Different Tokenization Schemes Lead to Comparable Performance in Spanish Number Agreement

Catherine Arnett*[1], Pamela D. Rivière*[2], Tyler A. Chang[2,3], Sean Trott[2]
[1]Department of Linguistics,
[2]Department of Cognitive Science,
[3]Halıcıoğlu Data Science Institute
UC San Diego
{ccarnett, pdrivier, tachang, sttrott}@ucsd.edu

Abstract

The relationship between language model tokenization and performance is an open area of research. Here, we investigate how different tokenization schemes impact number agreement in Spanish plurals. We find that morphologically-aligned tokenization performs similarly to other tokenization schemes, even when induced artificially for words that would not be tokenized that way during training. We then present exploratory analyses demonstrating that language model embeddings for different plural tokenizations have similar distributions along the embedding space axis that maximally distinguishes singular and plural nouns. Our results suggest that morphologically-aligned tokenization is a viable tokenization approach, and existing models already generalize some morphological patterns to new items. However, our results indicate that morphological tokenization is not strictly required for performance.

1 Introduction

In natural language processing (NLP) pipelines, **tokenizers** segment unstructured text into smaller, discrete constituents ("tokens") for further processing. Importantly, different tokenizers can incur performance and efficiency trade-offs. Assigning a unique token to each word in a corpus may lead to high-precision semantic representations, but the resulting models might be less robust to unseen words and require more computational resources.

Most existing tokenizers allow words to be decomposed into subword tokens (Sennrich et al., 2016; Kudo and Richardson, 2018). They can do so along morphological boundaries (e.g. *books* to ['book', '##s']), but this behavior is not guaranteed. Segmenting words into their lemmas and morphemes might simultaneously allow models to more robustly learn morphosyntactic patterns, more efficiently represent such patterns, and better

generalize to novel words. (An analogous question concerning the storage of whole words vs. learning generalizable rules exists within human psycholinguistics research, e.g., Ullman, 2016).

In the current work, we ask whether and how the tokenization strategy employed facilitates successful language model predictions. We evaluate the effect of three types of plural noun tokenization in Spanish—single-token plurals, morphemically-tokenized plurals, and non-morphemically-tokenized plurals—in the context of a masked article prediction task (§4).[1] We focus on tokenization schemes for plural forms in Spanish, as it offers relatively simple and frequent examples of morphologically complex words. Spanish leverages two primary plural marking strategies, which are highly predictable for any given lemma. We specifically focus on cases where the plural form is composed of the singular form with the addition of '-s' or '-es'.

We find that tokenization schemes are differentially successful, although the effect is small, and article agreement accuracy is high across all tokenization types. Artificial tokenization schemes, where we coerce an initially single-token or non-morphemically-tokenized plural into a morphemic representation, leads to successful task performance, but does not improve performance beyond the original tokenization scheme. In an exploratory analysis, we compare singular and plural form embeddings across all tokenization schemes. We find axes with high overlap between all plural forms (regardless of tokenization scheme) and high discriminability between plural and singular forms, but other axes can still separate different plural tokenization schemes. This work contributes to a growing literature examining the impact of tokenization on the language

---

*Equal contribution.

[1]Note that this categorization scheme mirrors an approach taken in contemporaneous work, using the labels "vocab", "morph", and "alien", respectively.

modeling objective. Code and data are available: `https://github.com/catherinearnett/spanish-plural-agreement`.

2 Related Work

Several studies have investigated morpho-syntactic agreement in BERT-style models across multiple languages (Linzen et al., 2016; Mueller et al., 2020; Edmiston, 2020; Pérez-Mayos et al., 2021, inter alia), finding generally high agreement accuracy. In a subject-verb agreement task, however, BETO incurs a relatively high rate of agreement errors for certain Spanish nouns (despite the ability to extend number agreement to novel words; Haley, 2020). It is unclear to what extent degraded performance is attributable to tokenization scheme, but the word "comanas"—listed as an example of a frequently mis-numbered word—is tokenized non-morphemically into ['coman', '##as'].

Indeed, recent work has demonstrated that morphologically-aware tokenization improves NLP model performance on a variety of downstream benchmarks (Park et al., 2020; Hofmann et al., 2021; Toraman et al., 2023; Jabbar, 2024; Uzan et al., 2024). Most relevantly, Batsuren et al. (2024) devise a tool to classify English words in terms of whether they are stored as single tokens ("vocab"), as multiple morphemic tokens ("morph"), or as multiple non-morphemic tokens ("alien"). The authors find that how multi-morphemic English words are tokenized is correlated with the language model's downstream performance on several tasks.

