and sentence embeddings can successfully predict functional neural data in the human language network (e.g., fMRI and ECoG recordings, see Caucheteux and King 2022; Schrimpf et al. 2021; Toneva and Wehbe 2019; Tuckute et al. 2024), behavioral measurements of incremental sentence processing difficulty (De Varda and Marelli 2023; Shain et al. 2024; Wilcox et al. 2020, *inter alia*), and word meaning judgments and processing times (Cassani et al. 2023). We must note, however, that there are still important differences in the ways humans and language-based statistical models process and represent language, including the different strength of the links to goals and beliefs (Lake and Murphy 2023), grounding to perception and action (Borghesani and Piazza 2017; Andrews, Vigliocco, and Vinson 2009; Glenberg and Robertson 2000; Chemero 2023), and different learning efficiency (Linzen 2020; Warstadt and Bowman 2022).

LMs represent contextualized token meanings as dense numerical vectors (also referred to as "embeddings") that reflect the distributional history of the tokens as encountered during training (Lenci et al. 2022). This distributional history of a word is highly informative of its meaning, as words with similar meanings also tend to be used similarly in the same linguistic contexts (Harris 1954; Lenci 2008). Using such distributional vectors or word embeddings as computational models of semantic representation stands in a long and empirically successful tradition, also in psychology and cognitive science (Günther, Rinaldi, and Marelli 2019; Jones, Willits, and Dennis 2015; Kumar, Steyvers, and Balota 2021; Landauer and Dumais 1997). Central to this process of meaning induction in LMs is the concept of sub-word tokenization, which defines how the models divide text sequences into smaller units. While early language models relied on word-level tokenization, treating each word as a distinct unit (Bengio et al. 2003), this approach had limitations, particularly in handling the vast vocabulary that characterizes several human languages with productive word formation processes (Sennrich, Haddow, and Birch 2016). Items that were not present in the training data (out-of-vocabulary items) posed a substantial challenge, as the models had no reference for their distributional properties, limiting the models' ability to process text involving these lexical items. Sub-word tokenization emerged as a solution to this problem. Techniques like Byte Pair Encoding (BPE; Sennrich, Haddow, and Birch 2016) or its variants are commonly used to pre-process the text that is fed to LMs. These methods split words into smaller sub-word units based on frequently occurring character sequences; more specifically, BPE iteratively merges the most frequent character pairs in the training data creating a vocabulary of larger character sequences. As an example, the BPE-based

GPT2 tokenizer breaks down the word *obfuscate* into the sub-word units *ob*, *fusc*, *ate*. This approach allows the tokenizer to process a wider range of items, including those not explicitly present in the training dataset. Crucially, sub-word tokenization enables the model to represent also lexical items that do not effectively exist. For instance, the GPT2 tokenizer can properly encode the pseudoword *arwarts* with the sub-word tokens *ar*, *warts*. The vector representation of such a pseudoword will reflect the distributional history of the sub-word tokens, each weighted with self-attention. This approach allows us to project words and pseudowords onto the same embedding space, which in turn allows us to compare similarity values between pseudoword representations and their speaker-produced free definitions.

Note that our study is not the first to use computational models that operate at the sub-word level to investigate semantic and syntactic effects in pseudoword processing. Cassani, Chuang, and Baayen (2020) showed through computational simulations that it is possible to infer word category (noun vs. verb) from the phonological form of pseudowords implementing a linear mapping between the words' form and semantic vectors. Chuang et al. (2021) utilized a linear discriminative learning model to demonstrate that the semantic neighborhoods of pseudowords, represented as numerical vectors, can predict reaction times in a lexical decision task (see also Hendrix and Sun 2021). In a similar vein, Gatti, Marelli, and Rinaldi (2023) derived semantic vectors for both word and pseudoword stimuli with fastText (Bojanowski et al. 2017), and used model-based similarity estimates to predict human response data in a semantic priming task. Bonandrini et al. (2023) compared fastText—an embedding model using sub-word tokens—and compositional semantic models in predicting human performance on a lexical decision task involving affixed pseudowords, highlighting the cognitive relevance of morphological structure in semantic access. Pugacheva and Günther (2024) utilized distributional semantic models in a taboo game setting to illustrate that model-based semantic proximity between target words and participant-generated novel words can facilitate word recognition and comprehension. These previous studies collectively demonstrate the feasibility and effectiveness of using computational tools to derive semantic representations for pseudowords that align closely with human cognitive processes. However, our study departs from prior approaches by utilizing contextualized language models, enabling us to derive embeddings not only for pseudowords and words but also for the sentences through which participants define them. This advancement allows us to assess the relationship between lexical items and the declarative semantic content they are associated with (expressed in the form of sentences), offering new insights into the ways participants produce free definitions of pseudowords.

1.3 Objectives

In the present work, utilizing a range of different LMs and exploiting their ability to approximate the meaning of pseudowords, we first re-analyzed free word and pseudoword definitions from a previous study by Gatti et al. (2023) in an exploratory analysis (Study 1). Gatti et al. (2023) asked participants to provide a written definition of the (possible) meaning of pseudowords following a decision on their affective content. In a subsequent confirmatory analysis (Study 2, which was preregistered after Study 1, `https://doi.org/10.17605/OSF.IO/C2QG3`), we aimed to replicate the results of Study 1 on a new set of pseudowords (and existing words) in a controlled and well-balanced experiment.

2. Study 1: Exploratory Analysis

2.1 Method

2.1.1 Dataset. In this first exploratory study, we used a dataset consisting of (pseudo)words and their definitions collected by Gatti et al. (2023, Experiment 3) in an online experiment. The pseudowords used as stimuli were constructed using Wuggy (Keuleers and Brysbaert 2010). Starting from a given word, Wuggy allows for the generation of written polysyllabic pseudowords that obey a given language's phonotactic constraints and that match its template in sub-syllabic structure. These stimuli are highly word-like but also not easily identifiable as related to existing words. In the Gatti et al. (2023) paper, Wuggy was set using its standard parameters, that is: orthographic English module, restricted match length of sub-syllabic segments, restricted match letter length, restricted match transition frequencies, and match segments 2 out of 3. This data was originally collected in a study investigating the valence of pseudowords, in which participants had to select the most negative and most positive out of sets of six pseudowords. In order to encourage participants to create some meaning representations for these pseudowords, they were then asked to provide a possible definition for the words they selected using an affirmative sentence (i.e., defining what it *is*, not what it is not). Participants were instructed that in cases where a word was new to them (i.e., also for pseudowords), they should still provide what they thought was a plausible meaning or interpretation.

In addition to 25 trials consisting only of pseudowords, each participant was also presented with the same two practice trials and four catch trials consisting of six existing words with clear correct responses (to check if participants actually performed the task as intended). The complete dataset, collected from 112 participants, includes 1,356 definitions for 31 different existing words (with between 1 and 112 definitions per word) and 5,640 definitions for 499 pseudowords (with between 1 and 29 definitions per pseudoword).

2.1.2 Modeling. Words and pseudowords were embedded along with their definition onto a shared vector space through statistical language models based on the Transformer architecture (Vaswani et al. 2017). Eighteen pre-trained Transformer models were considered. Out of these 18 models, 14 were standard causal (or auto-regressive) language models trained on next-word prediction. The causal language models considered were the original GPT model (Radford et al. 2018), four models in the GPT2 family (Radford et al. 2019), three models in the GPT-Neo family (Black et al. 2021), and six models in the Pythia family (Biderman et al. 2023). The additional four models were masked language models, two from the BERT family (Devlin and Toutanova 2019) and two from the RoBERTa family (Liu et al. 2019). The pre-trained models were used as out-of-the-box language representation models, and did not undergo any fine-tuning or adaptation process. The models were accessed through the `transformers` Python library (Wolf et al. 2020). A summary of the models considered in this study is provided in Table 1.

The LMs described above were utilized to derive data-driven semantic representations for both (pseudo)words and their definitions. The models we considered operate at the sub-word token level, generating embeddings for linguistic units equal to or smaller than a word. Sub-word vector representations were obtained from the last layer of the Transformer models (i.e., the layer prior to next- or masked-word prediction; see, for instance, Li et al. 2020). Out-of-vocabulary characters (e.g., curly quotation marks)

Table 1
Summary of the language models considered in the study.

Model	Type	Family	Parameters	Citation
GPT	Causal	GPT	117M	Radford et al. 2018
$GPT2_{124M}$	Causal	GPT2	124M	Radford et al. 2019
$GPT2_{355M}$	Causal	GPT2	355M	Radford et al. 2019
$GPT2_{775M}$	Causal	GPT2	775M	Radford et al. 2019
$GPT2_{1.5B}$	Causal	GPT2	1.5B	Radford et al. 2019
$GPT\text{-}Neo_{125M}$	Causal	GPT2	125M	Black et al. 2021
$GPT\text{-}Neo_{1.3B}$	Causal	GPT2	1.3B	Black et al. 2021
$GPT\text{-}Neo_{2.7B}$	Causal	GPT2	2.7B	Black et al. 2021
$Pythia_{70M}$	Causal	Pythia	70M	Biderman et al. 2023
$Pythia_{160M}$	Causal	Pythia	160M	Biderman et al. 2023
$Pythia_{410M}$	Causal	Pythia	410M	Biderman et al. 2023
$Pythia_{1B}$	Causal	Pythia	1B	Biderman et al. 2023
$Pythia_{1.4B}$	Causal	Pythia	1.4B	Biderman et al. 2023
$Pythia_{2.8B}$	Causal	Pythia	2.8B	Biderman et al. 2023
$BERT_{110M}$	Masked	BERT	110M	Devlin and Toutanova 2019
$BERT_{340M}$	Masked	BERT	340M	Devlin and Toutanova 2019
$RoBERTa_{125M}$	Masked	RoBERTa	125M	Liu et al. 2019
$RoBERTa_{360M}$	Masked	RoBERTa	360M	Liu et al. 2019

were replaced with their in-vocabulary counterparts (straight quotes). Occurrences of these cases were very rare (0.0023% of the characters in the definitions). Word-level representations for all words and pseudowords were obtained by averaging over the sub-word tokens composing the word. Sentence-level representations for the definitions were obtained by averaging over the word-level representations of the included words, in accordance with previous work using mean-pooling to derive definition embeddings (Giulianelli et al. 2023).

