

Byte-pair Encoding Captures Aspects of Word Meaning that Distributional Semantics Overlooks

David A. Haslett

City University
of Hong Kong

dhaslett@cityu.edu.hk

Antoni B. Chan

City University
of Hong Kong

abchan@cityu.edu.hk

Janet H. Hsiao

The Hong Kong University
of Science and Technology

jhhhsiao@ust.hk

Abstract

Large language models learn word meanings through distributional semantics (i.e., patterns in text), and distribution does not always capture information about taxonomy and function, which are essential to human representations of word meaning. However, sound and spelling often conveys such information, and language models decompose many words into subword tokens, which may identify category markers. For example, GPT-5 segments *fluorine* and *bromine* into *fluor* + *ine* and *brom* + *ine*, which share the category marker *ine*. We provide evidence that, across 18 languages, shared tokens predict greater semantic priming effects in humans (e.g., people recognize *bromine* faster after reading *fluorine*), beyond what distributional similarity explains. Furthermore, across seven languages, shared tokens boost semantic priming beyond what shared morphemes explain. This suggests that subword tokens could help language models arrive at human-like representations of word meanings. However, we also provide evidence that the trend towards larger vocabularies of tokens obscures category markers, and even when language models have access to subword tokens, they underestimate the semantic relatedness of words that share tokens.

1 Introduction

Large language models (LLMs) are a type of distributional semantic model, which is to say, they infer meaning from patterns in text. Distribution sometimes overlooks aspects of word meaning such as taxonomy and sensory experience (e.g., Andrews et al. 2009), which may hinder concept alignment, i.e., the goal of having words mean the same things to LLMs as to people (Rane et al., 2024). However, taxonomic and sensory information is often implied by word form (e.g.,

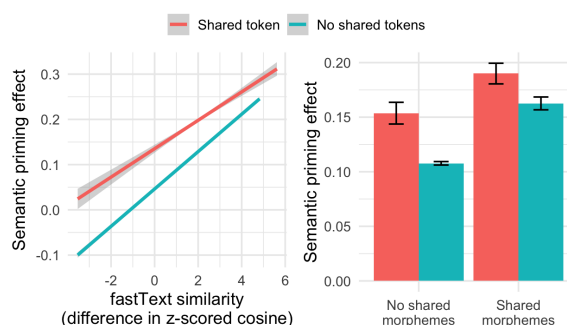


Figure 1: Semantic priming as a function of distributional semantic similarity (across 18 languages) and morphological overlap (across seven languages). Semantic priming is the difference in how quickly people recognize a target word (e.g., *abdominal*) following a semantically related prime (*spinal*) compared to an unrelated prime (*chemical*). Priming is greater when primes and targets occur in more similar contexts (left) or share morphemes (right). Crucially, according to twelve tokenizers, shared final tokens (as in *sp* + *inal* and *abdom* + *inal*) boost priming beyond what distributional similarity or morphological overlap predicts.

Kelly 1992), and LLMs segment many words into chunks (i.e., subword tokens), which could convey that information. For example, *fluorine* and *bromine* are members of a family of chemicals marked by *ine*, and when processed by GPT-5, *fluorine* and *bromine* share the subword token *ine*. In this study, we use semantic priming effects in human word processing, from Buchanan et al. 2026, to differentiate representations in distributional semantic models from multifaceted word meanings, across a variety of languages, and we investigate whether subword tokens identify semantic features that supplement the meanings implied by patterns in text.

1.1 Distributional semantics

Distributional semantic models, such as Latent Semantic Analysis (Landauer and Dumais, 1997) and word2vec (Mikolov et al., 2013), repre-

sent word meanings with embeddings, which are strings of numbers that situate words in semantic space. Embeddings are derived from patterns in text, based on the assumption that words which occur in similar contexts have similar meanings (Firth, 1957; Harris, 1954). For example, *abdominal* and *hernia* co-occur in many sentences, which can be quantified as the cosine similarity of their embeddings. LLMs are sometimes called “contextualized” distributional semantic models (e.g., Lenci et al. 2022), because whereas “static” models such as word2vec learn a single, stable embedding for each word, LLMs tailor embeddings to contexts (Devlin, 2018; Peters et al., 2018; Radford et al., 2019). For example, the representation of *abdominal* will differ in contexts that refer to exercise versus medicine. The cosine similarity of embeddings correlates with semantic priming effects in lexical decisions by humans: People recognize a target word faster after reading a semantically related prime word (Meyer and Schvaneveldt, 1971), and priming effects vary as a function of the proximity of primes to targets in distributional semantic space, for both static models (Auguste et al., 2017; Gatti et al., 2023; Jones et al., 2006; Lund and Burgess, 1996; Mandera et al., 2017) and LLMs (Cassani et al., 2023). Günther et al. 2016 conclude that distributional semantic models offer a “cognitively plausible measure of meaning similarities” (p. 11).

However, the meanings implied by distribution may be incomplete. Andrews et al. 2009 argue that people learn word meanings through two paths, distribution and sensory experience, and Hill et al. 2015 emphasize the difference between words that occur in similar contexts, such as *abdominal* and *hernia*, and words with similar functions or referents, such as *abdominal* and *spinal*. This distinction is evident in the semantic priming task, where the facilitatory effects of shared semantic features can be dissociated from the effects of co-occurrence (Hutchison, 2003; Lucas, 2000) and where morphology has larger effects than distribution (e.g., people will recognize *abdominal* faster after reading *spinal* than after *hernia*, other things being equal; Feldman 2000; Rastle et al. 2000). Furthermore, the word association task (e.g., asking participants to say the first word that comes to mind when they hear *abdominal*) provides evidence that meanings in the human mind gravitate towards function and reference: As children

mature, they shift away from co-occurring associates in favour of taxonomically related associates (Brown and Berko, 1960; Entwisle, 1966). Static distributional semantic models therefore fail to capture much of the variance in human associations (De Deyne et al., 2016, 2021), and LLM-generated associates differ systematically from human-generated associates (Xiao et al., 2025).

Despite these challenges, distributional semantic models have made progress at capturing taxonomy (Amenta et al., 2020; Erk, 2016; Marelli and Baroni, 2015), thanks in part to hyperparameters—most notably, leveraging the fact that adjacent words imply part of speech (e.g., words immediately following “the” or “his” are often nouns, whereas words immediately following “must” and “to” are more often verbs), while distantly co-occurring words imply thematic similarity (Bullinaria and Levy, 2007; Redington et al., 1998)—but thanks also to the information conveyed by word form. For example, across many languages, nouns and verbs differ systematically in sound and spelling (Kelly, 1992), as do words with abstract versus sensory meanings (Reilly et al., 2017; de Zubicaray et al., 2023). For reviews, see Dingemanse et al. 2015 and Haslett and Cai 2024. In some cases, these patterns in form–meaning mapping are subtle, but in many other cases, they are overt (Kallens and Christiansen, 2020; Pimentel et al., 2019). For example, Korean words that end with *다* tend to be infinitive verbs, Chinese words that end with *子* tend to be concrete nouns, and English words that end with *-ness* tend to be abstract nouns. In this paper, we refer to subword constituents that imply taxonomy or function as category markers. FastText models can identify such category markers by learning the distributions of subword strings (Bojanowski et al., 2017), which enables such models to situate low-frequency words and out-of-vocabulary words in semantic space (de Varda et al. 2024; Gatti et al. 2023; Hendrix and Sun 2021; cf. Cassani et al. 2020).

Still, whether subword distribution suffices to imply human-like meanings is an open question. Along with practical challenges—namely the need to encounter words and subwords across many contexts, which poses problems for both static models (Bullinaria and Levy, 2007; Sahlgren and Lenci, 2016) and LLMs (Algayres et al., 2025; Schick and Schütze, 2020)—it is unclear in princi-

ple what role inductive biases might play. People and languages have co-evolved such that different parts of words convey different sorts of information (e.g., in many languages, category markers tend to be suffixes), so to infer taxonomy, it helps to know where to look (Cutler et al., 1985; St. Clair et al., 2009; Sun and Poeppel, 2023). Furthermore, the sensory information implied by word form is underdetermined (e.g., the vowel /i/ could indicate size, speed, taste, or nothing at all; Winter et al. 2019), and many of those sound symbolic form–meaning mappings occur across a range of unrelated languages (Blasi et al., 2016; Ćwiek et al., 2022), which suggests that subword strings might cue pre-existing associations in people without their distribution sufficing to teach those associations to LLMs.

1.2 Tokenization

The way that LLMs represent words further complicates their ability to infer human-like meanings. LLMs segment text into chunks called tokens, and almost all LLMs use a compression algorithm called byte-pair encoding (Gage, 1994; Sennrich et al., 2016). Byte-pair encoding is determined entirely by orthographic regularities, and English is predominant on the internet, so common English words tend to be single tokens in state-of-the-art LLMs, while most other words are segmented into subword tokens. For example, *interdisciplinary* is represented by GPT-5 as a single token, whereas the rare word *intradisciplinary* is segmented into three tokens (`_intr` + `ad` + `isciplinary`), and the French translation *interdisciplinaire* is segmented into four tokens (`_inter` + `disc` + `ipl` + `inaire`). (The underscores indicate whitespace incorporated into a word’s first token.) Unlike fastText, which averages the embeddings of all the n-grams in a word (e.g., all the 4-letter strings in *interdisciplinary*), LLMs have access only to the subword tokens identified by byte-pair encoding, so a single-token word such as *interdisciplinary* obscures *inter*, while the tokenization `_intr` + `ad` + `isciplinary` obscures *disciplin*.

As `_intr` + `ad` + `isciplinary` illustrates, byte-pair encoding often produces tokens that diverge from morphemes (i.e., the smallest meaningful constituents in languages; Bostrom and Durrett 2020; Church 2020; Hofmann et al. 2021), and several studies report that non-morphological tokens hinder performance on downstream tasks,

such as categorizing words or pairing them with definitions (e.g., Batsuren et al. 2024; Hofmann et al. 2022; Jabbar 2023; Mager et al. 2022). However, other studies report that LLMs perform equally well regardless of whether tokens correspond to morphemes (Arnett et al., 2024a, 2025; Gutierrez et al., 2023). One explanation for these conflicting results might be that non-morphological tokenization still captures semantic information, such as the relationships reflected in the forms of *height* and *weight* or *junior* and *senior*. Haslett and Cai 2025 found that words which share tokens are more distributionally similar than expected by chance even when they do not share morphemes, and Gutierrez-Vasques et al. 2023 found that byte-pair encoding quickly comes to reflect linguistic typology, concluding that tokens “might not correspond to usual morphological analyses, but they are structural elements” (p. 944). This view of linguistic structure is consistent with non-compositional theories of morphology, which eschew morphemes as an intermediate level of representations; they instead posit that morphology emerges from systematic form–meaning mappings (e.g., Baayen et al. 2011, 2019; Plaut and Gonnerman 2000). In this way, non-compositional morphology accommodates “sub-morphemic” category markers (e.g., Bergen 2004; Haslett et al. 2026; Kwon and Round 2015), and it provides a theoretical grounding for the possibility that tokenizers identify meaningful strings and explain word processing effects without mapping onto morphemes as traditionally construed.