Following Batsuren et al. (2021), our work investigates how the tokenization of Spanish nouns affects language model predictions involving a specific morphosyntactic rule, providing insight into how morphologically-aware tokenization affects NLP model performance.

3 Model and Data

All experiments use BETO, a Spanish pre-trained BERT model (Cañete et al., 2020) with 110M parameters trained on approximately 3B words. BETO uses a SentencePiece tokenizer (Kudo and Richardson, 2018) with a 32K vocab size.

3.1 Data

All plural nouns and their singular form lemmas were extracted from the AnCora Treebanks (Alonso and Zeman, 2016). Plurals were categorized according to their affix. Nouns ending in vowels use

the plural suffix -s, while nouns ending in consonants use the suffix -es. Plurals were also annotated for their grammatical gender by a native Spanish speaker. Irregular nouns, misspellings, and words not listed in the Real Academia Española (RAE) online dictionary were excluded.

3.2 Identifying Tokenization Type

We created three lists of plurals: one-token (n=1247), multi-token morphemic (n=508), and multi-token non-morphemic (n=627). One-token plurals are stored as single tokens in the tokenizer's vocabulary. We then categorized multi-token plurals as morphemic or non-morphemic. If tokenization followed morpheme boundaries (e.g., *naranjas* as ['naranja', '##s']), the noun was categorized as morphemic; if not, it was categorized as non-morphemic (e.g., *neuronas* is tokenized as ['neuro', '##nas']).

3.3 Relationship of Tokenization to Frequency

Using oral frequency measures for 2071 target plural wordforms available in a corpus of over 3M spoken words (Alonso et al., 2011), we examined the relationship between a wordform's frequency and how it was tokenized. A linear model predicting Log Frequency from Tokenization Scheme explained significant variance $[R^2 = 0.33]$. With MORPHEMIC level as a reference class (i.e., intercept), the NON-MORPHEMIC plural nouns were significantly less frequent $[\beta = -0.18, SE = 0.03, p < .001]$, while the SINGLE-TOKEN plural nouns were significantly more frequent $[\beta = 0.59, SE = 0.03, p < .001]$. As expected, the frequency of a wordform was likely a major factor in how it was tokenized (see also Appendix A.2).

Due to the relationship between tokenization scheme and wordform frequency, we carried out several supplementary analyses to determine the extent to which frequency was a confound in the results presented in Section 4. We found two key results: first, BETO's predictions were indeed more accurate for more frequent wordforms; second, however, BETO's predictions were still more accurate for some of the original tokenization schemes than others, even controlling for wordform frequency (see Appendix A.2 for details).

3.4 Artificial Tokenization Procedure

To investigate the effect of tokenizing a wordform at the morpheme boundary, we artificially tokenized single-token and multi-token non-

morphemic plural nouns by concatenating the token for the appropriate affix (e.g., "##es") onto the token(s) for the singular noun (Table 1).

Morpheme Boundary	Original Tokenization	Artificial Tokenization
mujer+es	['mujeres']	['mujer', '##es']
patrono+s	['patr, '##onos']	['patr, '##ono', '##s']

Table 1: Artificial tokenizations for the words *mujeres* 'women' (*mujer*), and *patronos* 'employers' (*patrono*).

4 Study: Article-Noun Agreement

Our primary research question concerned the impact of the original tokenization (TOKENIZATION SCHEME) on an article agreement task, similar to that implemented by Linzen et al. (2016). In Spanish, articles must agree with the *number* of the noun (e.g., *la mujer* vs. *las mujeres*); learned representations for the target noun should thus be conducive to predicting article number. We asked:

1. How does the initial tokenization scheme of a plural noun impact the language model's ability to predict the correct article?

2. Does our *artificial* tokenization scheme provide sufficient information to facilitate successful agreement?

3. How does the success of our artificial tokenization scheme compare to the original tokenization scheme for those nouns?

4.1 Method

Implementational details for the masked article prediction task are available in Appendix A.1. Agreement was assessed by taking the logarithm of the relative probability of a plural vs. singular article as predicted by a given noun. For a given wordform (e.g., *mujeres*), a positive log-odds indicated a higher probability was assigned to the plural article, while a negative log-odds indicated a higher probability was assigned to the singular article. A *singular* noun should be associated with a more negative log-odds, while a *plural* noun should be associated with a more positive log-odds. We considered both DEFINITE and INDEFINITE articles (ARTICLE TYPE) for each wordform; the log-odds calculation was performed separately for each type.