In order not to artificially inflate the similarity between matching pairs of (pseudo)words and definitions, the occurrences of the (pseudo)word to be described were removed from the definitions. For instance, if a participant produced the sentence "Sibre is a noun referring to a musical instrument similar to a flute" as a definition of the pseudoword *sibre*, the string *sibre* was deleted from the definition. Then, for the same purpose, definitions that, after the removal of the target (pseudo)word, still contained sub-word tokens that were shared with the target (pseudo)word were eliminated from the following analyses (e.g., definitions containing *swing* when the target pseudoword was *baxswing*).

The word and the pseudoword data were analyzed separately. As a first step, we measured the cosine similarity of all the (pseudo)words with all the (pseudo)word definitions. These cosine similarity values represent our dependent variable, and we evaluated whether they varied as a function of the target-definition match. In other words, we evaluated whether the model-quantified similarity scores were different for actual target-definition pairs vis-à-vis all other pairs of elements. A schematic representation of our analytical approach is presented in Figure 1. To account for the hierarchical structure of the data, characterized by non-independence of observations at multiple levels, we conducted our analyses with linear mixed-effects models with random intercepts for (a) the (pseudo)word for which the definition was provided (*word_from_sent*), (b) the (pseudo)word being compared with the definition (*word_compared*), (c) the participant

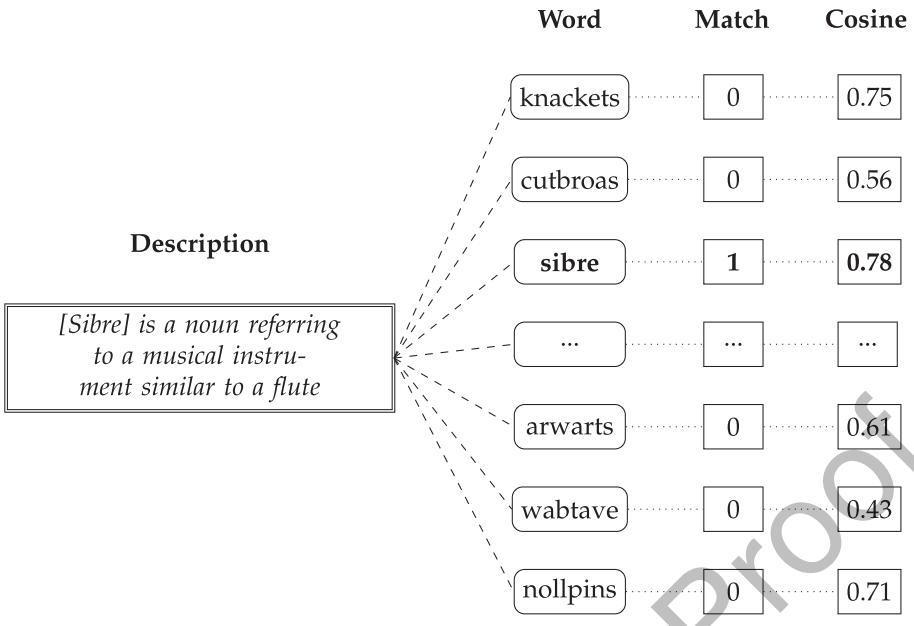


Figure 1
Graphical depiction of the experimental approach. Each definition embedded with a given language model is compared with the vector representation of all the (pseudo)words in the dataset, both matching and non-matching; the cosine similarity between the vector representation of the definition (the sentence embedding) and the (pseudo)word is the dependent variable in the analyses. The item within square brackets is excluded before deriving the sentence embedding.

id (*participant*), and (d) the definition id (*sentence_id*).[2] The linear mixed-effects models were fit in Julia with the package `MixedModels`. We expected matching (pseudo)words and definitions to display a higher level of similarity, thus we anticipated the coefficient and *t*-value associated with our predictor to be positive and statistically significant.

The analyses presented above test whether the similarity between matching pairs of (pseudo)words and definitions is above chance. To better interpret the extent to which such similarity scores outperform chance level, we considered three additional metrics based on hit at k (Hit@k). For each (pseudo)word definition, we ranked all the (pseudo)word embeddings according to their cosine similarity with the definitions. Then, we recorded whether the target (pseudo)word was listed among the top k (pseudo)words, with $k \in (1, 10, 20)$. We repeated this procedure for all definitions and averaged the scores. This is effectively a version of Hit@k with a single relevant document for each definition. Thus, Hit@k was calculated as detailed in (1):

$$\text{Hit@K} = \frac{1}{N}\sum_{j=1}^{N} \max_{i=1\ldots k} \begin{cases} 1 & \text{if } w_{ij} = t_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

2 The models were fitted with the formula *cosine* ∼ *match* + *(1 | word_from_sent)* + *(1 | word_compared)* + *(1 | participant)* + *(1 | sentence_id)*.

Where N is the number of definitions, k is the cutoff value for the rank of considered (pseudo)words, w_{ij} is the i^{th}-ranked pseudoword for the j^{th} definition, and t_j indicates the target (pseudo)word for the j^{th} definition. Simply put, this metric indicates how often the target (pseudo)word is listed among the top 1, 10, or 20 closest (pseudo)word neighbours with respect to the corresponding definitions.

The analyses described in this section allow us to test the semantic similarity between the pseudowords and the corresponding participants' definitions, as measured through the set of candidate models listed in Table 1. Using a set of candidate models makes it possible to assess the robustness of our findings with respect to some model properties such as their pre-training objective, data, and architecture. This entails that our effect of interest, namely, the participants' ability to produce (pseudo)word definitions that resemble the (pseudo)word meanings, could be quantified differently by different models. We chose to operationalize our effect of interest as measured with each model both on the word and the pseudoword data as the t-value associated with the main regressor of interest, that is, the independent variable indicating for a given observation whether word and definition were matched or not. T-values in regression analysis are standardized, meaning they are scaled to account for the variability in the data and the scale of the coefficients. This makes them a useful metric for comparison across different datasets because they provide a consistent measure of the strength of each predictor's relationship with the dependent variable. The quantification of the effect of interest according to each model allowed us to choose the most suitable architecture for the confirmatory part of our study (Study 2), as well as examine the relationship between the estimated effect according to each model in the word and the pseudoword data.

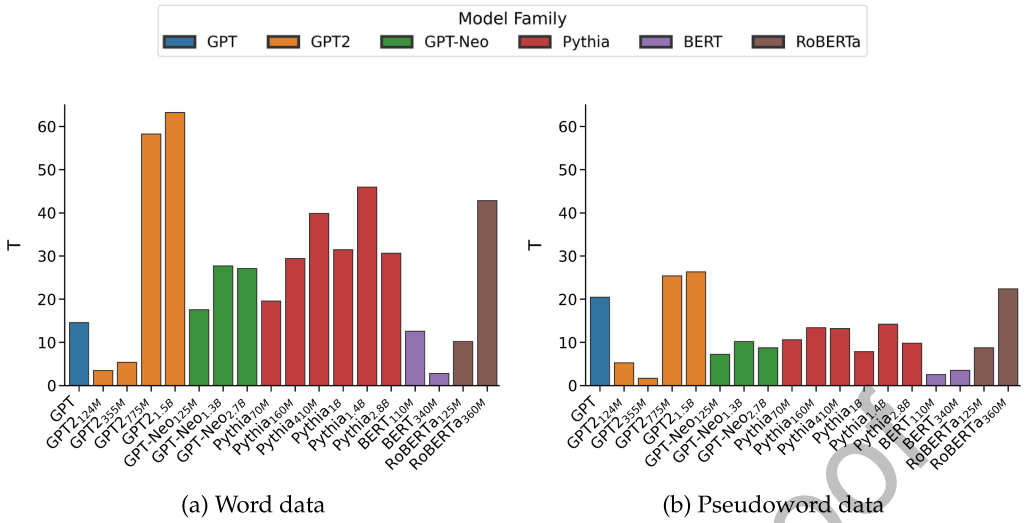
The code and data supporting our study can be found on OSF.[3]

2.2 Results

Figure 2 depicts the t-values associated with the categorical regressor of interest, both for words (2a) and pseudowords (2b). The results presented in Figure 2 are reported in numerical format in Table 2, along with the models' Hit@k scores. The results show that, across the word and the pseudoword data, matching (pseudo)word-definition pairs are associated with higher cosine similarity values than their non-matching counterparts. The results are significant for all the considered models, except for GPT2_{355M} in the pseudoword data (t = 1.6907, p = 0.0909). A clear pattern that emerges from Figure 2 is that the resemblance to the definitions is more evident in the word than in the pseudoword data, as expected. Note that t-values are sensitive to sample size, and the number of datapoints in the word data was substantially smaller than on the pseudoword data; thus, the figure under-represents the difference between the models' performances in the two data subsets.

Among the auto-regressive language models, the two largest models in the GPT2 family (GPT2_{775M} and $\text{GPT2}_{1.5B}$) estimated the strongest effect of our independent variable of interest, whereas the results based on RoBERTa_{360M} obtained the highest t-values among the bidirectional language representation models. $\text{GPT2}_{1.5B}$ achieved a Hit@1 score of 0.3247 for word definitions (with chance level being 31^{-1} = 0.0322) and a Hit@10 of 0.8077 (with chance level being 0.3226), meaning that the target word was the closest word to the definition in about one third of the cases, and among the

3 https://osf.io/5bv8r/.



(a) Word data (b) Pseudoword data

Figure 2
T-values obtained in the linear mixed effects regression with the representations from the models detailed in Table 1.

top 10 words in 80.77% of the cases. Despite the reliable statistical significance, Hit@k values were relatively low in the pseudoword data (although note that the Hit@k metrics are not comparable across word and pseudoword data because of the difference in the number of ranked alternatives). The highest Hit@1 and Hit@10 scores were obtained with $GPT2_{775M}$, with Hit@1 = 0.0066 (with chance level being $499^{-1} = 0.002$) and Hit@10 = 0.0371 (with 0.03 chance level). The strongest Hit@20 performance was obtained by $GPT2_{1.5B}$ (Hit@20 = 0.0682, chance level = 0.04). Overall, Hit@k and the *t*-values from the mixed effects models show a similar pattern of results. In the word data, the highest values in both metrics are obtained by $GPT2_{1.5B}$. In the pseudoword data, $GPT2_{1.5B}$ obtains the highest *t*-values and Hit@20, while $GPT2_{775M}$ is associated with higher Hit@1 and Hit@10 scores than its larger-capacity counterpart; this minor incongruity is not worrisome as $GPT2_{775M}$ and $GPT2_{1.5B}$ produced the second highest estimates for the *t*-values and Hit@1, Hit@10, respectively.