1.3 Semantic priming and form priming

In this paper, we use data from the Semantic Priming Across Many Languages Project (SPAM-L; Buchanan et al. 2026) to investigate whether tokens capture aspects of word meaning that distributional semantic models overlook—that is, whether primes which share tokens with targets, as in some representations of *spinal* and *abdominal*, speed up visual word recognition, beyond what is predicted by fastText and word2vec. Shared tokens entail orthographic overlap (although this is complicated in languages that do not use the Latin alphabet), and orthographic overlaps prime too (e.g., Forster and Davis 1984; Tanenhaus et al. 1980), so any boost from shared tokens might be better attributed to form than meaning. However, form priming requires very brief exposure to

prime words, typically less than 70 ms; at longer delays, primes that overlap in form but not meaning instead inhibit target recognition (e.g., [Feldman 2000](#); [Forster and Veres 1998](#); [Grainger et al. 1991](#); [Humphreys et al. 1987](#); [Rastle et al. 2000](#)). In SPAM-L, participants made lexical decisions about primes, with up to 3,000 ms per decision, followed by a 500 ms gap until the target. This suggests that any persistent boost from shared tokens will reflect semantic relatedness.

On the other hand, the roles of form and meaning are not easily disentangled. Incorporating formal relationships into semantic networks improves models of language acquisition, aphasia, and word meaning access ([Castro et al., 2020](#); [Kumar et al., 2022](#); [Stella et al., 2017](#)), and word form affects interpretation and motivates semantic change ([Ćwiek et al., 2022](#); [Haslett and Cai, 2023](#); [Lupyan and Casasanto, 2015](#)). Indeed, form priming reflects the co-evolution of people and language: Word endings prime, whereas onsets compete (e.g., [Norris et al. 2002](#); [Radeau et al. 1998](#)), much like how word endings help people lump words into categories, whereas onsets differentiate meanings ([Monaghan et al. 2011](#); [St. Clair et al. 2009](#); though these trends vary across languages; [Hammarström 2021](#)). We are therefore open to the possibility that any boost from shared tokens will blur the line between semantic and form priming, but we maintain that, because primes are presented overtly and with long lags until targets, these would be the sorts of formal effects that are salient to how people retrieve and represent word meanings.

1.4 The present study

We hypothesize that byte-pair encoding identifies constituents that supplement distributional semantics, thereby facilitating concept alignment. In Analysis 1, we regress semantic priming effects on the distributional semantic similarity of prime–target pairs in 18 languages, along with other lexical variables, and we investigate whether shared tokens explain additional variance, according to 12 tokenizers used by 12 LLMs. In Analysis 2, we investigate whether shared tokens explain semantic priming effects independently of shared morphemes, in seven languages. In Analysis 3, we train tokenizers of increasing vocabulary size on English data, and we investigate which vocabulary best explains semantic priming effects. In Analy-

Target	Related prime	Unrelated prime
English		
entrance	exit	soup
_entrance	_exit	_soup
Hungarian (unrelated = “commander”)		
bejárat	kijárat	parancsnok
_be+já+rat	_kij+á+rat	_par+anc+sn+ok
Korean (unrelated = “personal”)		
입구	출구	사적인
_입+구	_출+구	_사+적인
Serbian (unrelated = “bridge”)		
улаз	излаз	мост
_у+ла+з	_из+ла+з	_м+ост
Turkish (unrelated = “soup”)		
giriş	çıkış	çorba
_gi+ri+ş	_çı+k+ış	_ç+or+ba

Table 1: Examples of one semantic priming item in five languages, as tokenized by Kimi K2. The targets and related primes have the same meanings across languages (entrance/exit), but the unrelated primes vary.

sis 4, we compare semantic priming effects to how five state-of-the-art LLMs rate semantic relatedness and categorize words, to investigate whether LLMs take advantage of the information available in tokens when representing word meanings.

By using semantic priming effects from human participants ([Buchanan et al., 2026](#)), this study builds on [Haslett and Cai 2025](#), who focus on the distributional semantic information available in LLM tokens and so overlook the role that word form may play in conveying taxonomic information not captured by distributional semantic models. By using semantic priming effects, we also build on [Yang et al. 2022](#) and [Beinborn and Pinter 2023](#), who show that LLM tokens predict human word processing effects in reading and in lexical decisions, respectively. The stimuli for semantic priming effects from [Buchanan et al. 2026](#) were manipulated within items, which better accounts for confounding variables (e.g., high-frequency words comprise fewer tokens and are read faster): In each analysis, our dependent variable is not how quickly people recognize a target but, instead, how much *more* quickly people recognize a target following a related prime compared to an unrelated prime. Finally, our approach is distinct from [Batsuren et al. 2024](#) and others because, rather than evaluate the downstream impact of tokenization on LLM performance (though we do consider that in Analysis 4), we emphasize the meaningfulness of

tokens. We are interested in the information that byte-pair encoding supplies to LLMs and the big-picture implications for concept alignment, not the performance of LLMs on any particular task.

All materials are available at <https://osf.io/dnwqy>. Analyses were conducted in Python and R (R Core Team, 2025).

2 Analysis 1

If tokens identify semantic features, priming effects should be larger when prime–target pairs share tokens, and if the features are distinct from distributional semantics, shared tokens should explain additional variance in priming.

2.1 Materials

Semantic priming. SPAM-L comprises 1,000 target words from each of 18 languages, paired with a related prime and an unrelated prime. (Buchanan et al. 2026 report semantic priming effects for 19 languages, but because the stimuli for Brazilian and European Portuguese overlap substantially, we treat them as a single language.) For example, the Hungarian target *bejárat* (“entrance”) was paired with the related prime *kijárat* (“exit”) and the unrelated prime *parancsnok* (“commander”). See Table 1. Buchanan et al. 2026 filtered for primes and targets that are among the 10,000 most frequent concepts in each languages’ Open-Subtitles corpora (Lison and Tiedemann, 2016), and using subs2vec embeddings (Van Paridon and Thompson, 2021), they selected related primes by identifying words that are among the five most similar to a given target in all candidate languages. They selected unrelated primes by randomly shuffling words in each language until all pairs were relatively low in similarity. In each trial, a participant made a lexical decision about a prime word (e.g., deciding whether the string of letters *bejárat* forms a word), followed by a lexical decision about the target. The priming effect is the difference in response time (z-scored for each participant) between the related and unrelated condition for each target word. For example, Hungarian participants recognized *bejárat* faster after reading *kijárat* (516 ms, on average) than after reading *parancsnok* (604 ms).

Shared tokens. We tokenized each prime and target word according to 12 LLMs, four made in the USA (Gemini 3, GPT-4.1, Llama 4, Phi-4; Gemma Team et al. 2024; Singh et al. 2025; Ad-

cock et al. 2026; Abdin et al. 2024), four made in China (DeepSeek-V3.1, GLM-4.7, Kimi K2, Qwen3; Liu et al. 2024; GLM Team et al. 2024; Kimi Team et al. 2025; Yang et al. 2025), and four made in Europe (Apertus, EuroLLM, Mistral Large 3, and Teuken; Apertus et al. 2025; Martins et al. 2025; Karamcheti et al. 2021; Ali et al. 2024). Five were chosen because they are high-performing models that provide next-token probabilities via API and so were used for Analysis 4; the other seven were chosen to provide geographic diversity, which might affect multilingual tokenization. For 16 of the languages, we included leading whitespace when tokenizing words, as they would typically be encountered, excepting Japanese and Mandarin, which do not separate words with whitespace. We identified whether prime–target pairs share any tokens (e.g., according to Kimi, *kijárat* and *bejárat* share *rat*), and we coded items as sharing a token when the target shares a token with the related prime but not the unrelated prime. Note that this analysis does not involve LLM representations or output (though see Analysis 4); we simply determined whether pairs of words share tokens.

Distributional semantics. We measured the cosine similarity of language-specific embeddings (e.g., the tendency for the Hungarian words *bejárat* and *kijárat* to occur in similar contexts). The subs2vec models used by Buchanan et al. 2026 do not generate embeddings for out-of-vocabulary words, and Buchanan et al. 2026 tweaked the forms expressing some concepts in some languages, resulting in 1,190 targets (6.8% of items) with a missing subs2vec embedding in either the related or unrelated condition. We therefore used fastText embeddings from Grave et al. 2018, which handle out-of-vocabulary words and so have no missing values. These fastText embeddings from reflect the distribution of whole words as well as subword 5-grams (e.g., *bejárat* comprises *<beja, bejar, ...arat>*; *<* and *>* indicate word boundaries), so prime–target pairs that share 4- or 5-letter tokens will have more similar fastText embeddings, making for a more severe test of whether shared tokens boost priming than word2vec embeddings would (i.e., models which ignore subword distribution). However, the subs2vec embeddings from Van Paridon and Thompson 2021 incorporate more fine-grained n-grams than the embeddings from Grave et al. 2018, ranging from 3-

grams to 6-grams, which can align with shorter tokens, so we reproduce our analyses using those embeddings. For the sake of comparison, we also use word2vec embeddings, from Al-Rfou et al. 2013, which do not incorporate subword distribution. To estimate priming effects, we z-scored cosine similarity within each language (since different languages have different semantic spaces) and then subtracted the similarity of unrelated prime–target pairs (e.g., -0.63 for *bej́arat* and *paranscnok*) from related pairs (1.26 for *bej́arat* and *ki-j́arat*).

Lexical variables. Priming effects, which measure the difference in response time for the same target in the related versus unrelated conditions, better control for lexical variables such as word frequency than simply measuring lexical decision response time. However, word frequency affects priming (e.g., Becker 1979), and frequency correlates with the number of tokens per word in byte-pair encoding, which could confound our findings. We therefore included several co-predictors in our linear regression models (described below): logged word frequency (from Speer 2022), orthographic length, and number of tokens per word, for targets, related primes, and unrelated primes.