Accounting for the different presentations of each wordform (i.e., definite vs. indefinite article; original vs. artificial tokenization), our final

dataset had 13,276 observations in total, each with an accompanying *log-odds* ratio. All data and visualizations were analyzed in R; mixed effects models were fit using the *lme4* package (Douglas Bates et al., 2015). Maximal random effects structures were fit where possible, and reduced as needed for model convergence.

4.2 Results

4.2.1 Impact of Initial Tokenization

We first asked whether the original tokenization scheme used for plural nouns affected successful agreement. We fit a mixed model with Log Odds as a dependent variable, fixed effects of Tokenization Scheme and Word Number (and an interaction between the two), fixed effects of Article Type, and random intercepts for each word lemma and sentence. This model explained significantly more variance than a model omitting only the interaction $[\chi^2(2) = 6.54, p = .04]$, suggesting that different tokenization schemes were differentially successful in predicting the appropriate article. Note that this interaction was independent from the effect of wordform frequency (see Appendix A.2).

However, as depicted in Figure 1, this effect was quite small. Accuracy was near ceiling for all tokenization types, i.e., the Log Odds was larger than 0 for plural nouns and smaller than 0 for singular nouns (see also Table 2). Thus, our results do not suggest that morphologically-aligned tokenization is required for good agreement performance.

Original Tokenization	Original	Artificial
Morphemic	0.97	—
Non-morphemic	0.98	0.96
Single-Token	0.98	0.97

Table 2: Accuracy scores for *plural nouns* only, using either the original tokenization scheme for that class of nouns or the artificially-induced morphemic scheme.

4.2.2 Success of Artificial Tokenization

Next, we artificially tokenized plural nouns that would otherwise be tokenized non-morphemically or as a single-token. To quantify the success of this procedure, we fitted a linear mixed-effects model predicting Log Odds with fixed effects of Article Type, Word Number, Tokenization Scheme, and Affix ('##s' or "##es"), as well as random intercepts for word lemma and sentence.

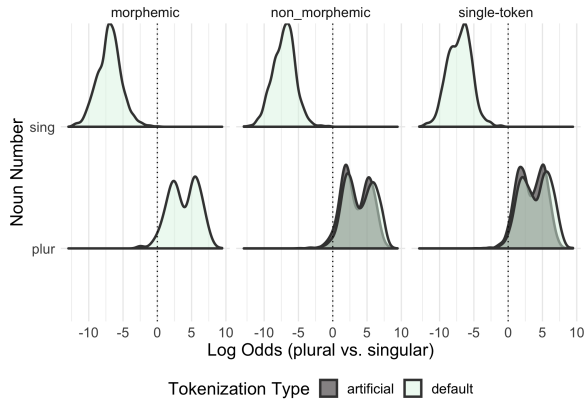This model explained significantly more variance than a model omitting only Word Number

Figure 1: Log-odds varied significantly as a function of noun number (*singular* vs. *plural*). The extent of this variance interacted (weakly) with initial tokenization (*morphemic* vs. *non-morphemic* vs. *single-token*) and with whether the *original* or *artificial* tokenization procedure was used. Larger log-odds indicate higher probabilities of the plural article.

$[\chi^2(1) = 11988, p < .001]$, indicating that the artificial tokenization procedure still led to good article number agreement performance: Log Odds were significantly different for singular nouns and artificially-tokenized plural nouns (see also Figure 1 and Table 2).

4.2.3 Comparing Default vs. Artificial Tokenization Schemes

Finally, restricting our analysis to plural forms, we asked whether a higher Log Odds was assigned to *artificially tokenized* plural nouns than ones using the default scheme. We fitted a linear mixed-effects model with fixed effects of Tokenization Scheme (artificial or original), Affix, and Original Tokenization Scheme (as well as random intercepts for word lemma, sentence, and wordform, and by-lemma random slopes for Tokenization Scheme). This model did explain more variance than a model omitting only Tokenization Scheme $[\chi^2(1) = 141.81, p < .001]$. Critically, however, the Log Odds for the artificially tokenized plural nouns was *lower* ($M = 3.38, SD = 2$) than when using the default tokenization ($M = 3.95, SD = 2.15$). In other words, the artificially-induced morphemic tokenization was successful, but less so than relying on the original scheme for those nouns.