A qualitative pattern that emerges from the model comparison is that, within each model family, representations from larger-capacity models display a tendency to estimate higher *t*-values than their smaller analogs, with the exception of the BERT model family. This is especially noticeable for the GPT2 family, where the two larger models ($GPT2_{775M}$ and $GPT2_{1.5B}$) estimate the highest *t*-values among all models tested, while the two smaller models ($GPT2_{124M}$ and $GPT2_{555M}$) are at the very low end.

Figure 3 represents the relationship between each model's estimate of the effect on the word and the pseudoword data. Our results clearly document a positive relationship between the effect estimated by the models in the two data subsets ($r = 0.81$, $p < .0001$).

2.3 Discussion

Using a dataset of (pseudo)words and their definitions (Gatti et al. 2023), in this first exploratory study we identified that pseudoword definitions are in fact not random: According to most language models we used, cosine similarities were higher for match-

Table 2
Results of the linear mixed effects models (Experiment 1), divided between words (top) and pseudowords (bottom). The table also reports Hit@k values, with $k \in (1, 10, 20)$. The highest Hit@k and t-values for words and pseudowords are highlighted in **bold**.

Model	B	t	p	SE	Hit@1	Hit@10	Hit@20
GPT	0.0196	14.6080	< 0.0001	0.0013	0.1276	0.7153	0.9159
$GPT2_{124M}$	0.0004	3.4545	0.0006	0.0001	0.0303	0.4135	0.7833
$GPT2_{355M}$	0.0006	5.3553	< 0.0001	0.0001	0.0843	0.3484	0.7448
$GPT2_{775M}$	0.0546	58.2387	< 0.0001	0.0009	0.3217	0.7618	0.9334
$GPT2_{1.5B}$	0.0557	**63.2299**	< 0.0001	0.0009	**0.3247**	**0.8077**	**0.9364**
$GPT\text{-}Neo_{125M}$	0.0254	17.5735	< 0.0001	0.0014	0.1087	0.4534	0.7892
$GPT\text{-}Neo_{1.3B}$	0.0348	27.7198	< 0.0001	0.0013	0.1317	0.4763	0.8706
$GPT\text{-}Neo_{2.7B}$	0.0466	27.1207	< 0.0001	0.0017	0.1146	0.4312	0.7670
$Pythia_{70M}$	0.0012	19.5572	< 0.0001	0.0001	0.0732	0.5551	0.7879
$Pythia_{160M}$	0.0191	29.4249	< 0.0001	0.0006	0.0687	0.4885	0.8544
$Pythia_{410M}$	0.0333	39.8707	< 0.0001	0.0008	0.0976	0.5196	0.7583
$Pythia_{1B}$	0.0388	31.4820	< 0.0001	0.0012	0.0517	0.5639	0.8115
$Pythia_{1.4B}$	0.0288	45.9791	< 0.0001	0.0006	0.1013	0.5499	0.9084
$Pythia_{2.8B}$	0.0424	30.6373	< 0.0001	0.0014	0.2047	0.5920	0.7664
$BERT_{110M}$	0.0233	12.6048	< 0.0001	0.0018	0.0363	0.4249	0.7565
$BERT_{340M}$	0.0103	2.8194	0.0048	0.0037	0.0222	0.2346	0.5433
$RoBERTa_{125M}$	0.0043	10.2119	< 0.0001	0.0004	0.1686	0.7611	0.8846
$RoBERTa_{360M}$	0.0013	42.8403	< 0.0001	0.0000	0.1043	0.4756	0.8299
GPT	0.0073	20.4520	< 0.0001	0.0004	0.0041	0.0402	0.0631
$GPT2_{124M}$	0.0002	5.2517	< 0.0001	0.0000	0.0032	0.0260	0.0529
$GPT2_{355M}$	0.0001	1.6907	0.0909	0.0000	0.0021	0.0234	0.0431
$GPT2_{775M}$	0.0090	25.3988	< 0.0001	0.0004	**0.0066**	**0.0371**	0.0676
$GPT2_{1.5B}$	0.0090	**26.3442**	< 0.0001	0.0003	0.0060	0.0363	**0.0682**
$GPT\text{-}Neo_{125M}$	0.0055	7.2277	< 0.0001	0.0008	0.0026	0.0273	0.0482
$GPT\text{-}Neo_{1.3B}$	0.0056	10.2236	< 0.0001	0.0005	0.0030	0.0252	0.0488
$GPT\text{-}Neo_{2.7B}$	0.0067	8.7534	< 0.0001	0.0008	0.0040	0.0258	0.0492
$Pythia_{70M}$	0.0003	10.6332	< 0.0001	0.0000	0.0036	0.0277	0.0531
$Pythia_{160M}$	0.0037	13.4106	< 0.0001	0.0003	0.0021	0.0264	0.0491
$Pythia_{410M}$	0.0050	13.2044	< 0.0001	0.0004	0.0044	0.0294	0.0541
$Pythia_{1B}$	0.0054	7.8603	< 0.0001	0.0007	0.0030	0.0256	0.0493
$Pythia_{1.4B}$	0.0045	14.1977	< 0.0001	0.0003	0.0036	0.0266	0.0558
$Pythia_{2.8B}$	0.0062	9.7680	< 0.0001	0.0006	0.0030	0.0277	0.0522
$BERT_{110M}$	0.0019	2.5638	0.0104	0.0007	0.0034	0.0290	0.0535
$BERT_{340M}$	0.0029	3.4849	0.0005	0.0008	0.0058	0.0320	0.0567
$RoBERTa_{125M}$	0.0012	8.7586	< 0.0001	0.0001	0.0026	0.0307	0.0574
$RoBERTa_{360M}$	0.0002	22.3841	< 0.0001	0.0000	0.0058	0.0333	0.0637

ing than for non-matching pseudoword-definition pairs (i.e., pseudowords are closer to their actual definitions than to the definitions of other pseudowords). Thus, the results are not unique to a specific model. We acknowledge that the Hit@k scores obtained on the pseudoword data are low in absolute terms; nonetheless, they are 55.2% to 120% as high as would be expected by chance. This finding thus still represents a substantial difference from complete randomness, with a t-value higher than 20 for several model types and thus ten times higher than the typical significance threshold of around 2. Complete randomness—that pseudowords presented in isolation should really not have

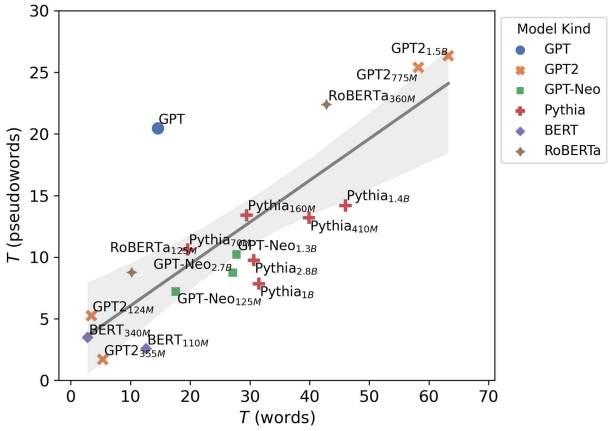


Figure 3
Relationship between the *t*-values associated with each model in the word (*x*-axis) and pseudoword (*y*-axis) data.

any discernable meanings and thus elicit completely random responses—would be a very reasonable assumption in our task, but the statistical analyses provide substantial evidence against this assumption.

The effects for pseudowords mirror the pattern we observe, in a more pronounced manner, for existing words. We wish to remark that the effect observed on the word data should be considered as a sanity check and a baseline for evaluating the effect on the pseudoword sample as estimated by a particular model. In fact, language models that are better able to discriminate between matching and mismatching pseudoword-definition pairs quite clearly tend to be those that are also able to discriminate between matching and mismatching word-definition pairs. Since the latter—accurately capturing word meaning definitions—can be considered a core competence that a language model ought to have, this suggests that the ability to also capture pseudoword meanings is a function of overall language model quality and capacity, rather than a specific niche capability. Among the models we tested, GPT2 (more specifically, GPT2$_{1.5B}$) estimated the strongest matching effect for both existing words and pseudowords. This qualifies GPT2$_{1.5B}$ as the best tool to detect our effect of interest, and we will accordingly use this model in our second, confirmatory study.

It is important to keep in mind that the dataset used in this exploratory analysis was not balanced (as already described in the *Dataset* section): Since participants only provided definitions for the words they selected as most positive or negative, the number of definitions per item varies greatly. This is especially the case for the existing words that were selected to have the clearly most negative and positive items in each trial, but also for the pseudowords where participants display convergent intuitions as to which of these are more positive and negative (Gatti et al. 2023). A second consequence of this setup is that the dataset only contains definitions for pseudowords that participants have already identified as positive or negative. On the one hand, this selects for specific pseudowords to receive more definitions than others; on the other hand, these definitions might be influenced by the decisions participants made (e.g., they might be more inclined to write something more positive after they already selected the word as positive).

Thus, in order to base the interpretation of our results on a firm empirical basis, we replicated this first exploratory study in a second (high-powered and pre-registered) confirmatory study. Here, the first study directly serves to inform the sample size justification, as well as the choice of the language model ($GPT2_{1.5B}$).

3. Study 2: Confirmatory Analysis

In order to generalize the findings of Study 1 and at the same time address the unbalanced nature of the dataset used there, we set up a second confirmatory study. To this end, we collected a well-balanced dataset of word and pseudoword definitions, and replicated the data analysis described in Study 1. Before any data collection, this confirmatory study was pre-registered, including detailed plans for data collection and analysis, at `https://doi.org/10.17605/OSF.IO/C2QG3`.

3.1 Method

3.1.1 Sample Size Motivation. The number of items to be considered in the second confirmatory study was set through a simulation-based bootstrapping power analysis. Starting from the original dataset released by Gatti et al. (2023), we iteratively:

1. Selected a number of participants N, with $N \in (2, 3, 5, 10, 20, 30, 40)$
2. Sampled 100 possible subsets of Gatti et al.'s (2023) data with N participants
3. Refitted our mixed effects model on the downsampled data
    (a) A mixed-effects model was fitted with the original, sampled response variable
    (b) A mixed-effects model was fitted on the simulated response variable
4. Calculated the power for N participants as the ratio of the 100 models that are significant, excluding cases with singular fit.