Language variables. We measured the prevalence of each language on the internet, as reported by the Observatory of Linguistic and Cultural Diversity (Pimienta, 2022). English and Chinese are the most common (20.08% and 19.04% of the text on the internet, respectively), while Hungarian and Danish are the least common in our sample (0.15% and 0.10%). We also recorded the script used to write each language. Twelve of the languages use the Latin alphabet; Russian and Serbian use Cyrillic; Mandarin uses Chinese; Japanese uses Kanji, Hiragana, and Katakana; Korean uses Hangul; and Greek uses Greek.

2.2 Analysis plan

First, we fit 30 pairs of linear regression models, one per tokenizer (each with data from 18 languages) and one per language (each with data from 12 tokenizers). We regressed semantic priming effects on fastText similarity, word frequency, orthographic length, and tokens per word. Then, we added the shared token factor as a predictor. If tokens identify semantic information beyond what is captured by distributional semantics, the shared token factor should increase adjusted R^2 . We used

ANOVA to test significance.

Second, we compared the priming effects of shared tokens that occur at the beginning versus end of a word. Category markers occur at the ends of words more often than the beginnings of words (Cutler et al. 1985; Greenberg 1957; though see Hammarström 2021 for evidence that this tendency has been overestimated), so if tokens convey taxonomic information, priming effects may be larger for words that share final tokens, such as *kij́arat* and *bej́arat*. To test this possibility, we fit the same regression models as above, but including the interaction of whether prime–target pairs share tokens with whether those overlaps occur at the beginnings or ends of words. We treatment coded the token and position factors such that the intercept predicts priming for pairs that share no tokens, the simple effect of shared tokens indicates whether priming is greater for pairs that share initial tokens, and the interaction of the token and position factors indicates whether predicted priming is greater when pairs share final rather than initial tokens. We again fit 30 regression models.

Finally, we investigated whether the semantic information captured by tokens varies by language and tokenizer. For each tokenizer, we fit a regression model with fastText similarity and the lexical variables, and we added interactions of the shared token factor with logged language prevalence and script type, treatment coded such that the simple effect of shared tokens refers to the 12 languages that use the Latin alphabet and the interaction with script type indicates the change in the languages that use different scripts. For each language, we added interactions of the shared token factor with tokenizer vocabulary size (ranging from 100,352 tokens for Phi-4 to 262,144 for Gemini 3) and with tokenizer country of origin, treatment coded such that the simple effect of shared tokens refers to tokenizers made in the USA and the interactions indicate how the effect changes for tokenizers made in China or Europe.

2.3 Results

As reported in Table 2 and the appendix, adding the shared token factor explains additional variance in semantic priming, beyond fastText similarity and the lexical variables, for all 12 tokenizers and for 13 of the 18 languages. The exceptions are German and Polish, where the effect is far from significant and trends in the opposite direc-

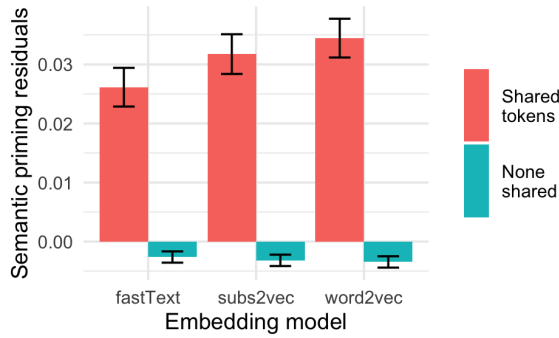


Figure 2: Semantic priming residuals, after regressing on word frequency, word length, tokens per word, and either fastText similarity, subs2vec similarity, or word2vec similarity. For the sake of comparison in this illustration, we include only items with embeddings for all three embedding models. In all cases, shared tokens capture variance in semantic priming unexplained by distributional semantic similarity. Larger effects of shared tokens imply less information captured by embeddings (e.g., for word2vec relative to fastText, since word2vec does not incorporate subword distribution).

tion, and Hungarian, Japanese, and Korean, where the additional variance is only “marginally” significant at the .05 level. We explore these language-level differences in the third set of regression models, below. When instead using subs2vec embeddings in the base models, the effects are very similar to fastText: highly significant for all 12 tokenizers, negligible for German and Polish, marginally significant for Turkish and Hungarian ($p = .064$ and $.096$, respectively), and significant for the 14 other languages. When using word2vec embeddings, which do not capture subword distribution, the effects are larger than for fastText or subs2vec: The additional variance explained by shared tokens is highly significant for all 12 tokenizers and all 18 languages. As illustrated in Figure 2 and reported online, the fastText embeddings from [Grave et al. 2018](#) capture more variance in semantic priming than the subs2vec and word2vec embeddings.

As illustrated in Figure 3 and reported in full in the appendix, prime–target pairs that share final tokens lead to more priming than pairs that share initial tokens. The interaction of whether prime–target pairs share tokens and whether they overlap at the beginnings or ends of words is highly significant for all 12 tokenizers (the “least significant” p -value is .0001, for Llama 4). The by-language effects are more mixed, but the trend holds: The interaction fails to reach significance only for Italian

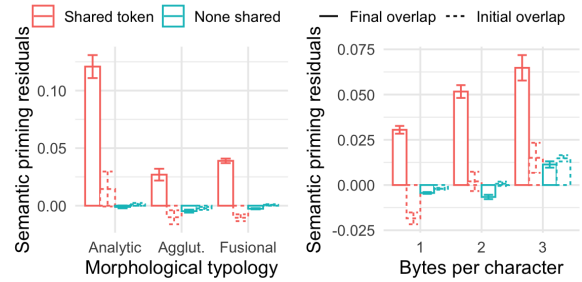


Figure 3: Semantic priming residuals (after regressing on distributional semantic similarity and lexical variables) as a function of whether prime–target pairs share tokens. Mandarin and English are analytic / isolating; Hungarian, Japanese, Korean, and Turkish are agglutinative; the rest are fusional. Japanese, Korean, and Mandarin use scripts with three bytes per character (when represented by UTF-8 encodings); Greek, Russian, and Serbian use two bytes per character; the rest use the Latin alphabet, with one byte per character (notwithstanding two-byte diacritics).

(where the interaction is marginally significant, and the simple effect of shared initial tokens is highly significant), for English and Serbian (where the simple effects are highly significant), and for Japanese and Korean (where the simple effects are not significant, either).

As illustrated in Figure 4 and reported in full in the appendix, the effect of shared tokens tends to be greater in languages that are more common on the internet. The interaction of the shared token factor with language prevalence is significantly positive for 10 of the 12 tokenizers, excepting Kimi and Teuken ($p = .473$ and $.262$, respectively). The interaction of the shared token factor with script type is significantly positive for Gemini, Llama, Apertus, and Mistral, indicating more priming in languages that use scripts other than the Latin alphabet, but only for those four tokenizers. In the by-language analyses, the interaction of the shared token factor with tokenizer region of origin is not significant in any case, for neither Chinese nor European tokenizers compared to American tokenizers (the “most significant” p -value is .187 for Italian, with a positive interaction for European tokenizers), and the interaction with tokenizer vocabulary size approaches significance only for Romanian ($p = .066$). We therefore provide no clear evidence that some types of tokenizers capture more semantic information than others.

Tokenizer	Base	Tokens	Δ	p
Czech	.0844	.0847	.0003	.029
Danish	.0341	.0433	.0093	2e-16
English	.137	.1385	.0015	3e-6
French	.0679	.0705	.0026	1e-8
German	.0956	.0955	-.0001	.946
Greek	.1409	.1437	.0028	.002
Hungarian	.0596	.0598	.0002	.098
Italian	.0643	.0664	.0021	2e-7
Japanese	.1134	.1138	.0004	.065
Korean	.0526	.0532	.0006	.055
Mandarin	.1286	.1308	.0022	1e-7
Polish	.0818	.0817	-.0001	.868
Portuguese	.0762	.0863	.0100	2e-16
Romanian	.0911	.0922	.0012	8e-4
Russian	.078	.0824	.0044	1e-13
Serbian	.0895	.0918	.0023	1e-6
Spanish	.0674	.0735	.0062	2e-16
Turkish	.0929	.0933	.0004	.027

Table 2: Adjusted R^2 from linear regression models predicting semantic priming, without (Base) and with (Tokens) the shared token factor.

2.4 Discussion

When prime–target pairs share tokens, the semantic priming effect is larger than predicted by distributional semantic similarity, word frequency, and word length. This suggests that byte-pair encoding identifies constituents that play a role in how people represent word meanings. The boost in priming is especially pronounced in word-final tokens: For all 12 tokenizers, word-final tokens significantly increase priming, while word-initial tokens do not. For 10 of 12 tokenizers, the boost is larger in languages that make up more of the text on the internet, suggesting that exposure to language-specific orthographic regularities helps byte-pair encoding identify meaningful constituents.

3 Analysis 2

Why do shared tokens correspond to larger semantic priming effects? Byte-pair encoding does not reliably identify morphemes (e.g., Bostrom and Durrett 2020), but the boost might nevertheless stem from the morphemes it does identify. This analysis therefore investigates whether shared tokens explain additional variance in semantic priming after accounting for shared morphemes.

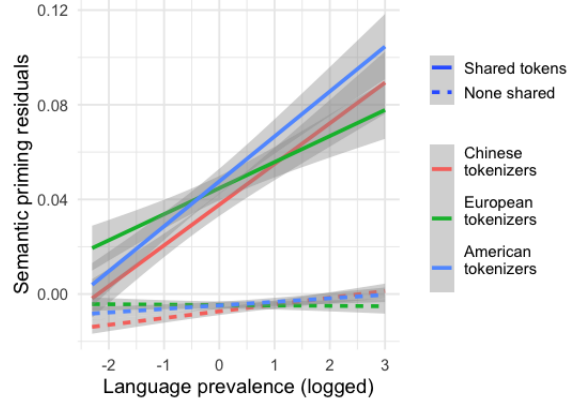


Figure 4: Semantic priming residuals (after regressing on distributional semantic similarity and lexical variables) as a function of the logged percentage of internet text that a language makes up.