5 Linear Discriminant Analysis (LDA)

To identify potential causes for the observed agreement patterns across noun types (singular vs. different plural tokenizations), we considered the em-
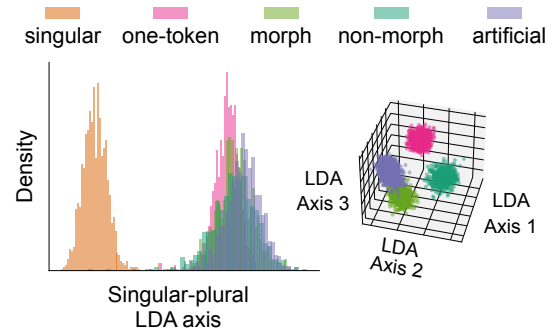

Figure 2: LDA for singular and plural embeddings reveals axes of overlap (*left*) and discriminability (*right*) for differentially tokenized plural forms.

beddings of those nouns in the language model representation space. We took each noun's mean embedding across the last four (out of twelve) BETO Transformer layers, averaging over all tokens in the noun. To minimize confounds from averaging embeddings over different numbers of tokens, we considered only two-token plurals in all multi-token scenarios for embedding analyses.

We first identified the linear axis that maximally separated single-token singular from plural nouns. To do this, we ran linear discriminant analysis (LDA) with two classes of embeddings: singular nouns (all single-token) and single-token plural nouns.[2] We then projected all noun representations linearly onto this axis, essentially projecting each embedding into a single value. As expected, we found that singular nouns clustered separately from plural nouns (Figure 2, *left*). Notably, all types of plurals (single-token, artificially tokenized, two-token morphemic, and two-token non-morphemic) patterned together and were not linearly discriminable along this axis. This suggests that the model could rely on similar number agreement mechanisms for different types of plurals, but future work would need to demonstrate causal impacts of this singular-plural axis on number agreement predictions (e.g. as in Mueller et al., 2022).

While the singular-plural LDA axis mapped different plural types to similar values, other axes could separate embeddings for the different plural types. We used LDA to identify the three linear axes that maximally separated the four types of plurals. As shown in Figure 2 (*right*), single-token plurals and two-token non-morphemic plu-

---

[2]Given n sets of representations, LDA computes $n - 1$ directions in the language model representation space that maximize separation between the sets.

rals were separable from one another and from all other plural types. The artificial and default morphemic plurals had distinct clusters, but they were not entirely separable from one another. This indicates that even though the artificial tokenization was never seen by the model during training, the representations were still quite similar (e.g. due to the presence of the '##s' or '##es' token). The slight separation between these clusters may be driven either by frequency effects or by veridical differences in how the models represent number in the two plural types.

6 Discussion and Conclusion

We assessed whether distinct tokenization schemes impacted the ability of BETO (a Spanish language model) to predict appropriate articles for Spanish plural nouns. Single-token representations facilitated slightly better predictions overall. However, the model did show evidence of generalization consistent with having learned morpheme-like "rules": artificially re-tokenizing plural nouns along morpheme boundaries produced representations amenable to article prediction—despite the language model never having previously observed that sequence of tokens (see Figure 1)—though this approach was slightly less accurate than relying on the original tokenization scheme. This provides further insight into work on language models generalizing morphological patterns (Haley, 2020); however, this does not work equally well for all languages or models (Weissweiler et al., 2023).

Notably, the similar agreement performance across single-token, morphological, non-morphological, and artificially-tokenized plurals could indicate multiple different agreement mechanisms in the model. At least on this task, tokenization along morpheme boundaries was not correlated with improved agreement performance; this is in contrast to other work suggesting that morphologically aware tokenization improves performance, e.g., in machine translation (Macháček et al., 2018) or similarity judgments (Batsuren et al., 2024). Future work might apply causal interventions on different embedding axes (as found in §5), to determine the extent to which the same model subnetworks are involved in number agreement for different types of plural tokenizations, shedding light on the impacts of tokenization on language model processing.

7 Limitations

A key limitation of the current work is scope. Future work could consider additional morphological phenomena, additional languages, and a larger range of language models or tokenization schemes. A second limitation is that the language model's performance was near-ceiling for each category considered. It is possible that different tokenization strategies do in fact impact agreement performance under more challenging conditions, but that the near-ceiling performance on this task made it difficult to detect those differences. Future work could work to develop more challenging tasks for which the model is not at ceiling (as in Linzen et al., 2016), or for which variance in how multimorphemic words are parsed might be expected to contribute more to downstream performance (Batsuren et al., 2024). Finally, our work does not demonstrate the extent to which different tokenizations rely on the same internal mechanisms for agreement in the model (§6), which is a valuable direction for future work.

