BEST methods for interpreting model performance on Spanish agreement prediction tasks

Julia Kundu **Madison Larter** **Ashlyn Winship** **Alison Yu**
Cornell University
{jk2578, mtl77, alw329, cy537}@cornell.edu

Abstract

Spanish morphemes explicitly mark number and gender, and these features must agree between nouns and the articles, verbs, and adjectives that are associated with them. However, language models often do not tokenize words along morpheme boundaries. Previous work (Arnett et al., 2024) shows that introducing an artificial tokenization scheme to BETO (Cañete et al., 2020) that tokenizes words in such a way that the plural morpheme is an individual token (e.g. *madre*: 'madre', 's'), ostensibly providing the model with plural information, does not improve the model's performance on article prediction tasks. This paper builds on Arnett et al.'s work to show similar results on verb and adjective prediction tasks, and extends all of these tasks to pseudoword and misspelled datasets, once again with similar results. Additionally, preventing the tokenizer from accessing non-plural final tokens does not improve performance on the article prediction task. Altogether, these results suggest that BETO's accuracy on plural prediction tasks does not rely on plural morpheme information that is accessible to humans.[1].

1 Introduction

Investigation of the ability of models like BETO to generalize grammatical rules like pluralization and subject-verb agreement may provide an increased understanding of models' linguistic intuition in comparison to humans. Previous work in the areas of psycholinguistics and child language acquisition has shown that increased awareness of morphology is correlated with increased reading, reading comprehension, and writing abilities (Carlisle, 2000; Deacon & Kirby, 2004; McCutchen & Stull, 2015). If models utilize morphological information similarly to humans, we might expect that providing a model with explicit morpheme information would improve model performance on tasks involving this information.

We replicate and further investigate the surprising finding in Arnett et al. (2024) that in Spanish, tokenization along the plural morpheme boundary is a viable strategy, but that it is not required for Spanish subject-article agreement and does not improve performance. We expand the testing to include subject-verb and subject-adjective agreement with both masked and unmasked articles and obtain similar results suggesting that even on more difficult tasks, tokenization according to morphological boundaries does not improve performance.

Since Arnett et al.'s proposed form of artificial tokenization is relatively simple and does achieve high accuracy on agreement tasks, it might be well-suited for noisy or low-resource language models which do not allow natural forms of tokenization along morpheme boundaries. We investigate this possibility by testing this scheme on noisy data in the form of both constructed pseudowords and character-replacement typos as in Matthews et al. (2024).

Finally, we explore a potential explanation for the lack of performance improvement of the artificial tokenization scheme by barring the tokenizer from using final tokens not corresponding to a plural affix. We find that under this restriction, the majority of words are tokenized along the same boundaries as the artificial scheme, which shows that this segmentation is present in the tokenizer but may not be preferred for reasons of efficiency or overall task performance.

[1] Arnett et al.'s (2024) original code and data available at: https://github.com/catherinearnett/spanish-plural-agreement
All code and data for this project are available at: https://github.com/jkjuliak/CL2-FinalProject

2 Background & Related Work

BETO (Cañete et al., 2020) is a Spanish variant of the BERT (Devlin et al., 2019) language model, pretrained on a large Spanish corpus comprised of diverse text sources including Wikipedia and numerous newspapers. It employs WordPiece tokenization (Wu et al., 2016), a subword-based method that splits words into smaller units if they do not appear in the model's vocabulary as a whole, allowing the model to handle a large array of Spanish vocabulary items—including inflected forms, named entities, and rare words more efficiently than a purely word-level model. The choice of WordPiece tokenization in BETO is rooted in the need to balance a manageable vocabulary size with broad lexical coverage. WordPiece splits are determined by frequency-based heuristics, aiming to maximize the likelihood of the pretraining data. As a result, morphologically meaningful boundaries (e.g., plural morphemes in Spanish) may or may not align with subword splits. This can pose challenges for modeling morphological agreement, as the information needed for accurate number or gender agreement may be strewn across multiple subwords, rather than being restricted to a single token that the model easily associates with specific morphological features.