Note that in the original dataset by Gatti et al. (2023), which was collected to investigate the valence of pseudowords, participants in each trial first selected two out of six items (the most positive and most negative) and then provided definitions for those items they selected. This leads to an unbalanced dataset where the number of definitions varies greatly between (pseudo)words. Our confirmatory study, however, explicitly and deliberately uses a balanced design, in which we collect the same number of definitions for each (pseudo)word. Thus, we use our power analysis to get an estimate for the overall number of observations (i.e., individual definitions) required.

The results of the power analysis are displayed in Figure 4. Our simulations (both with sampled and simulated response variable) showed that we already achieved perfect power (power = 1) with 5 participants, although with 60.15% of singular fits. With 30 participants, we achieved perfect power, with 30.56% of singular fits. With 30 participants in the original study we fitted our models on about 30 (participants) $\times$ 50 (pseudoword definitions per participant) $\times$ 499 (pseudowords being compared with the

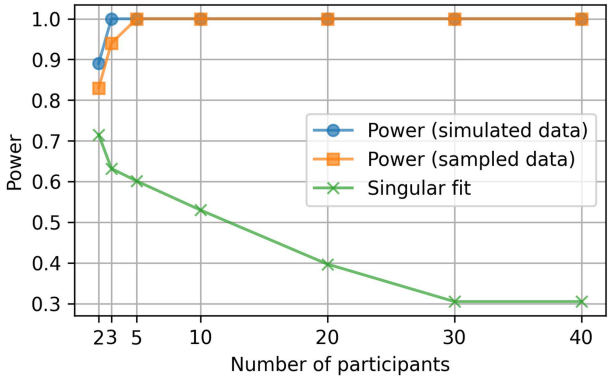


Figure 4
Results of the simulation-based power analysis. The figure reports the statistical power, both with simulated and sampled data, and the percentage of singular fits, as a function of the number of participants.

definitions) = 748,500 datapoints (note that the exact number of items might vary as a function of the tokenizer, as we excluded definitions that shared sub-word tokens with the target pseudoword). In our confirmatory study, we set our sample size in order to have 10 (responses) × 400 (pseudowords) × 400 (pseudowords compared) = 1,600,000 datapoints, ensuring that our study is properly powered even with the changes to the data structure from an unbalanced to a balanced design. Our calculations were based on the pseudoword data, where the effect was documented to be considerably weaker; thus, we are confident that our study is also properly powered for the word data, where we collect the same number of 1,600,000 datapoints.

3.1.2 Dataset. As described in the previous section, the dataset for this second confirmatory study consisted of 400 pseudowords and 400 existing words. The pseudowords were sampled from a list of originally 1,500 pseudowords created and manually curated by Gatti et al. (2023, Experiment 2), from which we excluded 499 pseudowords used in their Experiment 3 (and hence included in our first dataset). The existing words were sampled from the most well-known words of the English language (known by > 99% of participants in the word prevalence norms by Brysbaert et al. 2019), to ensure that participants would indeed know these words and their meanings. We only considered words and pseudowords that do not contain morpheme-like affixes longer than one character and that the $GPT2_{1.5B}$ tokenizer decomposes into 2-4 subword tokens.

These items were randomly arranged into 20 different lists of 40 items (20 pseudowords and 20 words) each. Each of these lists was presented to 10 different participants who provided a definition for each item in the list, with each participant being presented with exactly one list. The final dataset therefore contains 4,000 definitions for 400 pseudowords and 4,000 definitions for 400 words (in both cases, 10 definitions per (pseudo)word).

After informed consent to participate in the study and providing demographic information, participants were instructed to define the meaning of different words presented to them using an affirmative sentence. In cases where a word was new to them, they should still provide what they thought was a plausible meaning or interpretation.

We thus utilized the same instructions as Gatti et al. (2023), described above. After these instructions, participants were presented with one item after the other, and could type their definition into a text box. The instruction "Define the meaning of the following word using a short and affirmative sentence (avoiding negations: define what it is, not what it is not)" was always visible during a trial. After completing two practice trials (using the word item "mouse" and the pseudoword item "kaily" in random order), all 40 (pseudo)words were presented in random order. Only responses that consisted of at least two character strings (separated by a white space) were accepted.

The 200 participants required for this study design were recruited using the crowdsourcing platform Prolific. All participants were monolingual English speakers located in the United States. In this sample ($M_{age} = 40.6$ years, $SD_{age} = 13.5$ years), 114 participants self-identified as female, 83 as male, and 3 chose a different description. Data from an additional 30 participants were not included in the dataset or in any analyses, as these participants often (for at least 8 out of their 20 pseudowords) produced responses that did not comply with the instructions: (a) using the word in a sentence instead of providing a definition, (b) providing responses like "I don't know", (c) providing actual definitions for the pseudowords as words of a foreign language or proper names, most likely as the result of an Internet search, (d) providing responses that were very likely produced by a chatbot, or (e) providing descriptions of the word form ("This is a funny sounding word"). Participants were reimbursed with £3 for their participation in this study (mean completion time: 20.2 minutes).

3.1.3 Modeling. As stated before, in Study 2 we only considered $GPT2_{1.5B}$, the language model that displayed the strongest effects for both words and pseudowords on the exploratory data. Our modeling approach was left unaltered with respect to Study 1; we derived word-level representations by averaging over sub-word token vectors, and averaged over words to obtain definition-level embeddings. We compared definition embeddings with the representations of both matching and non-matching (pseudo)words, and assessed whether the cosine similarity for (pseudo)word-definition pairs was higher when the two matched. We based our analyses on mixed-effects regression models, and left our model structure unaltered with respect to Study 1.

3.2 Results and Discussion

The significant results obtained in Study 1 were successfully replicated and thus confirmed in Study 2. We were able to detect a significant effect of the binary variable of interest in the word data ($B = 0.05269$, $SE = 0.00043$, $t = 122.34$, $p < .0001$), as expected. The target word was the closest to the definition in about 6% of the cases (Hit@1 = 0.0560), and among the closest 10 and 20 words for about 21% and 30% of the definitions (Hit@10 = 0.2167, Hit@20 = 0.3063). Similarly, the effect was significant also for pseudowords ($B = 0.00821$, $SE = 0.0003$, $t = 25.00$, $p < .0001$), with Hit@1 = 0.0055, Hit@10 = 0.0407, and Hit@20 = 0.0776. Chance levels are 0.0025, 0.025, and 0.05 for Hit@1, Hit@10, and Hit@20, respectively. Figure 5 reports a visualization of the cosine similarity values between (pseudo)word embeddings and matching as well as non-matching definition embeddings. Unlike the previous exploratory study, the balanced structure of the dataset used in Study 2 makes the results obtained with the word and the pseudoword data directly comparable. The results clearly show that the effect is stronger in the word than in the pseudoword data, as documented by the *t*-values associated with the regressor of interest in the linear mixed effects models, the Hit@k metrics, and the similarity distributions across conditions.

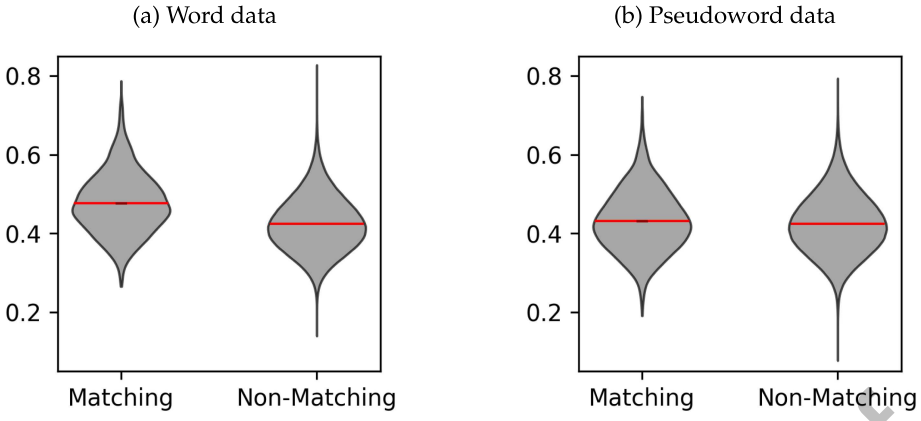


Figure 5
Visualization of the cosine similarity between (pseudo)word and matching and non-matching definition embeddings. The horizontal red lines indicate the means of the distributions. The width of the dark red lines report the 95% confidence intervals for the means. Given our large sample size, the confidence intervals for the non-matching similarities are very small (~ 0.0001) and thus not visible.

Table 3 lists some qualitative examples of the produced pseudoword definitions; in particular, we report the best and the worst definitions (as measured with the cosine similarity between pseudoword and definition embeddings) produced in response to a sample of pseudowords. The table shows that participants associate pseudowords to several semantic categories, including animals (*ottruds*, *kittings*, *skeb*, *silths*), substances (*sulfit*, *cryps*), foods (*teziels*), utensils (*visk*, *skeb*), and geographical locations (*tarvos*). Furthermore, pseudoword definitions convey both abstract (*cryps*, *groanfuls*, *prockack*) and concrete meanings (*sibre*, *skeb*, *visk*).

3.3 Robustness Check

In our main analyses, we assessed whether participants were systematic in assigning definitions to pseudowords by measuring the vector similarity between the pseudoword embeddings and their corresponding definitions embeddings. If participants produced pseudoword definitions based on sub-word information contained in the pseudowords' form, one would expect to find an above chance similarity between pseudoword and definition embeddings. Crucially, one additional expectation that would follow from a systematic form-meaning mapping mechanism linking pseudowords to their definitions is that the definitions for a given pseudoword should be more similar to each other than to all other definitions. For example, a given definition of the pseudoword *teziels* should be more similar to other definitions produced in response to *teziels* than to definitions produced for the word *visk*. Thus, as a robustness check, we conducted an additional analysis aimed at providing additional support for our claims without relying on pseudoword embeddings.