8 Acknowledgements

References

Héctor Martínez Alonso and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, 57.

María Angeles Alonso, Angel Fernandez, and Emiliano Díez. 2011. Oral frequency norms for 67,979 Spanish words. *Behavior Research Methods*, 43:449–458.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

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*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

MM Douglas Bates, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. *arXiv preprint arXiv:2004.03032*.

Coleman Haley. 2020. This is a BERT. Now there are several of them. Can they generalize to novel words? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 333–341.

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, Online. Association for Computational Linguistics.

Haris Jabbar. 2024. MorphPiece: A linguistic tokenizer for large language models. *arXiv*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Dominik Macháček, Jonáš Vidra, and Ondřej Bojar. 2018. Morphological and language-agnostic word segmentation for nmt. In *International Conference on Text, Speech, and Dialogue*, pages 277–284. Springer.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Aaron Mueller, Yu Xia, and Tal Linzen. 2022. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 95–109, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China. Association for Computational Linguistics.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for Turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Michael T Ullman. 2016. The declarative/procedural model: A neurobiological model of language learning, knowledge, and use. In *Neurobiology of language*, pages 953–968. Elsevier.

Omri Uzan, Craig W Schmidt, Chris Tanner, and Yuval Pinter. 2024. Greed is all you need: An evaluation of tokenizer inference methods. *arXiv preprint arXiv:2403.01289*.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Article Agreement Task: Model Inputs

Within the sequence of model inputs, only the article token was masked, and special tokens ([CLS], [SEP]) were included, as in the examples below:

- Example model inputs for original single-tokenizations: "[CLS] [MASK] mujeres [SEP]"

- Example model inputs for artificial (morphemic) tokenizations: "[CLS] [MASK] mujer ##es [SEP]"

- Example model inputs for original non-morphemic multi-tokenizations: "[CLS] [MASK] patr ##onos [SEP]"

- Example model inputs for artificial (morphemic) tokenizations: "[CLS] [MASK] patr ##ono ##s [SEP]"

For each sequence of inputs independently, we obtain BETO's output logits over the target token corresponding to the (1) definite singular, (2) indefinite singular, (3) definite plural, and (4) indefinite plural articles. We subsequently apply softmax normalization to each token's logits to obtain the log probabilities of filling the masked item with a particular article.

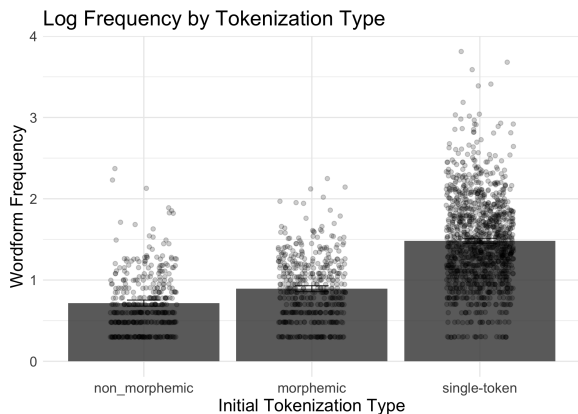## A.2 Supplementary Analysis with Log Frequency

Figure 3: Single-token plurals were significantly more frequent than those tokenized according to morphemic boundaries, which were more frequent than those tokenized according to non-morphemic substrings.

We ran a follow-up analysis asking whether the Log Frequency of a wordform was predictive of agreement success. This analysis had two key goals. First, because Log Frequency was correlated with Tokenization Scheme, we aimed to determine whether the effect of Tokenization Scheme on agreement success was in fact due to effects of token frequency. Second, we were independently interested in whether the language model made better predictions for more frequent wordforms.

We fitted a linear mixed-effects model including fixed effects of Tokenization Scheme, Word Number, and Log Frequency, as well as interactions between Word Number and Tokenization Scheme and between Word Number and Log Frequency. We also included random intercepts for word lemma and sentence. This model explained significantly more variance than a model omitting only the interaction between Log Frequency and Word Number $[\chi^2(1) = 17.89, p < .001]$. The interaction was negative $[\beta = -0.35, SE = 0.08, p < .001]$, i.e., the plural article log-odds were *more* negative for more frequent singular nouns. In other words, the language model made better predictions for more frequent nouns than less frequent nouns.

The full model also explained more variance than a model omitting the interaction between Word Number and Tokenization Scheme $[\chi^2(2) = 11.24, p = .004]$. This indicates that even controlling for wordform frequency, there was an independent effect of how the wordform was initially tokenized on the success of the language model's article predictions.