The paper we are replicating builds on previous research investigating the relationship between tokenization strategies and language model performance, specifically in the context of number agreement tasks in Spanish. BETO, the Spanish adaptation of BERT, uses a SentencePiece tokenizer (Kudo & Richardson, 2018) with a subword-based vocabulary of 32,000 tokens. While this tokenization strategy enables BETO to generalize across many lexical and morphological variations, it does not guarantee alignment with morphological boundaries. For example, plural suffixes such as "-s" and "-es" may or may not be separated into distinct tokens, affecting the model's ability to encode number and gender agreement.

The replicated paper (Arnett et al., 2024) investigates whether aligning tokenization with morphological boundaries improves BETO's performance on grammatical agreement tasks, such as article prediction. The authors introduce an artificial tokenization scheme that forces morpheme-aligned splits for plurals, hypothesizing that this approach may improve the model's ability to predict morphosyntactic patterns. This approach seems to align with prior findings that morphologically-aware tokenization does not always yield performance improvements (T. A. Dang et al., 2024; Haley, 2020). The article-agreement tasks executed yielded near-ceiling performance across all categories, making it challenging to detect significant differences between tokenization schemes, opening the floor for others to expand the investigation to additional morphological phenomena, languages, and models to better understand the impact of tokenization strategies.

While highlighting grammatical agreement errors as a key challenge in BETO's performance while introducing their methods, the replicated study does not directly address this issue. Instead, their primary focus is on article-noun agreement, which is less context-dependent compared to subject-verb agreement. While article-noun agreement provides clues to how tokenization schemes affect morphosyntactic prediction, it does not account for the broader implications of tokenization errors in more complex grammatical contexts.

3 Replication & extension

To investigate whether morphemic tokenization would improve model performance, we replicated the findings from Arnett et al. (2024) and also conducted several extension experiments on similar tasks. Arnett et al. conducted Spanish word tokenization on one single task, article-noun agreement, using a Spanish BETO model. The model tokenized words in three ways: single-token, morphemic tokenization, and non-morphemic tokenization. For example, the word *patronos* would just have the token '*patronos*' if it received single-token tokenization. The morphemic scheme would tokenize it as '*patrono*' and '*s*', separating the plural morpheme from the stem. The non-morphemic scheme might tokenize it as '*patr*' and '*onos*'. The word is separated into two tokens, but not by its morpheme boundary. Each word receives one of the three tokenizations.

Arnett et al. (2024) then developed an artificial tokenization scheme that operates morphologically. It re-tokenizes the words that were tokenized by the single-token and the non-morphemic scheme by separating the plural morpheme from the last token. The single token 'patronos' would be re-tokenized as '*patrono*' and '*s*', while the

Category	Default	Artificial
Morphemic	0.97	-
Non-morphemic	0.98	0.96
Single-token	0.98	0.97

Table 1: Replication results from Arnett et al (2024) on subject-agreement task

Category	Default	Artificial
Morphemic	0.99	-
Non-morphemic	0.98	0.99
Single-token	0.96	0.97

Table 2: Accuracy for the verb prediction task with masked articles

non-morphemic '*patr*' and '*onos*' would become '*patr*', '*ono*', and '*s*'.

The authors' analysis of their results was conducted by obtaining log-odds, which will also be utilized in this paper. Plural nouns are associated with positive log odds, and singular nouns are associated with negative log odds. If the prediction of a noun has high log-odds, it suggests that this noun is more likely to be predicted as plural, and vice versa. The default schemes were compared against each other, and the artificial results were compared against the default ones. The results showed that all default schemes performed almost perfectly, with all of them having near-ceiling accuracies. The artificial scheme also performed well but was slightly less successful than the original ones. It was then concluded by Arnett et al. (2024) that morphemic tokenization does not significantly improve model performance.

3.1 Original task

The replication experiment was conducted using the code from Arnett et al. (2024). Our replication was highly successful. The results yielded were exactly the same as that of Arnett et al. (2024). All tokenization schemes reached over 95% in accuracy (shown in Table 1), with the performances of the artificial tokenization scheme being slightly lower than the default schemes. While Arnett et al. (2024) claimed that these results suggest that morphemic tokenization does not improve performance, we believe that these near-ceiling results do not provide sufficient evidence for such claim. The results of the default schemes are already almost perfect, and there is little room to improve.

3.2 Additional tasks

BETO's performance on the article prediction task is already near-ceiling, even before the artificial tokenization scheme is introduced (Arnett et al., 2024, p. 36). For this reason, we designed additional tasks on which to test the model: a verb prediction task and an adjective prediction task.