3.3.1 Methods. In this robustness check, we measured whether the definition embeddings were more similar to other definition embeddings corresponding to the same (pseudo)word than to all other definitions. Our analytical approach was rather simple: We calculated the cosine similarity values between all definitions, and compared

Table 3
Qualitative examples of pseudoword descriptions, as provided by participants for the respective pseudowords. For each row, the first line reports the worst description provided for a given pseudoword (i.e., the one with the lowest cosine similarity to the pseudoword representation), and the second line reports the best one. Cosine similarity is calculated based on $GPT2_{1.5B}$ embeddings. The words within square brackets were excluded before deriving the sentence embeddings.

Pseudoword	Cosine	Description
teziels	0.3910	a food that is shaped like a pretzel but is something else
	0.5484	a pebble with irregular sides.
sulfit	0.3485	[sulfit] is a clothing brand
	0.5533	a salt or ester of sulfurous acid.
ottruds	0.3755	it is a small swimming mammal
	0.5445	[ottruds] are turkish footware.
clesp	0.3605	the little hook by the door where you hang your keys
	0.4462	[clesp] is a verb meaning to belittle an opponent's argument unfairly
cryps	0.3125	the gunk in the side of your eye
	0.3979	[cryps] is a synonym for clues in a puzzle
kittings	0.3992	strings or ropes used to hold a balloon down
	0.4972	small feline animals born in a barn
visk	0.3160	[visk] is a disease in which your legs fall off
	0.4619	a container for liquids
skeb	0.3092	a small bug or other pest
	0.5520	a viking weapon.
silths	0.3174	holding up something
	0.4059	a type of worms
tarvos	0.3019	flooring made of wooden planks
	0.5328	a region of Greece
groanfuls	0.4012	a [groanfuls] is a group of teenagers
	0.6635	a cry of pain
sibre	0.3887	[sibre] is a form of animal with a lot of fur
	0.5933	a flamboyant hat
prockack	0.3703	being a professional.
	0.4775	a small kangaroo-like animal

the pairs where definitions were about the same (pseudo)word vis-à-vis all the other pairings. We assessed by means of an independent-samples *t*-test whether the difference between the means of the two groups of similarity scores was statistically significant. As before, we separately conducted our analysis on both pseudowords and existing words. These analyses were not pre-registered, as they were suggested by an anonymous reviewer.

3.3.2 Results and Discussion. These analyses showed that definitions were indeed more similar to other definitions produced in response to the same (pseudo)word than to all other definitions of non-matching items. The effect was significant both for existing words ($t = 193.5664$, $p \ll 0.0001$) and for pseudowords ($t = 20.005$, $p \ll 0.0001$). Since the number of samples was comparable, we can again conclude that the effect was more pronounced for existing than for non-existing words, as expected.

This robustness check serves as a critical validation of our initial conclusions, utilizing an independent methodology that avoids altogether the reliance on pseudoword embeddings. By examining the similarity of definitions assigned to the same (pseudo)words in contrast to those assigned to different ones, without the intermediary of pseudoword embeddings, this analysis strengthens our results by leveraging direct comparisons between the semantic content of the definitions themselves. Once again, the fact that the effect is much more pronounced on the word data is not surprising, as participants providing word definitions can rely on a shared understanding and usage of word meanings. For pseudowords, the significant (albeit less pronounced) effect independently corroborates the notion that participants systematically generate explicit definitions based on the sub-word information that served as a cue, even in the absence of any pre-existing encounters with the lexical items.

4. Follow-up Analyses

In our pre-registered main analyses, we showed that pseudoword representations are closer to the embedding representations of their corresponding definitions than to other definitions. In order to get a better understanding of this phenomenon of pseudoword semantics, we proceeded with two (not pre-registered) follow-up analyses, where we shifted our focus to the word-level properties of words and pseudowords that might influence the ease with which participants are able to associate its form with a meaning.

4.1 Sub-word Units in Pseudoword Semantics

In the first follow-up analysis, we assessed whether pseudowords composed of specific sub-word units were more successfully associated with a free definition of their meaning.[4] This analysis was not driven by explicit predictions about which specific subword units might be particularly beneficial for the task. Instead, it was meant as an exploration aimed at identifying the most meaning-bearing sub-word units. As for all our previous tests, we separately conducted this exploratory analysis for words and pseudowords.

4.1.1 Methods. In order to derive a (pseudo)word-level metric summarizing the semantic association between the representations of each (pseudo)word and all its corresponding definitions, we started from each definition and calculated the distribution of cosine similarity values between its embedding representation and all the (pseudo)word embeddings. This first step is analogous to the base procedure of our main analyses (see Figure 1). Then, we calculated the z-score of the cosine value of the matching (pseudo)word with respect to the similarity distribution. This conversion to z-scores was done to take into account the possibility that some definitions might be, in general, closer to all (pseudo)words, whereas our measurement of interest is intended to capture how close the definitions are to the target word, compared to all other words. As a last step, we averaged the z-scores of all definitions corresponding to a given (pseudo)word, and used this measure as a (pseudo)word-definition proximity metric (henceforth "proximity") at the item level. Table 4 reports some examples of pseudowords and their associated proximity scores for the four quartiles of the proximity measure distribution.

4 We thank an anonymous reviewer for the suggestion.

Table 4
Examples of pseudowords with their associated proximity score, divided by quartiles. The 1st and the 4th quartiles report the highest and lowest-scoring items, respectively, while the items from 2nd and 3rd quartiles are randomly sampled.

1st Quartile		2nd Quartile		3rd Quartile		4th Quartile	
Item	**Proximity**	**Item**	**Proximity**	**Item**	**Proximity**	**Item**	**Proximity**
siseals	2.5333	asarps	0.8060	cruns	0.1270	wheam	−2.1600
aruds	2.3096	lylik	0.6227	febs	−0.1209	clact	−2.1892
groanfuls	2.2754	edsows	0.5914	ellachs	−0.1839	liflo	−2.2280
peisels	2.2579	chufo	0.4741	mundats	−0.2573	freapt	−2.6228
swimacks	2.2395	mambs	0.4658	thrun	−0.4408	splil	−3.2274

Then, we extracted from the $GPT2_{1.5B}$ tokenizer the individual sub-word units composing each (pseudo)word. We filtered out all sub-word units appearing less than 5 times in our dataset, in order to have a reasonable number of occurrences for each sub-word unit. Following this criterion, this analysis considered 29 sub-word units for the word data and 41 units in the case of pseudowords. For each (pseudo)word form, we dummy-coded whether it contained each of these units, and fitted a linear regression model to predict the proximity metric described above on the basis of the presence of each sub-word token.

4.1.2 Results and Discussion. Both our regression model based on the word data (F = 2.7104, R^2 = 0.1643. $p < 0.0001$) and the one based on the pseudoword data (F = 1.8696, R^2 = 0.1724. $p = 0.0016$) were significant, indicating that at least *some* of the considered sub-word units were associated with definitions with different similarity to their corresponding (pseudo)word. Table 5 reports the individual sub-word units that were significantly predictive of our proximity metric (with $p < .05$, uncorrected). The table shows that, in the case of existing words, only the sub-word unit "s" is associated with higher word-definition proximity. Five other units are associated with lower proximity. When considering pseudowords, the sub-word unit "s" was once again associated with higher average pseudoword-definition proximity. We speculate that this consistent finding might be related to the morpheme *-s* indicating plurality in English. Thus, the plural marker could be a meaningful cue both when retrieving the semantic content of a known lexical item and when inferring the meaning of an unknown letter string, and participants might consistently associate words ending with *-s* to definitions referring to plural entities. Apart from "s", several other sub-word units are associated with both higher (e.g., "ons", "h") and lower proximity scores ("amb", "raud").

The outcome of this analysis shows that some specific sub-word units influence the ease with which participants are able to retrieve or infer the meaning of letter strings, both in the case of words and pseudowords. Thus, not only the combination of these units, but also their identity, is associated with the the ability of participants to derive free pseudoword interpretations. Interestingly, most of the units that have a significant impact on pseudoword definition proximity are not morphemes. This finding highlights the fact that the cognitive process underlying the interpretation of novel words is flexible and can utilize various sub-lexical cues for meaning inference.

Table 5
Sub-word units predictive of (pseudo)word-definition proximity.

		B	SE	t	p
Words	s	1.2418	0.3219	3.8582	0.0001
	j	−0.8437	0.3749	−2.2503	0.0250
	sc	−0.9719	0.4026	−2.4143	0.0162
	sn	−1.1731	0.3999	−2.9334	0.0036
	qu	−1.4616	0.4734	−3.0878	0.0022
	sp	−1.5796	0.4011	−3.9383	0.0001
Pseudowords	s	0.6970	0.1793	3.8871	0.0001
	ons	1.1757	0.4266	2.7556	0.0062
	h	0.7900	0.3268	2.4174	0.0161
	fts	< 0.0001	< 0.0001	2.2417	0.0256
	e	0.9087	0.4247	2.1397	0.0330
	amb	< 0.0001	< 0.0001	1.9691	0.0497
	raud	< -0.0001	< 0.0001	−1.9736	0.0492
	olls	< -0.0001	< 0.0001	−2.0680	0.0394
	be	< -0.0001	< 0.0001	−2.2130	0.0275
	night	< -0.0001	< 0.0001	−2.2301	0.0264
	ab	−0.9986	0.3683	−2.7109	0.0070

4.2 Length and Sequential Predictability

In these analyses, we restricted our scope to two main factors: the orthographic length of a pseudoword, and its sequential predictability. We hypothesized that the richness of a pseudoword's form, operationalized as its number of characters, should be positively associated with the participants' success in freely describing its meaning (and thus, with higher proximity between the pseudoword and the definition embeddings). Indeed, if the human ability to describe the semantics of a novel word is supported by a form-to-meaning mapping system, the bare quantity of a pseudoword's form should endow participants with more sub-word material to support such mapping. Regarding the effects of sequential predictability, we expected a tension between two opposing tendencies. Letter sequences that are unpredictable carry more information (Shannon 2001); thus, strings with low sequential predictability should be better associated with their meaning since their word form is more informative (A). On the other hand, very improbable letter sequences will be more distant to the form of existing words, making it more difficult to exploit existing form-meaning mappings (B). Thus, (A) would entail a negative relationship between a pseudoword's sequential predictability and the proximity of its definitions. Conversely, (B) predicts a positive relationship between the two variables. Alternatively, the prediction stemming from (A) and (B) acting simultaneously would be a non-linear relationship between the two constructs, with a local increase in pseudoword-definition proximity when sequential predictability is relatively low (following B), and a decrease for high predictability values (following A).