3.1 Materials

To identify shared final morphemes, we used human-annotated word-level morpheme segmentations from the Sigmorphon 2022 Shared Task (Batsuren et al., 2022). Segmentations are available for seven of the 18 languages from Analysis 1: Czech, English, French, Hungarian, Italian, Russian, and Spanish. Examples of English items that share final morphemes are *marketing* and *advertising*, which share the inflection -ing; *senator* and *governor*, which share the derivation -or; and *Batman* and *Superman*, which share the free morpheme -man. Examples of French items that share final morphemes are *tourner* (“to turn”) and *pivoter* (“to pivot”), which share the (infinitive) inflection -er; *complètement* (“completely”) and *absolument* (“absolutely”), which share the derivation -ment; and *nom* (“name”) and *surnom* (“surname”), which share the free morpheme -nom. 10.5% of the items share a final morpheme, ranging from 6.4% in Spanish to 35.1% in Czech. Of the items that share final morphemes, 29.0% also share final tokens, ranging from 6.6% in English to 54.0% in Russian, and ranging from 18.3% for Llama 4 to 44.4% for Phi-4. Examples of English items that share both final morphemes and final tokens, according to EuroLLM, are *briefcase* and *suitcase*, which share the token -case, and *dusty* and *rusty*, which share the token -y.

One instructive example of an English item that does not share a final morpheme but does share a final token (according to EuroLLM) is *surname* and *nickname*. The Sigmorphon annotations segment *surname* into *sur* + *name*, but

Tokenizer	Base	Tokens	Δ	p
DeepSeek	.0394	.0407	.0013	.009
GLM-4.7	.0385	.0408	.0022	.0008
Kimi K2	.0400	.0432	.0032	8e-5
Qwen3	.0393	.0411	.0018	.002
Gemini 3	.0390	.0399	.0009	.023
GPT-4.1	.0379	.0396	.0017	.003
Llama 4	.0385	.0398	.0013	.009
Phi-4	.0391	.0410	.0019	.002
Apertus	.0377	.0396	.0019	.002
EuroLLM	.0368	.0398	.0030	.0001
Mistral	.0377	.0396	.0019	.002
Teuken	.0382	.0404	.0022	.001

Table 3: Adjusted R^2 from linear regression models predicting semantic priming, without (Base) or with (Tokens) the shared token factor. Adding whether prime–target pairs share final tokens improves fit, beyond shared morphemes and other lexical variables.

they treat *nickname* as holistic. In contrast, MorphoLex, a database of English morphological segmentations (Sánchez-Gutiérrez et al., 2018), decomposes *nickname*. Similarly, Sigmorphon treats *microscope* as holistic but *telescope* as compositional, whereas MorphoLex decomposes both. As we show in the appendix, there are many discrepancies between Sigmorphon and MorphoLex (though we find the same pattern of effects when using MorphoLex in Analysis 3). In other cases, shared final tokens unambiguously differ from morphemes—e.g., *sergeant* and *lieutenant*, *espionage* and *sabotage*—but as touched on in the introduction, whether such etymological or analogical relationships count as morphological depends on your theoretical commitments (e.g., Baayen et al. 2011). This analysis therefore tests the hypothesis that traditional morphological decompositions are ambiguous and incomplete and that byte-pair encoding might be well suited to capture semantic information which even human annotations miss.

3.2 Analysis plan

First, we fit 19 pairs of linear regression models, 12 by tokenizer and seven by language, regressing semantic priming on whether related prime–target pairs share final morphemes, plus prime and target frequency, length, and number of tokens. As in Analysis 1, we added whether related prime–target pairs share final tokens, and we used ANOVA to compare adjusted R^2 with versus without that fac-

tor. If tokens identify semantic information beyond what is captured by Sigmorphon, the shared token factor should increase adjusted R^2 .

Second, because morphological annotations are ambiguous (i.e., decomposition differs across resources, and definitions of morphemes acknowledge grey areas), we compared tokens to other shared strings, across all 18 languages, ignoring morphology. We identified the longest final overlap in related prime–target pairs, up to 10 characters in length (because that is the longest overlapping token in SPAM-L, namely when using Teuken to tokenize the Hungarian words *gyilkosság*, “murder”, and *öngyilkosság*, “suicide”), and at each length, we coded whether the prime and target share that final string and whether they end with the same token. For example, the Czech words *reaktor* (“reactor”) and *generátor* (“generator”) end with the same three letters and end with different GPT-4.1 tokens, whereas *zadní* (“rear”) and *přední* (“front”) end with the same three letters and the same GPT-4.1 token (ní). (To avoid, e.g., counting four-letter overlaps as two-letter tokens, which could underestimate token overlap and so overestimate token priming boosts, we code shared tokens rather than compare final overlap length to token length.) We fit 28 linear regression models, 12 by tokenizer and 10 by final overlap length, regressing priming on the interaction of whether related prime–target pairs overlap and whether they end with the same token (plus frequency, length, and number of tokens). If tokenizers selectively identify psychologically salient word endings, we should observe significant positive interactions of overlaps with shared tokens, such that shared tokens predict larger priming effects than other overlapping strings.

3.3 Results

First, As reported in Table 3, shared final tokens explain variance in semantic priming after accounting for shared final morphemes and other lexical variables: ANOVA comparing models with and without the shared token factor are significant, for all 12 tokenizers and all seven languages. However, as illustrated in Figure 5, this impact of tokens varies across languages. As reported in full in the appendix, the effect of shared final tokens is significantly positive for all 12 tokenizers, but in the by-language analyses, it is significantly *negative* for Hungarian. Hungarian differs

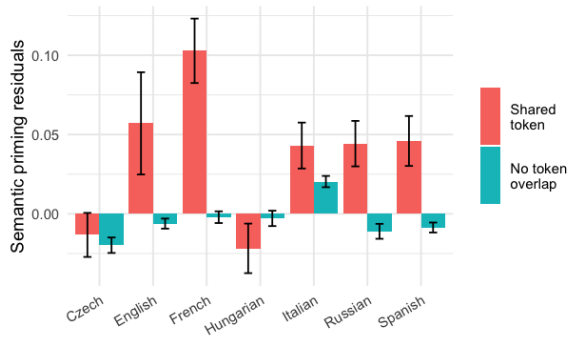


Figure 5: Semantic priming residuals (after regressing priming effects on whether prime–target pairs share final morphemes and lexical variables) as a function of whether prime–target pairs share final tokens.

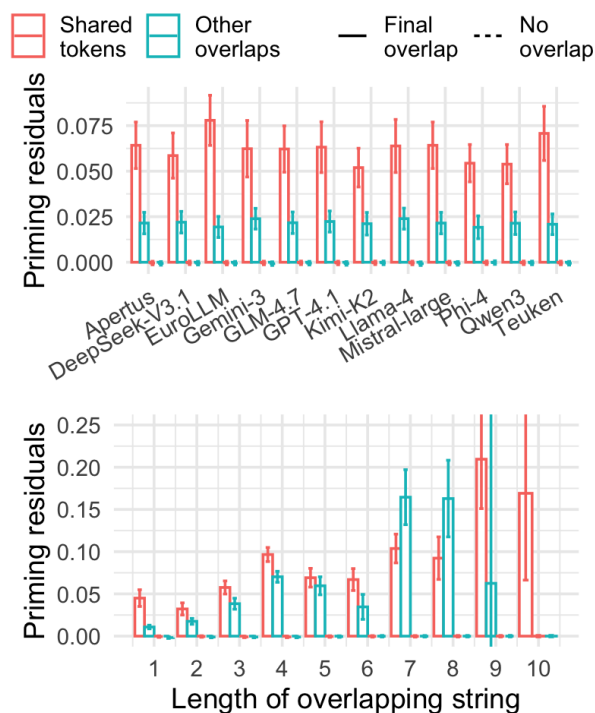


Figure 6: Semantic priming residuals as a function of whether prime–target pairs share final strings and whether they end with the same token, by tokenizer (top) and string length (bottom). The dashed lines, indicating residuals for prime–target pairs that do not end with the same string, are hard to see because after accounting for frequency and length, the priming effects do not systematically trend high or low.

from the other six languages not only because it is Uralic rather than Indo-European but also because it is highly agglutinative (i.e., complex words comprise chains of morphemes), making it especially amenable to morphological decomposition, whereas the other languages are fusional (i.e., different combinations of semantic features

are expressed by different morphemes, so a given feature does not consistently map onto the same string). Consequently, the shared token factor is collinear with the shared morpheme factor in Hungarian, leaving little variance in semantic priming to be explained by byte-pair encoding.

Second, as illustrated in Figure 6 and reported in full online, final tokens also predict greater priming than non-token final strings do, independently of morphology. The simple effect of shared final strings is significant and positive for all 12 tokenizers, and crucially, so is the interaction with shared tokens. The simple effect indicates that shared final strings predict a boost in priming, even when pairs do not share tokens, but the interaction indicates that tokenizers selectively identify overlaps that correspond to still larger priming effects. However, when fitting separate models for each length of overlapping string, these effects hold only up to four-character overlaps, after which there are far fewer observations and far fewer opportunities to differentiate category markers from meaningless chunks of words, at least among semantically related pairs of words in SPAM-L.

4 Analysis 3

So far, we have shown that byte-pair encoding identifies chunks of words that convey semantic information, beyond what is captured by fastText embeddings, subs2vec embeddings, word2vec embeddings, Sigmorphon segmentations, and other overlapping strings. However, these tokenizers seem to be trained on predominantly English data, which makes it more appropriate to evaluate them on English words and yet more difficult to do so: Only 0.5% of English prime–target pairs share tokens, by far the least of any language (the next lowest are French and Spanish, at 4.1% and 5.6%, respectively), and 91.6% of English targets comprise single tokens, by far the most of any language (the next highest are French and Spanish, at 43.2% and 42.6%). [Haslett 2025](#) and [Uzan and Pinter 2026](#) provide evidence that larger vocabularies of longer tokens obscure information in word form, and [Gutierrez-Vasques et al. 2023](#) provide evidence that information about linguistic typology emerges quickly when training a byte-pair encoding tokenizer, after only a few hundred tokens. This analysis therefore seeks to identify the optimal tokenizer vocabulary size for representing English words, which

is to say, the vocabulary of tokens that best supplements distributional semantics in explaining semantic priming effects.