In the verb prediction task, the model is asked to predict the correct form of the verb *ser* ("to be"): singular *es*, or plural *son*. In the adjective prediction task, the model is asked to predict the correct form of the adjective *grande* ("big"), chosen because it takes the same form in both grammatical genders, unlike many Spanish adjectives: singular *grande*, or plural *grandes*. There are two versions of each task. In the first version, the model is provided with the correct article:

(1) {*el/la/los/las*} {noun token(s)} [MASK]

At the mask, the model is asked to predict the verb or adjective. In the second version, the article is also masked:

(2) [MASK] {noun token(s)} [MASK]

The model is asked to predict the verb or adjective at the second mask. We predict that the model may perform better on the unmasked version of both tasks, because it could make use of the number information provided by the form of the correct article, which it lacks in the masked version.

3.2.1 Verb prediction

Both the unmasked and the masked version of verb prediction task received near-ceiling results. The unmasked version received 100% accuracy in all tokenization schemes. The masked results of the task are listed in Table 2. By a tiny margin, the unmasked version performed better on the task. This aligned with our prediction, although this margin was not significant.

3.2.2 Adjective prediction

The two versions of the adjective prediction task both received 100% accuracy in all tokenization schemes. Since both the replication and the extension tasks had near-ceiling performances, more difficult tasks are required to test if morphemic tokenization would improve model performance.

4 Testing robustness to noise

The tasks tested in the previous section demonstrate that the artificial tokenization scheme does not significantly improve performance on multiple agreement prediction tasks when using real Spanish data. The tasks in this section seek to determine whether the artificial tokenization scheme can improve robustness to noisy data, by testing both pseudowords and misspelled Spanish words.

4.1 Pseudowords

We used the Python package Gibberish[2] to generate pseudowords ending in vowels. To simplify the agreement prediction tasks, the words were filtered to those ending only in "-a" (matched with feminine articles) or "-o" (matched with masculine articles). These words were then tokenized with BETO's original tokenizer. This resulted in a dataset consisting of 43 single-token pseudowords, 255 multi-token morphemic pseudowords, and 1368 multi-token non-morphemic pseudowords.[3] This dataset was used in the same tasks described in the previous section.

4.1.1 Results

Tables 3 – 7 report the accuracies of the default and artificial tokenizers on all of the tasks. To summarize, the artificial tokenizer does not significantly improve performance on the original article prediction task, or on most of the additional tasks, whether or not the article is unmasked. In many conditions throughout the tasks, performance with the artificial tokenization scheme is actually worse.

The exception is in the masked adjective task, for single-token words (see Table 7). Accuracy for the artificial tokenization is substantially higher on this task. However, we should recall the small number of single-token pseudowords compared to multi-token pseudowords: There are 5 times as many multi-token morphemic pseudowords, and 28 times as many multi-token non-morphemic pseudowords. It is possible that with a higher number of single-token pseudowords, the accuracy pattern would begin to resemble the results of the other tasks, all of which trend toward the artificial tokenizer not showing any significant improvement.[4]

Category	Default	Artificial
Morphemic	0.81	-
Non-morphemic	0.90	0.53
Single-token	0.61	0.33

Table 3: Accuracy: Pseudoword article task

Category	Default	Artificial
Morphemic	1.0	-
Non-morphemic	1.0	1.0
Single-token	0.99	1.0

Table 4: Accuracy: Pseudoword verb task - unmasked article

Category	Default	Artificial
Morphemic	0.78	-
Non-morphemic	0.89	0.30
Single-token	0.60	0.12

Table 5: Accuracy: Pseudoword verb task - masked article

Category	Default	Artificial
Morphemic	1.0	-
Non-morphemic	1.0	1.0
Single-token	1.0	1.0

Table 6: Accuracy: Pseudoword adjective task - unmasked article

Category	Default	Artificial
Morphemic	0.98	-
Non-morphemic	0.99	0.90
Single-token	0.69	0.84

Table 7: Accuracy: Pseudoword adjective task - masked article

4.2 Typos

Our second approach to reintroducing noise takes inspiration from the test vocabulary generation

[2]https://pypi.org/project/gibberish/

[3]This dataset is much smaller than the original Spanish dataset used by Arnett et al. 2024. This is the result of an apparent upper limit of the number of original pseudowords that Gibberish can produce. With a requested output of 300,000 pseudowords, ∼131,000 ended in "-o" or "-a". Of those, 1666 contained at least one duplicate in the dataset, suggesting that Gibberish began repeating pseudowords.