4.2.1 Methods. As in the previous follow-up analysis, we used as a measure of (pseudo)word-definition proximity the average z-score of the cosine value of the

definition and the matching (pseudo)word with respect to the similarity distribution among the definition and all other (pseudo)words.

Word and pseudoword length was simply calculated as the number of letters contained in each string. Sequential predictability was calculated with character trigrams; in particular, we used character-level negative log-probability values, with the base of the logarithm set to 2. Trigrams, sequences of three characters, offer a more fine-grained estimation of sequence predictability than bigrams (two-character sequences) without the substantial increase in computational complexity and data sparsity associated with higher-order *n*-grams. Furthermore, character-level sequential predictability has been calculated with trigrams in a previous study (Thompson et al. 2022). We calculated trigram probabilities starting from the vocabulary in $\text{SUBTLEX}_{\text{US}}$ (Brysbaert and New 2009). Trigram probabilities were computed at the type—as opposed to token—level; this means that the trigrams composing each word were counted once for each word, irrespective of its number of occurrences in the corpus (as in Thompson et al. 2022). Character-level log-probabilities were summed to derive a word-level estimate of sequential predictability.

To be able to detect non-linear relationships between the variables we consider, we analyzed our data with Generalized Additive Models (GAMs), a regression technique designed for measuring curve shapes (Wood 2006). The GAM models were fitted with the Python package `pyGam`; both sequential predictability and orthographic length were modeled as non-linear splines with a penalty on their second derivative, with identity link and normal error distribution. The λ parameter, which controls the strength of the penalty for the splines, was set with a grid search. GAMs are known to overfit, and, when smoothing parameters have to be estimated, significance testing can reject the null too readily. Thus, to ensure that our findings were robust, we ran a permutation test to obtain more reliable *p*-value estimates. To do so, we focused on each of the two predictors separately, and shuffled its values while keeping the other unmodified. This procedure was repeated 1,000 times for each of the two predictors, and a non-parametric *p*-value was calculated as the proportion of simulations yielding a pseudo-R^2 value higher or equal than the non-shuffled model.

4.2.2 Results. Our analyses showed that, for the pseudoword data, the orthographic length and transitional probability of pseudowords were significantly associated with their proximity scores (length: EdoF = 2.2, $p = 0.0454$; sequential predictability: EdoF = 4.3, $p = 0.001$; pseudo-$R^2 = 0.1649$, $N = 400$). The statistical significance of the smooth terms was corroborated by the non-parametric permutation test (length: $p = 0.005$; sequential predictability: $p < 0.001$). The functional form of the effects of the two predictors is graphically depicted in Figure 6. There is a positive, monotonic relationship between orthographic length and proximity (6b). The steepness of the relationship is higher in the 6-8 letter range. As for the transitional probability effects, we detected a non-linear relationship between trigram negative log-probability and pseudoword-definition proximity (6d). The relationship appears to be quadratic and concave from visual inspection, with a local increase in proximity around the 30-70 information range, and a decrease from around 80 to 130 bits. In the word data, neither orthographic length (6a) nor sequential predictability (6c) were significantly associated with average word-definition proximity (length: EdoF = 1.1, $p = 0.136$; sequential predictability: EdoF = 3.4, $p = 0.496$; pseudo-$R^2 = 0.0744$, $N = 400$). The absence of statistical significance in the word data was supported by the permutation test (length: $p = 0.588$; sequential predictability: $p = 0.960$).

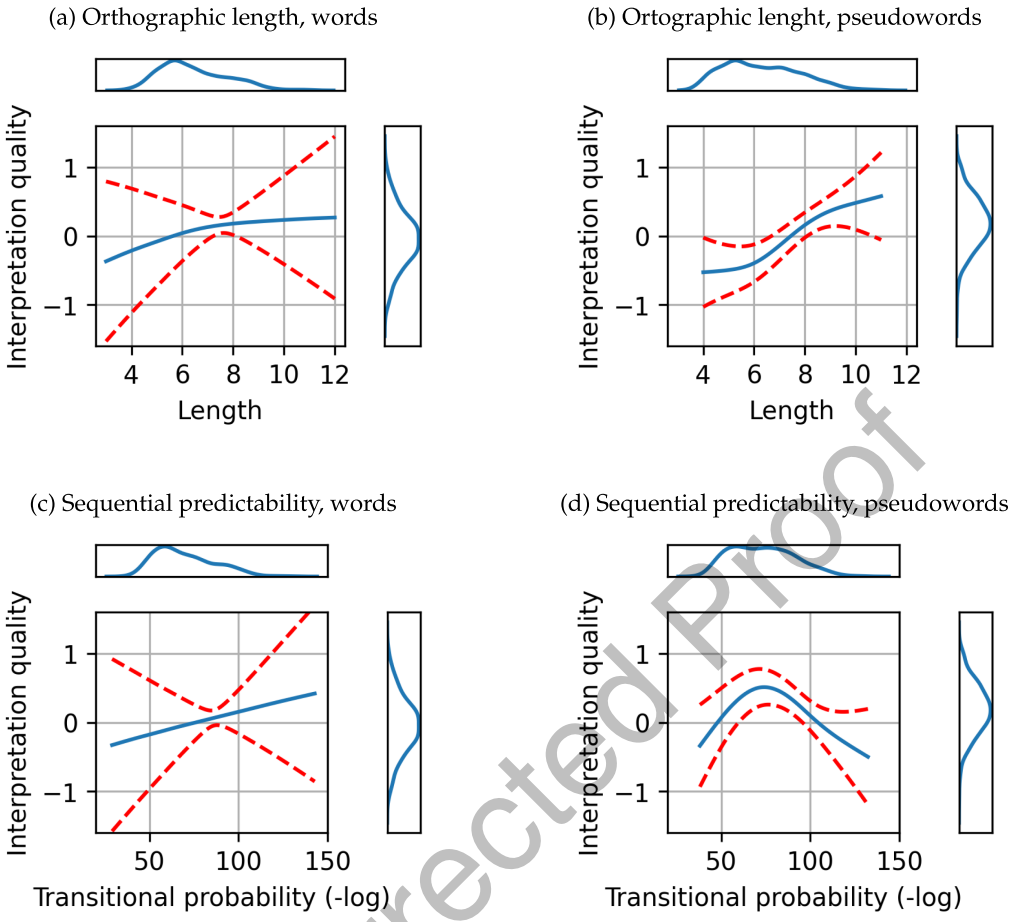


Figure 6
Spline terms associated with orthographic length (top) and transitional probability (bottom) for both words (left) and pseudowords (right). On the *y*-axis, interpretation quality is displayed in terms of the marginal effect of the predictor in question. The dashed red lines indicate the 95% confidence intervals. The subplots on top and on the right of each plot display kernel density estimates of the data distributions corresponding to the the variables on the two axes.

4.2.3 Discussion. Our follow-up analysis explored the role of form quantity and predictability on form-meaning mappings in language, focusing on how orthographic length and sequential predictability of words and pseudowords affect their semantic interpretations. The results are compatible with the idea that participants exploit statistical form-to-meaning regularities when inferring the meaning of words unknown to them. The positive and monotonic relationship between the orthographic length of pseudowords and their proximity to their definitions aligns with our hypothesis: Longer pseudowords provide more sub-word material, facilitating the formation of richer semantic representations for novel lexical entries. This finding contributes to our understanding of how the human language system makes use of the form of a novel word to derive useful clues about its meaning, emphasizing the importance of the raw amount of a word's form content in semantic expansion. The results concerning sequential predictability offer a more complex view of the role of the characters' distributional properties. Our findings align with one of our hypotheses, namely, that two

antithetical pressures might shape the relationship between sequential predictability and pseudoword-definition proximity. At higher predictability (i.e., low information) levels, additional informativeness of the pseudoword's form enhances proximity, supporting the idea that information-richer word forms facilitate access to specific semantic representations (A). However, as the pseudowords' informativeness increases further, the increased distance from known word forms appears to hinder existing form-meaning mappings, reflecting the difficulty in relating highly unfamiliar letter strings to prior semantic knowledge (B). This non-linear trend highlights the tension between informativeness and familiarity in the understanding of novel words. An alternative (and possibly complementary) explanation of this pattern can be understood in terms of the pseudowords' orthographic proximity to known English words. Easily predictable pseudowords may resemble a broad range of English words, leading to difficulties for participants when attempting to interpolate meanings based on this extensive set of familiar lexical items. Conversely, pseudowords with low sequential predictability might bear too little resemblance to familiar English words, challenging the participants' ability to leverage their existing linguistic knowledge for meaning induction.

Intriguingly, the observed patterns did not extend to real words, where neither orthographic length nor sequential predictability had a significant impact on word-definition semantic proximity. The reported disparity between words and pseudowords highlights the flexibility of semantic interpretation processes in response to novel linguistic stimuli compared to the access to the meaning of established lexical items. This asymmetry might be attributed to the more rigid form-to-meaning relationships that drive the understanding of real words. The meaning of a word is already known, and needs to be retrieved—as opposed to inferred—from the form; this entails that form-level properties should have a reduced impact on non-speeded semantic access.