4.1 Materials

We trained byte-pair encoding tokenizers on the wikitext-103 dataset (see huggingface.co/tokenizers), with vocabulary sizes of 500, 1,000, 2,000, 4,000, 8,000, 16,000, 32,000, 64,000, 128,000, and 256,000 tokens. For example, *interdisciplinary* comprises nine tokens in the 500-token vocabulary, six tokens in the 4,000-token vocabulary, four tokens in the 32,000-token vocabulary, and one token in the 64,000-token vocabulary. For each vocabulary size, we coded whether prime–target pairs share final tokens. We also updated the models and variables used in the first two analyses, which were constrained by the need for multilingual databases. We replaced the fastText model that uses 5-gram subwords with a 4-gram model trained on the Google News dataset; replaced word frequencies from wordfreq with SUBTLEX-US frequencies (Brysbaert and New, 2009); and replaced Sigmorphon segmentations with MorphoLex (Sánchez-Gutiérrez et al., 2018). We also used the Unimorph umLabeller, which was expressly designed to determine whether tokenizations of English words are morphological or not (Batsuren et al., 2024).

4.2 Analysis plan

First, we fit 10 pairs of linear regression models, one per vocabulary size, with fastText similarity, word frequency, word length, and tokens per word as predictors of semantic priming, and we compared the variance explained when adding shared final tokens as a predictor. Second, we fit those same models but with the shared final morpheme factor instead of distributional semantics. For example, the related prime–target pairs *client* and *patient*, *height* and *weight*, and *nausea* and *diarrhea* share final tokens according to tokenizers of several vocabulary sizes, but according to MorphoLex, they do not share final morphemes. Finally, we measured the proportion of targets deemed morphological by the Unimorph umLabeller at each vocabulary size. For example, *activity* is labelled “alien” in the 500-token vocabulary, when it is tokenized as *a + ct + iv + ity* and so splits *act* into infelicitous pieces; it is labelled “morph” in the 1,000-, 2,000-, and 4,000-token vocabularies, tokenized as *act + iv + ity*; and

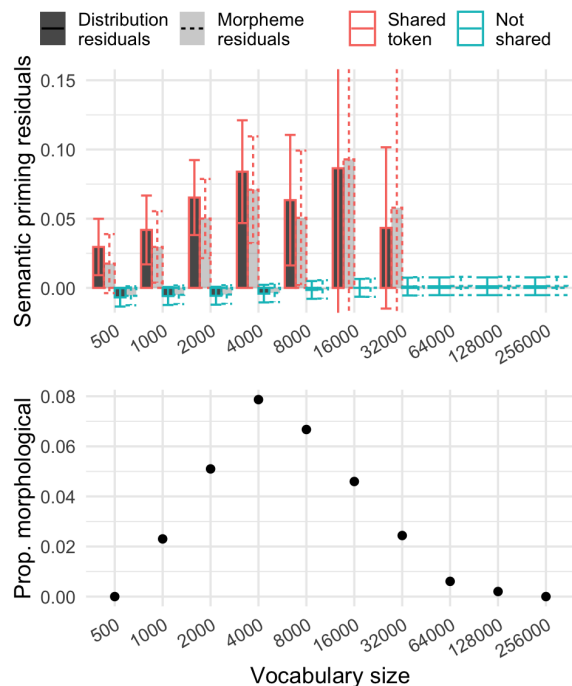


Figure 7: Additional variance explained by shared final tokens in English items (top) and the proportion of target tokenizations deemed morphological by the Unimorph umLabeller (bottom).

it is labelled “vocab” in 8,000-token vocabularies and up, when it is a single token (i.e., the token is lexical). We hypothesize that byte-pair encoding identifies meaningful constituents, so larger vocabularies will explain more variance in priming and have a larger proportion of morphological tokenizations, but only up to a certain point, after which, word-like tokens will obscure constituents.

4.3 Results

Consistent with our hypothesis, tokenizers with larger vocabularies explain more variance in semantic priming, but only up to the 16,000-token vocabulary, after which, tokens cease to add value. As reported in Table 4, shared final tokens capture the most additional information in the 4,000-token vocabulary, beyond what is explained by distributional similarity or shared final morphemes. According to Unimorph and as illustrated in Figure 7, tokens are most likely to align with morphology in the 4,000-token vocabulary, which suggests that when tokens add the most value, they correspond best to traditional morphological decomposition. This underscores the need to account for both conventional morphemes, as in *surname* or *governor*, as well as etymological or analogical re-

Vocab.	Sem. Δ (p)	Morph. Δ (p)
500	.0130 (2e-4)	.0076 (.005)
1,000	.0162 (4e-5)	.0126 (4e-4)
2,000	.0262 (2e-7)	.0233 (2e-6)
4,000	.0270 (1e-7)	.0244 (1e-6)
8,000	.0065 (.007)	.0042 (.028)
16,000	.0017 (.100)	.0021 (.089)
32,000	-.0010 (.944)	-.0011 (.880)
64,000	0 (NA)	0 (NA)
128,000	0 (NA)	0 (NA)
256,000	0 (NA)	0 (NA)

Table 4: Change in adjusted R^2 from linear regression models predicting semantic priming in English, when including fastText similarity (Sem.) or shared morphemes (Morph.) in the base model.

relationships that do not lend themselves to morphological decomposition, as in *espionage* and *sabotage* or *height* and *weight*. Byte-pair encoding is well equipped to identify such relationships, but only with small vocabularies that decompose words. Larger vocabularies represent most word holistically and so fail to identify meaningful subword constituents. In those regression models, the shared token factor is invariant.

5 Analysis 4

Do LLMs take advantage of the semantic information available in tokens, incorporating it into their representations? If so, shared final tokens should not explain additional variance when using LLM representations to predict semantic priming.

5.1 Materials

We measured the semantic similarity of prime–target pairs according to five LLMs: DeepSeek-V3.1, GPT-4.1, Kimi K2, Llama 4 Maverick, and Qwen3-235B. Each of these models returns `top_logprobs` values when accessed via API, which allows us to gather next-token probabilities while using commercial LLMs that are closed-source or too large to run locally. Related work has trained models from the ground up (e.g., [Batsuren et al. 2022](#); [Bunzeck and Zarri   2025](#)) or measured the cosine similarity of contextualized embeddings (e.g., [Cassani et al. 2023](#)). These are valuable lines of research, but they require convergent evidence from state-of-the-art LLMs, whose behaviour may differ qualitatively (cf. [Wei et al. 2022](#)).

We gathered LLM output on two tasks, one eliciting explicit semantic relatedness ratings and the other probing implicit similarity of word meanings. First, we asked each LLM to rate the semantic relatedness of each prime–target pair. For example, *On a scale of 1 to 7, with 1 indicating an unrelated meaning, 4 indicating a somewhat related meaning, and 7 indicated a very closely related meaning, how related are the meanings of the Hungarian words be  rat and kij  rat?* We included a system prompt: *Respond only with a number between 1 and 7.* We averaged the likelihood of the LLM responding with the values 1 through 7 (i.e., multiplying each scale value by its logprob and then summing all seven products, for a probabilistic rating between 1 and 7). As with fastText cosine similarity, we subtracted ratings for unrelated pairs from related pairs. Larger rating differences should have a positive relationship with semantic priming in humans.

Second, for the implicit representations, we conducted an odd-one-out task with each item’s target, related prime, and unrelated prime. For example, *Which one of these three Hungarian words is not like the other two? A: be  rat, B: kij  rat, C: parancsnok* (with the order of primes and targets randomized, each on a new line). We included the system prompt *Respond only with one of the following letters: A, B, or C.* We measured the likelihood of each LLM deciding that the unrelated prime (e.g., *parancsnok*, C) was the odd word out. Larger odd-one-out probabilities for unrelated primes imply stronger similarity of targets to related primes and so should have a positive relationship with semantic priming in humans (i.e., those related primes should speed up target recognition more, relative to unrelated primes).

Notice that for both relatedness ratings and odd-one-out decisions, we record logprobs of single placeholder tokens (i.e., 1–7 or A, B, or C). In this way, we avoid making (arguably) circular claims about the impact of subword tokens on semantic representations because we do not measure the cosine similarity of a token’s embedding in two different contexts or the surprisal of a token conditioned on itself. For example, according to Kimi K2, *be  rat* and *kij  rat* share one of their three tokens, which could inflate the similarity of those words or deflate the surprisal of one following the other. Research that works directly with token embeddings is an important part of the literature, but

	DeepSeek	GPT	Kimi	Llama	Qwen
Adjusted R^2 for relatedness ratings					
Base	.068	.068	.069	.068	.069
+Rate	.068	.068	.071	.069	.069
p	.421	.180	1e-9	1e-4	.334
+Tok.	.071	.071	.075	.071	.072
p	3e-12	2e-12	2e-14	1e-9	3e-13
Adjusted R^2 for odd-one-out probabilities					
Base	.068	.068	.069	.068	.068
+Prob.	.068	.069	.071	.072	.070
p	.208	1e-4	6e-9	3e-16	7e-6
+Tok.	.071	.072	.074	.074	.073
p	6e-12	5e-12	2e-12	2e-8	9e-14

Table 5: Results of ANOVA comparing linear regression models. Adding whether prime–target pairs share final tokens (+Tok.) improves fit over models that include LLM representations (+Rate / +Prob.) and distributional similarity and other lexical variables (Base).

it has limitations, including the potential circularity of the impact of subword tokens on word meanings and, again, the exclusion of the most popular closed-source LLMs.

5.2 Analysis plan

First, we regressed priming effects on fastText similarity and the other lexical variables. Next, we added relatedness ratings and odd-one-out probability as predictors of priming, in separate linear regression models. Finally, as above, we added whether prime–target pairs share final tokens, for each LLM, to see whether the token factor improves fit.

5.3 Results

As reported in Table 4, adding relatedness ratings significantly improves model fit only for Kimi K2 and Llama 4, while adding odd-one-out probability improves fit for every LLM other than DeepSeek-V3.1. Crucially, as illustrated in Figure 8, the shared token factor substantially improves model fit for all five LLMs, beyond what is captured by relatedness ratings and odd-one-out probabilities. As we report in the appendix, the results vary substantially by language, but they do not vary systematically; the interaction of the shared token factor with language prevalence is significant only for GPT-4.1 and Qwen3 for relatedness ratings, and only for GPT-4.1 for odd-one-out probabilities, with positive coefficients indicating that the token boost is greater in languages that are

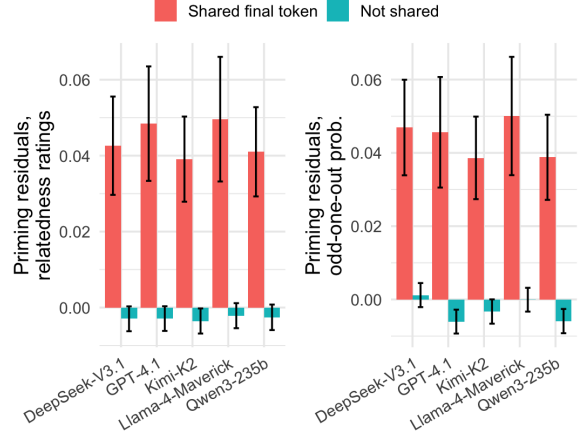


Figure 8: Semantic priming residuals (regressed on relatedness ratings / odd-one-out probability, distributional similarity, and other lexical variables) as a function of whether prime–target pairs share final tokens.

more common on the internet in these three cases.