[4]The only other apparent improvement is in the single-token category of the unmasked verb task (see Table 4). However, the default performance is near-ceiling at 0.99, so the artificial improvement is not significant.

methods presented by Matthews et al., where typographical errors were reflected by randomly selecting a character in each word and replace it with another randomly chosen alphabetic character of the same case, producing "edited" words. For each noun existing in the Arnett et al. vocabulary, we randomly selected a character position, replacing it with another randomly selected character within the spanish alphabet, simulating a "typo", and rerunning the article, verb and adjective prediction tasks on such stimuli.

4.2.1 Results

Figures 8 - 11 report the log-odds (plural vs. singular) of the default and artificial tokenizers on all of the tasks. To summarize the data, the artificial tokenization scheme once again does not significantly improve performance on the original article prediction task, or on the other tasks that were defined. Figures 9-10 and 12-11 show a similar pattern of the original tokenizer yielding better performance on the verb/adjective task while unmasked and similar ($\leq$ 0.02) performance while masked. Similar performance for the masked-article tasks points to both tokenizers being equally as çonfused having decreased morphological information given the typo and the lack of an article to make a correct prediction. Unmasking the article reveals the original tokenizer's ability to integrate syntactical cues from the article to handle the noise for the corresponding noun.

An exception presents itself in the article prediction task in 8, where the artificial tokenizer beats the default in prediction performance. A possible explanation for this the nature of the task itself. For articles, the agreement patterns are simpler and less context-dependent compared to verbs or adjectives. The model only needs to recognize that the subject is plural or singular. Even if a typo alters tokenization, the artificial segmentation can sometimes isolate or highlight the morphological component (the affix) that signals plurality. This makes it easier to choose the correct article, paradoxically improving performance given irregular tokenization due to typos.

Category	Default	Artificial
Morphemic	0.83	0.91
Non-morphemic	0.85	0.94
Single-token	0.85	0.94

Table 8: Accuracy: Typo article task

Category	Default	Artificial
Morphemic	0.83	0.73
Non-morphemic	0.87	0.84
Single-token	0.86	0.58

Table 9: Accuracy: Typo verb task - masked article

Category	Default	Artificial
Morphemic	0.99	0.73
Non-morphemic	0.99	0.84
Single-token	1.0	0.58

Table 10: Accuracy: Typo verb task - unmasked article

Category	Default	Artificial
Morphemic	0.91	0.88
Non-morphemic	0.93	0.92
Single-token	0.93	0.80

Table 11: Accuracy: Typo adjective task - masked article

Category	Default	Artificial
Morphemic	1.0	0.99
Non-morphemic	1.0	0.92
Single-token	1.0	0.91

Table 12: Accuracy: Typo adjective task - unmasked article

5 Investigating the tokenizer

The original finding that a more linguistically motivated artificial tokenization scheme doesn't seem to improve BETO's performance on number agreement tasks (Arnett et al., 2024), which was reinforced by our additional investigation, doesn't match the intuition that explicitly providing plural morpheme information should allow the model to better predict plurals. We hypothesize that this lack of improvement could be at least partially due to this information already being present in the tokenizer.

Most tokenizers retain the ability to produce multiple tokenization possibilities for the same word (Provilkov et al., 2020), so it's possible that the segmentation corresponding to Arnett et al. (2024)'s artifical tokenization scheme is already stored in the tokenizer, but is not the most preferred tokenization strategy for overall model per-

formance. We decided to remove the final tokens of plurals for which the preferred tokenization didn't encode the plural morpheme '##s' or '##es' as a token and determine the tokenizer's next best option for how to segment these plurals.

5.1 Final token identification & deletion

We first identified all final tokens across the pretrained tokenizer's representation of the plurals in the original dataset not corresponding to the plural morpheme '##s' or '##es'. This included the entire word for all of the single-token plurals and the final tokens of each of the multi-token non-morphemic plurals. For the multi-token non-morphemic words, we hypothesized that there would be a larger subset of final tokens representing the model's interpretation of the plural morpheme. However, out of 646 plurals, there were 251 final tokens, which doesn't seem to support this idea. We then deleted all of these final tokens, including the single-token plurals, from the pretrained tokenizer's vocabulary.