5. General Discussion

Pseudowords, orthotactically viable letter strings that however do not occur in the lexicon of a given language, are traditionally considered as lexical items with a valid form but without any meaning. As such, they play a crucial role for example as control stimuli in word processing studies (Balota et al. 2007; Hutchison et al. 2013). Recent studies have challenged this characterization by collecting evidence that these words do indeed carry semantic content and display some of the same semantic effects as existing words (Gatti, Marelli, and Rinaldi 2023; Gatti et al. 2023; Hendrix and Sun 2021), without however elucidating whether participants can have explicit and deliberate access to detailed meaning representations for pseudowords. In other words, previous research has mainly focused on detecting meaning-related processing correlates in pseudoword recognition, without assessing whether humans can characterize their semantic content with explicit definitions. In the present work, we use LMs to extend these findings, showing across two studies utilizing an exploratory-confirmatory setup that speakers can assign non-random free meaning interpretations to pseudowords: On average, definitions generated for a given pseudoword are closer to that pseudoword than those definitions generated for other pseudowords (as measured by the cosine similarity between their respective vector representations). This reflects, albeit to a lesser degree, a pattern that emerges very clearly also for existing words. The results of our studies not only confirm the presence of a flexible form-to-meaning mapping system in humans, but also reveal a critical parallel between pseudoword representations in humans and language models. This suggests that the cognitive processes enabling humans to infer meanings from novel word forms operate under principles that are, to some extent,

mirrored by the statistical learning processes used by modern language models. This correspondence provides compelling evidence for the cognitive plausibility of distributional learning mechanisms at the sub-word level. When faced with unknown lexical items, humans might decompose them into smaller units and leverage their distributional properties to construct plausible meanings. Such a mechanism would represent a highly efficient strategy for navigating the linguistic environment, enabling speakers to adaptively expand their lexicon by extrapolating from known linguistic patterns to interpret new words. In addition to our main results, we identified in follow-up studies that interpretations are, on average, closer to their respective pseudowords for pseudowords that were longer, and whose sequential trigram probabilities were in the medium range. Here, we identified patterns that are unique to pseudowords and do not generalize to words.

As speakers encountered the pseudowords for the first time and without any relevant context information, their meaning had to be inferred purely from their form. For pseudowords that are not morphologically parsable (see Appendix A for a replication of our results when excluding the four parsable pseudowords from our dataset), such inferences have to be made based on a very general form-to-meaning correspondence, rather than from specific and well-delineated morpho-semantic information. While on a very fundamental level the correspondence between forms and meanings in language has long been considered arbitrary (de Saussure 1916), this view has increasingly come under pressure from empirical studies (for overviews, see Dingemanse et al. 2015; Haslett and Cai 2023). On the one hand, cases of iconicity (where the linguistic form reflects physical characteristics of the associated meaning, as in *meow*) introduce a certain ground-level connection between perceptual experience and linguistic forms that is not specific to a given language (Perniss and Vigliocco 2014). On the other hand, once a given linguistic system is established, there is a non-negligible tendency for words that are similar in form to also be more similar in meaning than would be expected by chance (Marelli and Amenta 2018; Monaghan et al. 2014). Our results fall well into line with these perspectives, suggesting that speakers are able to pick up these systematic patterns and regularities, and use this information to make informed guesses about novel stimuli. Since encountering new words is not a rare phenomenon (Brysbaert et al. 2016, estimate than an average speaker learns a new lemma every two days) this ability allows us to make some predictions that help us make sense of non-familiar stimuli, which allows us to navigate our complex and changing world.

Of course, both our Study 1 and Study 2 show that the relative match of pseudoword definitions to their actual targets is far smaller than it is for existing words. To put this into context, we have to highlight again that for existing words speakers can rely on an established and conventionalized meaning, while for pseudowords they can only rely on their form; and although it exists, systematic form-meaning mapping is a rather weak phenomenon in natural languages (Marelli and Amenta 2018; Monaghan et al. 2014). In addition, for the purpose of our study we had to ensure that all existing words we used are generally known to all English speakers (Brysbaert et al. 2019). These extremely well-known and frequent words, the meaning of which is clear to every speaker, form the toughest possible comparison for the pseudowords. We expect that the discrepancy between words and pseudowords will diminish for words with lower prevalence or frequency. An especially interesting comparison would be to consider words that follow the same criteria as our pseudoword sample but are unknown to most speakers, like *alnico*, *stotinka*, *gomuti*, or *feoff* (very low-prevalence examples from Brysbaert et al. 2019). This would allow us to conclude if there is anything special about existing words at all (which might be the case due to etymological connections, or the

word in principle referring to an actual existing concept for which it made sense to coin it in the first place), or if pseudowords and unknown words are really functionally equivalent.

The fact that we can model human definitions of pseudowords using LMs has relevant implications for current (psycho)linguistic theories of form-to-meaning mapping and representation systems (Stevens and Plaut 2022). Here, the last years have seen an intense debate (see, for example, Bonandrini et al. 2023; Stevens and Plaut 2022) between morphological models (e.g., Marelli and Baroni 2015; Taft and Forster 1975), in which the relevant representational units at the sub-word level are morphemes (the smallest meaning-bearing units in a language), amorphous models (Baayen et al. 2011, 2019) that reject this notion and instead take the embedded letter *n*-grams as their basic representational unit that map directly onto meaning (which is also adapted in word embedding models like fastText; Bojanowski et al. 2017), and distributed connectionist models that start from the same assumption as amorphous models but further assume hidden layers between basic form units and meaning that allow for the emergence of more complex representational units like morphemes (Plaut and Gonnerman 2000). On the one hand, our results show that a traditional morpheme-based representation system cannot be a complete account for a form-to-meaning mapping system, because pseudowords do not have a morphemic structure and cannot be parsed into what one would traditionally consider morphemes. To fully uphold such a morpheme-based view, one would have to accept the notion that the sub-word units identified by the BPE—such as *ob*, *fusc*, and *ate* in *obfuscate*—essentially have morpheme status: In the model architecture, they are the smallest units for which vector representations are available and thus very literally the smallest meaning-bearing units. On the other hand, the success of the BPE algorithm—demonstrated by its widespread adoption in contemporary language models—suggests that it may be useful to not simply consider *all* embedded *n*-grams as meaning-bearing representational units. However, this depends on whether models using BPE generally outperform across a wide range of representative tasks models based on *n*-gram tokenizers, which remains to be tested in future studies. If this indeed turns out to be the case, the BPE-based sub-word representations used by language models would align best with the hidden-layer representations in distributed connectionist models: They differ in length and are more structured than *n*-grams, they emerge from experience with the linguistic system, but the exact set of these units can also differ substantially from the morphemes identified in the traditional literature.

6. Future Directions

Our approach provides experimental evidence in support of a flexible form-meaning mapping system and opens new avenues for research on the mechanisms that support it. Starting from these results, future studies could address which properties of the pseudoword forms contribute to particular semantic interpretations, thus providing an algorithmic account on how the system operates. This component is a relevant aspect of the pseudoword interpretation process and raises further questions about the exact nature of the cognitive processes involved. Thus, while our study sheds light on the capacity of individuals to ascribe meaning to pseudowords, it opens up avenues for further exploration into the intricate dynamics of form-meaning relationships in language. The data we have made available offers a valuable starting point for this continued investigation.

An additional line of research where language models could be used as a critical tool in understanding pseudoword interpretation processes is the analysis of the role of

the (linguistic) context in understanding novel lexical items. In this study, we deliberately focused on the word-internal mechanisms supporting the extraction of semantic information from form alone. On the other hand, previous research has also shown that the (sentence) context in which a novel lexical item is encountered plays a major part in forming a semantic representation for said item (e.g., Borovsky, Kutas, and Elman 2010; Günther et al. 2020; Lazaridou, Marelli, and Baroni 2017). Future studies should thus investigate the interplay between such contextual information and the word-level information encoded in the pseudowords' form: Language models have the ability to process and represent language based on context. By incorporating contextual analysis, future research could explore how surrounding words, sentences, or even larger text structures contribute to the meaning that humans attribute to pseudowords.

From a practical standpoint, our approach could be helpful in providing additional experimental control in studies utilizing pseudowords as stimuli. For instance, in pseudoword learning experiments, researchers might want to ensure that the sub-word content of a pseudoword is not associated a priori with certain semantic properties that the participants are expected to learn during the experiment (which so far required additional data collection from human participants; see for example Günther, Dudschig, and Kaup 2018). More generally, it is likely that researchers in psycholinguistics might want to control for the semantic content that is encoded in a pseudoword's form. In cases where such semantic content might be expressed as a single word, simpler alternatives such as fastText might suffice; however, if the semantic content to be controlled is more complex and needs to be articulated at the sentence level, our approach might provide a straightforward way to measure in a data-driven fashion the semantic association between a pseudoword's form and declarative meanings.

A: Removal of pseudo-compounds

To ensure that the effect we reported could be ascribed to general sub-word form-to-meaning relationships (as opposed to consolidated word-level semantic knowledge), we took care to re-run our main confirmatory analysis after the exclusion of pseudo-compound words. Indeed, semantic composition relies on established word-level semantic features, and this process of conceptual combination is outside of the scope of the paper. In the data we collected for Study 2, there were four morphologically parsable pseudowords (*switchcraps*, *punstack*, *popend*, *danchunk*). Thus, as a control, we repeated the main analysis after removing those words and their corresponding definitions. Our results showed that the main effect of our independent variable of interest was not critically dependent on the presence of pseudo-compounds in our data, as it remained significant after their exclusion ($B = 0.00809$, $SE = 0.00032$, $t = 24.56$, $p < .0001$). The target pseudoword was the closest to its associated definition in about .5% of the cases (Hit@1 = 0.0056; chance level: 0.0025), and among the closest 10 and 20 words for about 4% and 8% of the definitions (Hit@10 = 0.0408, Hit@20 = 0.0782, with chance level being 0.0253 and 0.0505, respectively).

Acknowledgments

Fritz Günther received funding from the German Research Foundation (DFG), Emmy-Noether grant "What's in a name?" (project number 459717703). The contribution of Marco Marelli was supported by the European Union (ERC-COG-2022, BraveNewWord, 101087053). Views and opinions expressed are however those of the authors only and do not necessarily reflect

those of the European Union, the European Research Council Executive Agency, or the German Research Foundation. Neither the European Union nor the granting authorities can be held responsible for them.

References

Andrews, Mark, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463. https://doi.org/10.1037/a0016261, PubMed: 19618982

Aryani, Arash, Erin S. Isbilen, and Morten H. Christiansen. 2020. Affective arousal links sound to meaning. *Pschological Science*, 31(8):978–986. https://doi.org/10.1177/0956797620927967, PubMed: 32662741

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019:4895891. https://doi.org/10.1155/2019/4895891

Baayen, R. Harald, Petar Milin, Dušica Filipović Đurđević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438–481. https://doi.org/10.1037/a0023851, PubMed: 21744979

Balota, David A., Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. 2007. The English lexicon project. *Behavior Research Methods*, 39:445–459. https://doi.org/10.3758/BF03193014, PubMed: 17958156

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Biderman, Stella, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430.