6 Discussion and conclusion

Distributional semantic models are indispensable tools, able to approximate word meanings automatically and efficiently. However, their representations are incomplete, and because LLMs are a type of distributional semantic model, this could result in concept misalignment, such that words mean different things to people versus LLMs. In Analysis 1, we found that byte-pair encoding—the compression algorithm LLMs use to segment text into chunks—identifies meaningful constituents. Across 18 languages, people recognized target words, such as *abdominal* and *sabotage*, faster after reading primes that end with the same token, such as *spinal* and *espionage*, and this boost in priming is not explained by fastText embeddings, word2vec embeddings, word frequency, word length, or the number of tokens per word.

In Analysis 2, we provided evidence that byte-pair encoding captures semantic information independent of morphemes. For example, the forms of *height* and *weight* reflect their semantic relatedness, but they do not share morphemes. In this way, byte-pair encoding is consistent with contemporary theories of morphology, which emphasize systematic form–meaning mappings (Stevens and Plaut, 2022), which gives byte-pair encoding an advantage over morpheme-based decomposition. As shown in the appendix, Unimorph labels many semantically informative tokenizations as “alien”, either because they correspond to morphemes but

differ in their surface forms (e.g., *inal* subsumes but differs from the morpheme *al* in *spinal*), or because they correspond to sub-morphemic category markers (e.g., *ar* in *solar* is not considered a morpheme by Sigmorphon or Unimorph). We further found that the variance in priming explained by tokens is not merely due to the flexibility of tokenization relative to morphological annotation: Tokens predict larger boosts in priming than other overlapping strings of characters do, providing further evidence that byte-pair encoding selectively identifies chunks of words that are salient to human language processing.

Byte-pair encoding is based on orthographic regularities, so the fact that it identifies meaningful constituents simply demonstrates that those constituents recur in language. That is, the value added by tokens reflects the information implicit in word form, and as we show in Analysis 3, large vocabularies obscure that information. We trained tokenizers with increasing vocabularies, and they ceased to identify meaningful constituents beyond the meager size of 16,000 tokens, an order of magnitude smaller than the vocabularies of state-of-the-art LLMs. This raises questions about the tradeoff between compressing text and identifying semantic information. Would a language-specific tokenizer with a vocabulary of less than 16,000 tokens defeat the purpose of compression by failing to represent as much as text as possible with as few chunks as possible? We doubt it. Brysbaert et al. 2016 report that when accounting for morphological relationships, the average English speaker knows less than 12,000 words, so it should take only a small vocabulary of tokens to approach the optimal vocabulary size, after which, words that get segmented into multiple tokens would be precisely the sorts of low-frequency words that people tend to decompose (cf. Stemberger and MacWhinney 1986). This is especially important in low-resource languages, which lack the abundance of training data needed for distributional semantics and on which LLMs perform worse (e.g., Li et al. 2024). Bear in mind that state-of-the-art LLMs use multilingual tokenizers, which confound direct comparisons to our 16,000-token model. Still, those tokenizers prioritize English, Chinese, and other high-resource languages (Arnett et al., 2024b; Limisiewicz et al., 2024; Petrov et al., 2023), as exemplified by long, single-token words such as *interdisciplinary*, so there is lots of

room to facilitate access to meaningful subword constituents while, at the same time, improving fairness across languages by devoting less space in a tokenizer’s vocabulary to high-resource languages.

The meaningfulness of tokens may be a moot point, though, since all five LLMs failed to take advantage of the semantic information available in subword tokens, in Analysis 4. As with fast-Text and word2vec, representations of word meanings from LLMs systematically underestimated the similarity of prime–target pairs that share tokens, which speaks to the limits of statistical learning. LLMs are, fundamentally, neural networks that learn patterns in tokens, but using tokens in a human-like fashion—knowing which subwords to seize on, and when and how—seems to require inductive biases that LLMs lack. The co-evolution of people and languages has led to onsets that discriminate between meanings and suffixes that function as category markers, and people tend to extract the right sorts of information from the right parts of words (Cutler et al., 1985; St. Clair et al., 2009). Without this inclination, LLMs may struggle to transcend the narrow worldview implied by distributional semantics, despite the fact that byte-pair encoding is a font of taxonomic information.

Limitations

First, the facilitatory effects of shared tokens could be purely orthographic, i.e., form priming, in which case, shared tokens may not correspond to semantic features. Future work should investigate this possibility, but the experimental setup of SPAM-L and the nature of the tokens that boost priming strongly suggest that we have detected semantic relationships. The boost follows lexical decisions about primes, whereas form priming typically requires subliminal exposure to primes; the tokens often correspond to morphemes or category markers; and related prime–target pairs share tokens much more often (14% of items) than unrelated pairs do (3%). We conclude that tokens capture semantic information that distribution overlooks. Second, by analyzing shared tokens and LLM output, we have not directly investigated contextualized token embeddings or positional embeddings. Future work should consider these intermediate stages of LLM processing.

Acknowledgements

This work was supported by the Research Grant Council of Hong Kong (AoE/E-601/24-N).

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Aaron Adcock, Aayushi Srivastava, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pande, Abhinav Pandey, Abhinav Sharma, Abhishek Kadian, Abhishek Kumawat, Adam Kelsey, et al. 2026. The llama 4 herd: Architecture, training, evaluation, and deployment notes. *arXiv preprint arXiv:2601.11659*.
- Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.
- Robin Algayres, Charles-Éric Saint-James, Mahi Luthra, Jiayi Shen, Dongyan Lin, Youssef Bencheikroun, Rashel Moritz, Juan Pino, and Emmanuel Dupoux. 2025. Longtail-swap: benchmarking language models’ abilities on rare words. *arXiv preprint arXiv:2510.04268*.
- Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, et al. 2024. Teuken-7b-base & teuken-7b-instruct: Towards european llms. *arXiv preprint arXiv:2410.03730*.
- Simona Amenta, Fritz Günther, and Marco Marelli. 2020. A (distributional) semantic perspective on the processing of morphologically complex words. *The Mental Lexicon*, 15(1):62–78.
- Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.
- Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, et al. 2025. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*.
- Catherine Arnett, Tyler Chang, and Sean Trott. 2024a. Different tokenization schemes lead to comparable performance in spanish number agreement. In *Proceedings of the 21st SIG-MORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–38.
- Catherine Arnett, Tyler A Chang, and Benjamin K Bergen. 2024b. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 1–9.
- Catherine Arnett, Marisa Hudspeth, and Brendan O’Connor. 2025. Evaluating morphological alignment of tokenizers in 70 languages. *arXiv preprint arXiv:2507.06378*.
- Jeremy Auguste, Arnaud Rey, and Benoit Favre. 2017. Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd workshop on evaluating vector space representations for NLP*, pages 21–26.
- R Harald Baayen, 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(1):4895891.
- R Harald Baayen, Petar Milin, Dusica 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.
- Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinović, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina

- Pelegriová, et al. 2022. The sigmorphon 2022 shared task on morpheme segmentation. *arXiv preprint arXiv:2206.07615*.
- Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating subword tokenization: Alien subword composition and oov generalization challenge. *arXiv preprint arXiv:2404.13292*.
- Curtis A Becker. 1979. Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 5(2):252.
- Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478–4486.
- Benjamin K Bergen. 2004. The psychological reality of phonaesthemes. *Language*, 80(2):290–311.
- Damián E Blasi, Søren Wichmann, Harald Hammarström, Peter F Stadler, and Morten H Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 113(39):10818–10823.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.
- Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624.
- Roger Brown and Jean Berko. 1960. Word association and the acquisition of grammar. *Child development*, pages 1–14.
- Marc Brysbaert 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.
- Marc Brysbaert, 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.
- Erin M Buchanan, Kelly Cuccolo, Tom Heyman, Niels Van Berkel, Nicholas A Coles, Aishwarya Iyer, Kim Peters, Anna E van’t Veer, Maria Montefinese, Nicholas P Maxwell, et al. 2026. Measuring the semantic priming effect across many languages. *Nature human behaviour*, 10(1):182–201.
- John A Bullinaria and Joseph P Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.
- Bastian Bunzeck and Sina Zarriß. 2025. Subword models struggle with word learning, but surprisal hides it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 286–300.
- Giovanni Cassani, Yu-Ying Chuang, and R Harald Baayen. 2020. On the semantics of non-words and their lexical category. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(4):621.
- Giovanni Cassani, Fritz Günther, Giuseppe Atanasio, Federico Bianchi, and Marco Marelli. 2023. Meaning modulations and stability in large language models: An analysis of bert embeddings for psycholinguistic research. *osf.io/preprints/psyarxiv/b45ys*.
- Nichol Castro, Massimo Stella, and Cynthia SQ Siew. 2020. Quantifying the interplay of semantics and phonology during failures of word retrieval by people with aphasia using a multiplex lexical network. *Cognitive science*, 44(9):e12881.
- Kenneth Ward Church. 2020. Emerging trends: Subwords, seriously? *Natural Language Engineering*, 26(3):375–382.
- Anne Cutler, John A Hawkins, and Gary Gilligan. 1985. The suffixing preference: A processing explanation. *Linguistics*, 23.

- Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, Sofia Koutalidis, Manfred Krifka, Pärtel Lippus, et al. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, 377(1841):20200390.
- Simon De Deyne, Danielle J Navarro, Guillem Collell, and Andrew Perfors. 2021. Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, 45(1):e12922.
- Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 1861–1870.
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Mark Dingemanse, Damián 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.
- Doris R Entwisle. 1966. Form class and children’s word associations. *Journal of verbal learning and verbal behavior*, 5(6):558–565.
- Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9:17–1.
- Laurie Beth Feldman. 2000. Are morphological effects distinguishable from the effects of shared meaning and shared form? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6):1431.
- John Firth. 1957. A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis, Special Volume/Blackwell*.
- Kenneth I Forster and Chris Davis. 1984. Repetition priming and frequency attenuation in lexical access. *Journal of experimental psychology: Learning, Memory, and Cognition*, 10(4):680.
- Kenneth I Forster and Csaba Veres. 1998. The prime lexicality effect: form-priming as a function of prime awareness, lexical status, and discrimination difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(2):498.
- Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.
- Daniele Gatti, 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.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv:2403.08295*.
- GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Jonathan Grainger, Pascale Colé, and Juan Segui. 1991. Masked morphological priming in visual word recognition. *Journal of memory and language*, 30(3):370–384.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.
- Joseph H Greenberg. 1957. *Essays in linguistics*. University of Chicago Press.
- Fritz Günther, Carolin Dudschig, and Barbara Kaup. 2016. Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, 69(4):626–653.
- Bernal Jimenez Gutierrez, Huan Sun, and Yu Su. 2023. Biomedical language models are robust to sub-optimal tokenization. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

- Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. Languages through the looking glass of bpe compression. *Computational Linguistics*, 49(4):943–1001.
- Harald Hammarström. 2021. Measuring prefixation and suffixation in the languages of the world. In *Proceedings of the third workshop on computational typology and multilingual NLP*, pages 81–89.
- Zellig Harris. 1954. Distributional structure.
- David A Haslett. 2025. Tokenization changes meaning in large language models: Evidence from chinese. *Computational Linguistics*, pages 1–30.
- David A Haslett and Zhenguang G Cai. 2023. Similar-sounding words flesh out fuzzy meanings. *Journal of Experimental Psychology: General*, 152(8).
- David A Haslett and Zhenguang G Cai. 2024. Systematic mappings of sound to meaning: A theoretical review. *Psychonomic Bulletin & Review*, 31(2):627–648.
- David A Haslett and Zhenguang G Cai. 2025. How much semantic information is available in large language model tokens? *Transactions of the Association for Computational Linguistics*, 13:408–423.
- David A Haslett, Bo Yao, Xufeng Duan, and Zhenguang G Cai. 2026. Incipient constituents: Phonesthemes facilitate word processing in english. *Open Mind*, 10:827–856.
- Peter Hendrix 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.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.
- Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves bert’s interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608.
- Valentin Hofmann, Hinrich Schuetze, and Janet B Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Association for Computational Linguistics*.
- G. W. Humphreys, P. T. Quinlan, L. J. Evett, and D. Besner. 1987. [Orthographic priming: Qualitative differences between priming from identified and unidentified primes](#). In M. Coltheart, editor, *Attention and Performance 12: The Psychology of Reading*, pages 105–125.
- Keith A Hutchison. 2003. Is semantic priming due to association strength or feature overlap? a microanalytic review. *Psychonomic bulletin & review*, 10(4):785–813.
- Haris Jabbar. 2023. Morphpiece: A linguistic tokenizer for large language models. *arXiv preprint arXiv:2307.07262*.
- Michael N Jones, Walter Kintsch, and Douglas JK Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of memory and language*, 55(4):534–552.
- Pable Contreras Kallens and Morten H Christiansen. 2020. Phonological cues to semantic class membership across hundreds of languages. *Proceedings of the 13th International Conference on the Evolution of Language*.
- Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, et al. 2021. Mistral—a journey towards reproducible language model training. *Centre for Research on Foundation Models, Stanford University*.
- Michael H Kelly. 1992. Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological review*, 99(2):349.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue

- Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.
- Abhilasha A Kumar, Nancy B Lundin, and Michael N Jones. 2022. Mouse-mole-vole: The inconspicuous benefit of phonology during retrieval from semantic memory. In *Proceedings of the annual meeting of the Cognitive Science Society*.
- Nahyun Kwon and Erich R Round. 2015. Phonaesthemes in morphological theory. *Morphology*, 25(1):1–27.
- Thomas K Landauer and Susan T Dumais. 1997. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.
- Alessandro Lenci, Magnus Sahlgren, Patrick Juniaux, 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).
- Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *arXiv preprint arXiv:2404.11553*.
- Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. Myte: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076.
- Pierre Lison and Jörg Tiedemann. 2016. Open-subtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 923–929.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Margery Lucas. 2000. Semantic priming without association: A meta-analytic review. *Psychonomic bulletin & review*, 7(4):618–630.
- Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208.
- Gary Lupyan and Daniel Casasanto. 2015. Meaningless words promote meaningful categorization. *Language and Cognition*, 7(2):167–193.
- Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2022. Bpe vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. *Findings of the Association for Computational Linguistics*.
- Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.
- Marco Marelli and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122(3):485.
- Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, et al. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.
- David E Meyer and Roger W Schvaneveldt. 1971. Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Padraic Monaghan, Morten H Christiansen, and Stanka A Fitneva. 2011. The arbitrariness of the sign: learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3):325.

- Dennis Norris, James M McQueen, and Anne Cutler. 2002. Bias effects in facilitatory phonological priming. *Memory & Cognition*, 30(3):399–411.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.
- Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.
- Tiago Pimentel, Arya D McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 1751–1764.
- Daniel Pimienta. 2022. Resource: Indicators on the presence of languages in internet. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 83–91.
- David C Plaut 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.
- R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Monique Radeau, Mireille Besson, Elisabeth Fonteneau, and Sao Luis Castro. 1998. Semantic, repetition and rime priming between spoken words: behavioral and electrophysiological evidence. *Biological psychology*, 48(2):183–204.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).
- Sunayana Rane, Polyphony J Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L Griffiths. 2024. Concept alignment. *arXiv preprint arXiv:2401.08672*.
- Kathleen Rastle, Matt H Davis, William D Marslen-Wilson, and Lorraine K Tyler. 2000. Morphological and semantic effects in visual word recognition: A time-course study. *Language and cognitive processes*, 15(4-5):507–537.
- Martin Redington, Nick Chater, and Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science*, 22(4):425–469.
- Jamie Reilly, Jinyi Hung, and Chris Westbury. 2017. Non-arbitrariness in mapping word form to meaning: Cross-linguistic formal markers of word concreteness. *Cognitive Science*, 41(4):1071–1089.
- Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980.
- Claudia H Sánchez-Gutiérrez, Hugo Mailhot, S Hélène Deacon, and Maximiliano A Wilson. 2018. Morpholex: A derivational morphological database for 70,000 english words. *Behavior research methods*, 50(4):1568–1580.
- Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

- Robyn Speer. 2022. rspeer/wordfreq: v3. 0.
- Michelle C St. Clair, Padraic Monaghan, and Michael Ramscar. 2009. Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, 33(7):1317–1329.
- Massimo Stella, Nicole M Beckage, and Markus Brede. 2017. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific reports*, 7(1):46730.
- Joseph Paul Stemberger and Brian MacWhinney. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition*, 14(1):17–26.
- Patience Stevens and David C Plaut. 2022. From decomposition to distributed theories of morphological processing in reading. *Psychonomic Bulletin & Review*, 29(5):1673–1702.
- Yue Sun and David Poeppel. 2023. Syllables and their beginnings have a special role in the mental lexicon. *Proceedings of the National Academy of Sciences*, 120(36):e2215710120.
- Michael K Tanenhaus, Helen P Flanigan, and Mark S Seidenberg. 1980. Orthographic and phonological activation in auditory and visual word recognition. *Memory & Cognition*, 8(6):513–520.
- Omri Uzan and Yuval Pinter. 2026. Charbench: Evaluating the role of tokenization in character-level tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33296–33304.
- Jeroen Van Paridon and Bill Thompson. 2021. subs2vec: Word embeddings from subtitles in 55 languages. *Behavior research methods*, 53(2):629–655.
- Andrea Gregor de Varda, Daniele Gatti, Marco Marelli, and Fritz Günther. 2024. Meaning beyond lexicality: Capturing pseudoword definitions with language models. *Computational Linguistics*, 50(4):1313–1343.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Bodo Winter, Paula Pérez-Sobrino, and Lucien Brown. 2019. The sound of soft alcohol: Cross-modal associations between interjections and liquor. *PLoS One*, 14(8):e0220449.
- Bei Xiao, Xufeng Duan, David A Haslett, and Zhenguang Cai. 2025. Human-likeness of llms in the mental lexicon. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 586–601.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Jinbiao Yang, Antal Van den Bosch, and Stefan L Frank. 2022. Unsupervised text segmentation predicts eye fixations during reading. *Frontiers in Artificial Intelligence*, 5:731615.
- Greig I de Zubicaray, Joanne Arciuli, Elaine Kearney, Frank Guenther, and Katie L McMahon. 2023. On the roles of form systematicity and sensorimotor effects in language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(3):431.

A Appendix

Tokenizer	Base	Tokens	Δ	p
DeepSeek	.0681	.0704	.0023	2e-9
GLM-4.7	.0678	.0691	.0014	3e-6
Kimi-K2	.0690	.0711	.0021	1e-8
Qwen3	.0683	.0696	.0014	3e-6
Gemini-3	.0678	.0692	.0014	2e-6
GPT-4.1	.0675	.0697	.0021	6e-9
Llama-4	.0676	.0692	.0016	4e-7
Phi-4	.0684	.0700	.0016	3e-7
Apertus	.0680	.0701	.0022	4e-9
EuroLLM	.0676	.0705	.0029	2e-11
Mistral-large	.0680	.0701	.0022	4e-9
Teuken	.0707	.0732	.0025	4e-10

Table 6: Analysis 1: Adjusted R^2 from linear regression models, without (Base) and with (Tokens) the shared token factor, by tokenizer. The tokenizers average across languages. The Base models include fast-Text similarity and other lexical variables. Compare this to Table 2, on page 8, which reports the same analyses by language.

Tokenizer	Prevalence	Script
DeepSeek	.011 (.064)	.013 (.215)
GLM	.014 (.021)	.013 (.240)
Kimi	.004 (.473)	.008 (.430)
Qwen	.016 (.006)	.001 (.921)
Gemini	.017 (.015)	.031 (.015)
GPT	.022 (8e-4)	.005 (.685)
Llama	.017 (.017)	.037 (.005)
Phi	.021 (1e-5)	-.007 (.462)
Apertus	.013 (.025)	.028 (.015)
EuroLLM	.010 (.076)	.016 (.138)
Mistral	.013 (.025)	.028 (.015)
Teuken	.007 (.262)	-.010 (.384)

Table 7: Analysis 1: Coefficients (and p-values, in parentheses) for the interaction of the shared token factor with language prevalence (i.e., the change in priming as a function of the logged percentage of the internet that a language makes up) and with script type (i.e., the change in priming for languages that use a script other than the Latin alphabet). Positive interactions with language prevalence indicate tokenizers that better identify constituents with larger effects on human word processing in higher resource languages. The mostly null interactions with script type indicate that the effects of shared tokens differ little in languages that use the Latin alphabet versus other scripts.