5.2 Updated tokenizations

After identifying and deleting all final tokens across these plurals except for the plural morpheme '##s' or '##es', we re-encoded all of the plurals in the single-token and multi-token non-morphemic categories with this updated tokenizer vocabulary. We found that across the 1363 originally single-token plurals, the segmentation of 1308 exactly matched the artificial tokenization scheme, and the same was true for 577 of the 646 originally multi-token non-morphemic plurals.

Of the 55 originally single-token plurals for which the new tokenization did not match the artificial scheme, 20 still had the plural affix '##s' (1 had the plural affix '##es'). Of the 69 originally multi-token non-morphemic plurals for which the new tokenizations did not match the artificial scheme, 57 had the plural affix '##s' and 2 had the plural affix '##es'.

5.3 Impacts on performance

We tested the model on the original article prediction task presented by Arnett et al. (2024) after modifying the tokenizer to investigate the impacts of the new tokenizations on the model's number agreement performance. We found that the model's performance was not significantly impacted by the modification to the tokenizer, with prediction accuracy closely matching the model's accuracy with the unmodified tokenizer as well as Arnett et al.'s artificial tokenization scheme (Table 13, Figure 1).

Category	Default	Artificial	New
Morphemic	0.97	–	0.98
Non-morphemic	0.98	0.96	0.97
Single-token	0.98	0.97	0.98

Table 13: Accuracy for plural nouns using the original tokenization, artificial scheme, and modified tokenization (after final token deletion).

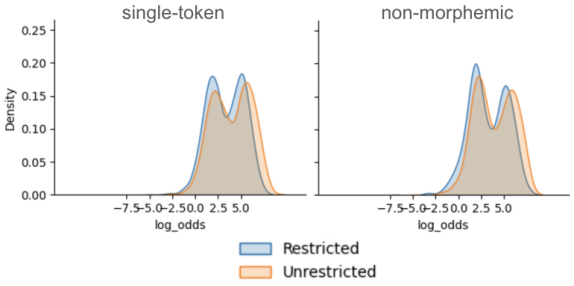

Figure 1: Log-odds (plural vs. singular) for originally single-token and multi-token non-morphemic plurals before (unrestricted) and after (restricted) tokenizer modification.

This seems to suggest that the plural morpheme is already recognized by the model, and that the default tokenizations that don't separate the plural morpheme as a token are simply chosen for overall model performance on other language tasks that do not explicitly involve number agreement. It could also at least partially explain why the artificial tokenization scheme did not improve model performance, as the new linguistic information it supposedly provided to the model in the form of the plural affix was already present.

6 Discussion

The purpose of this paper was to investigate whether the tokenization scheme developed by Arnett et al. (2024), which seems intuitive to humans, could improve BETO's performance in any meaningful way. Their original study found that it could not improve performance on an article prediction task. Our additional work shows that it does not improve performance on verb or adjective prediction tasks, either. The model does not appear to be using the kind of plural information that is available to humans through morphological knowledge. This is supported by the similarity

between the results of the masked and unmasked versions of the verb and adjective prediction tasks: Giving the model explicit number information in the form of an unmasked singular or plural adjective does not improve the model's performance. It is also further supported by the fact that the tokenization scheme has no effect on the model's robustness to noise, whether that noise is introduced through randomly generated pseudowords, or through a more realistic misspelling schematic.

The investigation conducted in section 5 suggests that this kind of morphological number information is simply not useful to the model. Removing the original suffixes from the tokenizer's vocabulary resulted in a new set of tokens that very closely matches those created by the artificial tokenization scheme – implying that the artificial tokenization scheme is very nearly the model's "next best" option for tokenizing. This shows that the tokenizer does possess the ability to split words up in this way. For whatever reason, however, it is not the most useful tokenization scheme for the model. The tasks we have focused on in this investigation are solely concerned with number agreement in very short, 2- to 3-word phrases. Models such as BETO must perform well on a wide variety of tasks. For instance, we asked the model to predict the correct form of *grande(s)* for every noun in the Spanish dataset, without regard for whether that adjective semantically (or even syntactically) makes sense in combination with any given noun in that dataset. A model must not only predict the correct morphological agreement for factors such as number and grammatical gender, but the words it predicts must also make sense in context. The results of this investigation in combination with the work done by Arnett et al. (2024) suggest that it may be unreasonable to expect models to extract the same knowledge or usefulness from morphology that a human being does. This difficulty with morphology is reflected in some current research (cf. Ismayilzada et al. (2024), A. Dang et al. (2024), Moisio et al. (2024)), while other studies show that morphological information *does* improve performance on certain tasks (cf. Vania et al. (2018), Vania (2020)). More work on probing models for morphological encoding is needed to paint a fuller picture of whether and to what extent models make use of morphological information.