Black, Sid, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. https://doi.org/10.5281/zenodo.5297715

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Bonandrini, Rolando, Simona Amenta, Simone Sulpizio, Marco Tettamanti, Alessia Mazzucchelli, and Marco Marelli. 2023. Form to meaning mapping and the impact of explicit morpheme combination in novel word processing. *Cognitive Psychology*, 145:101594. https://doi.org/10.1016/j.cogpsych.2023.101594, PubMed: 37598658

Borghesani, Valentina and Manuela Piazza. 2017. The neuro-cognitive representations of symbols: The case of concrete words. *Neuropsychologia*, 105:4–17. https://doi.org/10.1016/j.neuropsychologia.2017.06.026, PubMed: 28648571

Borovsky, Arielle, Marta Kutas, and Jeff Elman. 2010. Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition*, 116(2):289–296. https://doi.org/10.1016/j.cognition.2010.05.004, PubMed: 20621846

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

Brysbaert, Marc, Paweł Mandera, Samantha F. McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51:467–479. https://doi.org/10.3758/s13428-018-1077-9, PubMed: 29967979

Brysbaert, Marc and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990. https://doi.org/10.3758/BRM.41.4.977, PubMed: 19897807

Brysbaert, Marc, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. 2016.

How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7:1116. https://doi.org/10.3389/fpsyg.2016.01116, PubMed: 27524974

Cassani, Giovanni, Yu-Ying Chuang, and R. Harald Baayen. 2020. On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46:621. https://doi.org/10.1037/xlm0000747, PubMed: 31318232

Cassani, Giovanni, Fritz Günther, Giuseppe Attanasio, Federico Bianchi, and Marco Marelli. 2023. Meaning modulations and stability in large language models: An analysis of bert embeddings for psycholinguistic research. *psyArXiV preprint: 10.31234/osf.io/b45ys*. https://doi.org/10.31234/osf.io/b45ys

Caucheteux, Charlotte and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134. https://doi.org/10.1038/s42003-022-03036-1, PubMed: 35173264

Chemero, Anthony. 2023. LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour*, 7(11):1828–1829. https://doi.org/10.1038/s41562-023-01723-5, PubMed: 37985905

Chierchia, Gennaro and Sally McConnell-Ginet. 2000. *Meaning and grammar: An introduction to semantics*. MIT Press.

Chuang, Yu Ying, Marie Lenka Vollmer, Elnaz Shafaei-Bajestan, Susanne Gahl, Peter Hendrix, and R. Harald Baayen. 2021. The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*, 53(3):945–976. https://doi.org/10.3758/s13428-020-01356-w, PubMed: 32377973

de Saussure, Ferdinand. 1916. Nature of the linguistic sign. *Course in General Linguistics*, 1:65–70.

De Varda, Andrea and Marco Marelli. 2023. Scaling in cognitive modelling: A multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149. https://doi.org/10.18653/v1/2023.acl-short.14

Delfitto, Denis, Roberto Zamparelli, et al. 2009. *Le strutture del significato*. Il mulino.

Devlin, Jacob, Ming Wei Chang, and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Dingemanse, Mark, Damian E. Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan. 2015. Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, 19(10):603–615. https://doi.org/10.1016/j.tics.2015.07.013, PubMed: 26412098

Gatti, Daniele, Marco Marelli, and Luca Rinaldi. 2023. Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology: General*, 152(3):851–863. https://doi.org/10.1037/xge0001304, PubMed: 36174173

Gatti, Daniele, Laura Raveling, Aliona Petrenco, and Fritz Günther. 2023. Valence without meaning: Investigating form and semantic components in pseudowords valence. *psyArXiV preprint: 10.31234/osf.io/sfzgr*.

Giulianelli, Mario, Iris Luden, Raquel Fernández, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148. https://doi.org/10.18653/v1/2023.acl-long.176

Glenberg, Arthur M. and David A. Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 43(3):379–401. https://doi.org/10.1006/jmla.2000.2714

Günther, Fritz, Carolin Dudschig, and Barbara Kaup. 2018. Symbol grounding without direct experience: Do words inherit sensorimotor activation from purely linguistic context? *Cognitive Science*, 42:336–374. https://doi.org/10.1111/cogs.12549, PubMed: 29052241

Günther, Fritz, Tri Nguyen, Lu Chen, Carolin Dudschig, Barbara Kaup, and Arthur M. Glenberg. 2020. Immediate sensorimotor grounding of novel concepts learned from language alone. *Journal of Memory and Language*, 115:104172. https://doi.org/10.1016/j.jml.2020.104172

Günther, F., L. Rinaldi, and M. Marelli. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14:1006–1033. https://doi.org/10.1177/1745691619861372, PubMed: 31505121

Harris, Z. 1954. Distributional structure. *Word*, 10:146–162. https://doi.org/10.1080/00437956.1954.11659520

Haslett, David A. and Zhenguang G. Cai. 2023. Systematic mappings of sound to meaning: A theoretical review. *Psychonomic Bulletin & Review*, 31(2):627–648. https://doi.org/10.3758/s13423-023-02395-y, PubMed: 37803232

Hendrix, Peter and Ching Chu Sun. 2021. A word or two about nonwords: Frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(1):157–183. https://doi.org/10.1037/xlm0000819, PubMed: 31999159

Hutchison, Keith A., David A. Balota, James H. Neely, Michael J. Cortese, Emily R. Cohen-Shikora, Chi-Shing Tse, Melvin J. Yap, Jesse J. Bengson, Dale Niemeyer, and Erin Buchanan. 2013. The semantic priming project. *Behavior Research Methods*, 45:1099–1114. https://doi.org/10.3758/s13428-012-0304-z, PubMed: 23344737

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jones, Michael N., Jon Willits, and Simon Dennis. 2015. Models of Semantic Memory. Oxford University Press. 232–254. https://doi.org/10.1093/oxfordhb/9780199957996.013.11

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Keuleers, Emmanuel and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42:627–633. https://doi.org/10.3758/BRM.42.3.627, PubMed: 20805584

Kumar, Abhilasha A., Mark Steyvers, and David A. Balota. 2021. Semantic memory search and retrieval in a novel cooperative word game: A comparison of associative and distributional semantic models. *Cognitive Science*, 45(10):e13053. https://doi.org/10.1111/cogs.13053, PubMed: 34622483

Lake, Brenden M. and Gregory L. Murphy. 2023. Word meaning in minds and machines. *Psychological Review*, 130(2):401. https://doi.org/10.1037/rev0000297, PubMed: 34292021

Landauer, T. K. and S. T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lazaridou, Angeliki, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, 41:677–705. https://doi.org/10.1111/cogs.12481, PubMed: 28323353

Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Lenci, Alessandro, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, 56(4):1269–1313. https://doi.org/10.1007/s10579-021-09575-z

Li, Bohan, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130. https://doi.org/10.18653/v1/2020.emnlp-main.733

Linzen, Tal. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217. https://doi.org/10.18653/v1/2020.acl-main.465

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marelli, Marco and Simona Amenta. 2018. A database of orthography-semantics

consistency (OSC) estimates for 15,017 English words.*Behavior Research Methods*, 50:1482–1495. https://doi.org/10.3758/s13428-018-1017-8, PubMed: 29372490

Marelli, Marco and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, 122(3):485–515. https://doi.org/10.1037/a0039267, PubMed: 26120909

Monaghan, Padraic, Richard C. Shillcock, Morten H. Christiansen, and Simon Kirby. 2014. How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130299. https://doi.org/10.1098/rstb.2013.0299, PubMed: 25092667

Murphy, Gregory. 2004.*The Big Book of Concepts*. MIT Press.

Perniss, Pamela and Gabriella Vigliocco. 2014. The bridge of iconicity: From a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130300. https://doi.org/10.1098/rstb.2013.0300, PubMed: 25092668

Plaut, David C. and Laura M. Gonnerman. 2000. Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, 15(4–5):445–485. https://doi.org/10.1080/01690960050119661

Pugacheva, Vasilisa and Fritz Günther. 2024. Lexical choice and word formation in a taboo game paradigm. *Journal of Memory and Language*, 135:104477. https://doi.org/10.1016/j.jml.2023.104477

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Romberg, Alexa R. and Jenny R. Saffran. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):906–914. https://doi.org/10.1002/wcs.78, PubMed: 21666883

Sabbatino, Valentino, Enrica Troiano, Antje Schweitzer, and Roman Klinger. 2022. "splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words. *arXiv preprint arXiv:2202.12132*.

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118. https://doi.org/10.1073/pnas.2105646118, PubMed: 34737231

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162

Shain, Cory, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121. https://doi.org/10.1073/pnas.2307876121, PubMed: 38422017

Shannon, Claude Elwood. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55. https://doi.org/10.1145/584091.584093

Stevens, Patience and David C. Plaut. 2022. From decomposition to distributed theories of morphological processing in reading. *Psychonomic Bulletin & Review*, 29(5):1673–1702. https://doi.org/10.3758/s13423-022-02086-0, PubMed: 35595965

Sulpizio, Simone, Eleonora Pennucci, and Remo Job. 2021. The impact of emotional content on pseudoword recognition. *Psychological Research*, 85:2980–2996. https://doi.org/10.1007/s00426-020-01454-6, PubMed: 33337511

Taft, Marcus and Kenneth I Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14(6):638–647. https://doi.org/10.1016/S0022-5371(75)80051-X

Thompson, Arthur L., May Pik Yu Chan, Ping Hei Yeung, and Youngah Do. 2022. Structural markedness and depiction: The case of lower sequential predictability in cantonese ideophones. *The Mental Lexicon*, 17(2):300–324. https://doi.org/10.1075/ml.21016.tho

Toneva, Mariya and Leila Wehbe. 2019. Interpreting and improving

natural-language processing (in machines) with natural language-processing (in the brain). In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 14954–14964.

Tuckute, Greta, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. 2024. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, *8*(3):544–561. https://doi.org/10.1038/s41562-023-01783-7, PubMed: 38172630

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Vidal, Yamil, Eva Vivani, Davide Zoccolan, and Davide Crepaldi. 2021. A general-purpose mechanism of visual feature association in visual word identification and beyond. *Current Biology*, 31(6):1261–1267. https://doi.org/10.1016/j.cub.2020.12.017, PubMed: 33417881

Warstadt, Alex and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*. CRC Press, pages 17–60. https://doi.org/10.1201/9781003205388-2

Wilcox, Ethan, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wood, Simon N. 2006.*Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420010404