Lang.	Chinese	European	Vocab.
Czech	.004 (.71)	.012 (.35)	.002 (.68)
Danish	-.001 (.98)	.007 (.71)	.005 (.56)
German	.004 (.76)	.005 (.71)	.006 (.27)
Greek	.024 (.33)	.034 (.22)	.008 (.51)
Spanish	-.016 (.35)	-.006 (.74)	.012 (.16)
French	-.012 (.61)	-.001 (.71)	-.001 (.94)
Hungar.	-.001 (.94)	-.011 (.59)	-.001 (.88)
Italian	.008 (.56)	.020 (.19)	.007 (.24)
Polish	.014 (.23)	.011 (.35)	.001 (.92)
Roman.	-.010 (.59)	.022 (.29)	.016 (.07)
Russian	-.009 (.39)	-.002 (.86)	.005 (.20)
Serbian	.011 (.36)	.005 (.69)	-.001 (.87)
Turkish	.002 (.91)	-.002 (.88)	-.006 (.31)
Korean	.015 (.60)	.014 (.63)	.003 (.79)
English	.039 (.50)	.007 (.88)	-.011 (.40)
Japan.	-.000 (.99)	.0163 (.42)	.006 (.45)
Mandar.	-.053 (.21)	-.005 (.81)	-.007 (.32)
Portug.	-.006 (.71)	-.013 (.49)	-.003 (.71)

Table 8: Analysis 1: Coefficients (and p-values) for the interaction of the shared token factor with tokenizer region of origin (i.e., the change in predicted priming for European or Chinese tokenizers, compared to American tokenizers), and the interaction of the shared token factor with tokenizer vocabulary size (i.e., the change in priming for languages that use a script other than the Latin alphabet). For the sake of space, p-values are reported only to two decimal places. As discussed on page 7, the only (marginally) significant effect is the interaction with vocabulary size for Romanian.

Tokenizer	Initial	Final
DeepSeek	-.001 (.860)	.044 (3e-5)
GLM-4.7	-.019 (.119)	.065 (2e-9)
Kimi K2	-.008 (.291)	.048 (3e-7)
Qwen3	-.004 (.048)	.059 (4e-9)
Gemini 3	-.010 (.321)	.055 (1e-5)
GPT-4.1	-.010 (.320)	.056 (2e-6)
Llama 4	-.004 (.704)	.050 (1e-4)
Phi-4	-.010 (.170)	.052 (1e-8)
Apertus	-.018 (.049)	.067 (3e-9)
EuroLLM	.007 (.444)	.056 (8e-6)
Mistral	-.018 (.049)	.067 (3e-9)
Teuken	.008 (.316)	.034 (.001)

Table 9: Analysis 1: Coefficients (and p-values) for the interaction of the shared token factor with overlap position, by tokenizer. “Initial” refers to the simple effect of shared tokens (because initial overlap is the reference level); “Final” refers to the interaction of shared tokens with overlap position (i.e., the change in predicted priming for final tokens). As illustrated in Figure 2 on page 6, initial tokens have negligible or small effects on priming, while final tokens boost priming.

Language	Base	Tokens	Δ	p
Czech	.0938	.0947	.0008	.021
English	.1136	.1146	.0009	.014
French	.0503	.0560	.0057	5e-15
Hungar.	.0421	.0428	.0008	.009
Italian	.0434	.0443	.0009	.001
Russian	.0523	.0663	.0140	< 2e-16
Spanish	.0489	.0533	.0044	2e-12

Table 10: Analysis 2: Adjusted R^2 from linear regression models, without (Base) or with (Tokens) the shared token factor. The languages average across 12 tokenizers. The Base models include final morphemes and other lexical variables. Compare this to Table 3, on page 8, which reports the same analyses by tokenizer.

Language	B	SE	p
Czech	.020	.009	.021
English	.052	.021	.014
French	.100	.013	4e-15
Hungarian	-.034	.013	.009
Italian	.023	.007	.001
Russian	.069	.008	< 2e-16
Spanish	.059	.008	2e-12

Table 11: Analysis 2: The effects of shared final tokens, by language. These linear regression models include shared morphemes, word frequency, word length, and tokens per word as co-predictors. This table corresponds to Figure 5, on page 10.

LLM	Shared tokens	Tokens : Prev.
Relatedness ratings		
DeepSeek	.044 (3e-7)	.013 (.125)
GPT-4.1	.051 (3e-6)	.030 (.002)
Kimi K2	.040 (3e-8)	.005 (.451)
Llama 4	.043 (1e-4)	.013 (.214)
Qwen3	.039 (3e-6)	.013 (.083)
Odd-one-out decision probability		
DeepSeek	.044 (4e-7)	.013 (.115)
GPT-4.1	.051 (2e-6)	.031 (.002)
Kimi K2	.038 (2e-7)	.006 (.426)
Llama 4	.041 (3e-4)	.013 (.183)
Qwen3	.042 (.2e-7)	.015 (.053)

Table 12: Analysis 4: Coefficients (and p-values, in parentheses) for the simple effect of shared final tokens and its interaction with logged language prevalence, beyond what is explained by relatedness ratings (top) or by odd-one-out decisions (bottom), along with distributional semantic similarity and lexical variables. This table corresponds to Figure 6, on page 11.

Tokenizer	Base	Tokens	Δ	p
Relatedness ratings				
Czech	.0992	.1003	.0011	.022
Danish	.0513	.0677	.0165	1e-16
German	.1778	.1801	.0023	.146
Greek	.1564	.1774	.0211	9e-7
English	.1400	.1398	-.0001	.616
Spanish	.0631	.0681	.0060	4e-8
French	.0689	.0746	.0057	2e-7
Hungarian	.0621	.0623	.0002	.196
Italian	.0578	.0580	.0002	.145
Japanese	.3211	.3225	.0014	.213
Korean	.0468	.0469	.0001	.284
Polish	.0844	.0846	.0002	.172
Romanian	.0939	.0938	.0000	.357
Russian	.0804	.0851	.0047	6e-5
Serbian	.0824	.0957	.0133	2e-11
Turkish	.1014	.1032	.0018	.006
Portuguese	.0687	.0822	.0135	3e-13
Mandarin	.1140	.1175	.0035	6e-5
Odd-one-out decision probability				
Czech	.1002	.1015	.0013	.014
Danish	.0511	.0678	.0167	2e-16
German	.1773	.1797	.0024	.142
Greek	.1551	.1752	.0201	2e-6
English	.1419	.1418	-.0001	.592
Spanish	.0622	.0682	.0060	4e-8
French	.0698	.0753	.0056	2e-7
Hungarian	.0629	.0631	.0002	.204
Italian	.0057	.0558	.0001	.204
Japanese	.3130	.3145	.0014	.211
Korean	.0467	.0468	.0001	.295
Polish	.0839	.0841	.0002	.163
Portuguese	.0630	.0773	.0143	2e-15
Romanian	.0941	.0941	.0000	.315
Russian	.0795	.0840	.0045	9e-5
Serbian	.0818	.0941	.0123	1e-10
Turkish	.1017	.1032	.0015	.011
Mandarin	.1124	.1160	.0036	5e-5

Table 13: Analysis 4: Adjusted R^2 from linear regression models, without (Base) or with (Tokens) the shared token factor. The Base models include odd-one-out probabilities from five LLMs (top) or odd-one-out probabilities from five LLMs (bottom), along with distributional semantic similarity and lexical variables. As mentioned on page 11, the shared token factor almost always improves model fit (it is negative only for English), but the magnitude and significance of effects vary. As reported in Table 12 (left), the effects tend to be larger in higher resource languages (i.e., the interaction of the shared token factor with language prevalence is positive in all cases), but this positive effect reaches significance only for GPT-4.1.

Vocab size	Cue / target	Final token	Final Sigmorphon	Final MorphoLex	Unimorph label
4,000	senior	ior	senior	senior	alien
	junior		junior	junior	alien
	fresh	h	fresh	fresh	alien
	refresh		fresh	fresh	alien
	lunar	ar	lunar	ar	alien
	solar		solar	ar	alien
	firmly	ly	ly	ly	alien
8,000	strongly		ly	ly	morph
	racist	ist	ist	ist	morph
	communist		ist	ist	morph
	rusty	y	y	y	morph
	dusty		y	y	morph
	accidentally	ally	ly	ly	morph
	intentionally		ly	ly	morph
	physics	ics	s	ics	alien
	mathematics		mathematics	mathematics	alien
	patient	ient	patient	patient	alien
	client		client	client	alien
	liquid	id	liquid	liquid	alien
	fluid		fluid	fluid	alien
	telescope	ope	scope	scope	alien
	microscope		microscope	scope	alien
	shooter	er	er	er	morph
	sniper		er	er	alien
	nickname	ame	nickname	name	alien
	surname		name	name	alien
	suddenly	ly	ly	ly	morph
	unexpectedly		ly	ly	alien
16,000	abdominal	inal	al	al	alien
	spinal		al	al	alien
	ignorance	ance	ance	ance	alien
	arrogance		ce	arrogance	alien
	africa	a	africa	africa	alien
	america		america	america	alien
	espionage	age	espionage	espionage	alien
32,000	sabotage		sabotage	sabotage	alien
	suitcase	case	case	case	morph
	briefcase		case	case	morph
	superman	man	man	case	morph
	batman		man	man	morph
	december	ember	december	december	alien
	november		november	november	alien
	tomato	ato	tomato	tomato	alien
	potato		potato	potato	alien

Table 14: Analysis 3: Selected English prime–target pairs that share final tokens, according to four tokenizers of different vocabulary sizes. Words are listed in the largest vocabulary where they still share tokens. For example, *senior* and *junior* share a final token in the 4,000-token vocabulary but not the 8,000-token vocabulary, and *tomato* and *potato* share a final token according to all four tokenizers. The Sigmorphon and MorphoLex columns report the final morpheme in each word, according to those resources. The Unimorph umLabeller classifies tokenizations as morphological or alien.