Limitations

Perhaps the largest limitation with this work is still the scope. Our work addressed only plural agreement (although with several tasks of varying difficulty) and regular plural morphology in Spanish. We also tested only a single verb and a single adjective for each of those agreement tasks in an attempt to avoid introducing additional Spanish morphological phenomena into the mix that could complicate our results. Future work investigating the impacts of tokenization on downstream task performance could include additional or more complex morphological phenomena and a broader set of more difficult tasks.

In modifying the tokenizer vocabulary, we found that the artificial tokenization scheme was the next best segmentation option for the majority of the plurals in the dataset. It could be informative to investigate tokenization schemes that are not already present in the tokenizer of the model being used (or at least further down in preference order) to determine whether schemes that introduce more novel word segmentation information can improve model performance on tasks involving particular morphemic characteristics. We also used a pre-trained model and tokenizer with no additional fine-tuning for all of the tasks (as in Arnett et al. (2024)) - future work could investigate whether fine-tuning a model on particular morphemic information impacts downstream performance on linguistic tasks that involve this information.

References

Arnett, C., Rivière, P. D., Chang, T. A., & Trott, S. (2024). Different tokenization schemes lead to comparable performance in spanish number agreement. *arXiv preprint arXiv:2403.13754*.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pretrained bert model and evaluation data. *PML4DC at ICLR 2020*.

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing: An Interdisciplinary Journal, 12: 169-190*.

Dang, A., Raviv, L., & Galke, L. (2024). Morphology matters: Probing the cross-linguistic morphological generalization abilities of

large language models through a wug test. *13th edition of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2024)*, 177–188.

Dang, T. A., Raviv, L., & Galke, L. (2024). Tokenization and morphology in multilingual language models: A comparative analysis of mt5 and byt5. https://arxiv.org/abs/2410.11627

Deacon, S. H., & Kirby, J. R. (2004). Morphological awareness: Just "more phonological"? the roles of morphological and phonological awareness in reading development. *Applied Psycholinguistics, vol. 25, 2: 223–238*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805

Haley, C. (2020, November). This is a BERT. now there are several of them. can they generalize to novel words? (A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, & H. Sajjad, Eds.) [Presented at the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP]. https://doi.org/10.18653/v1/2020.blackboxnlp-1.31

Ismayilzada, M., Circi, D., Sälevä, J., Sirin, H., Köksal, A., Dhingra, B., Bosselut, A., van der Plas, L., & Ataman, D. (2024). Evaluating morphological compositional generalization in large language models. *arXiv preprint arXiv:2410.12656*.

Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. https://arxiv.org/abs/1808.06226

Matthews, J. A., Starr, J. R., & van Schijndel, M. (2024). Semantics or spelling? probing contextual word embeddings with orthographic noise. *arXiv preprint arXiv:2408.04162*.

McCutchen, D., & Stull, S. (2015). Morphological awareness and children's writing: Accuracy, error, and invention. *Reading and Writing, vol. 28, 2: 271-289*.

Moisio, A., Creutz, M., & Kurimo, M. (2024). Llms' morphological analyses of complex fst-generated finnish words. https://arxiv.org/abs/2407.08269

Provilkov, I., Emelianenko, D., & Voita, E. (2020). Bpe-dropout: Simple and effective subword regularization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1882–1892*.

Vania, C. (2020). On understanding character-level models for representing morphology.

Vania, C., Grivas, A., & Lopez, A. (2018). What do character-level models learn about morphology? the case of dependency parsing. *arXiv preprint arXiv:1808.09180*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. https://arxiv.org/abs/1609.08144