Computational Morphology Partial Task

Adrián Cuadrón Cortés
acuadron001@ikasle.com

1 Introduction

In this work, I will summarize the paper ”Different Tokenization Schemes Lead to Comparable Performance in Spanish Number Agreement” (Arnett et al., 2024). The authors investigate how various tokenization schemes affect number agreement in Spanish plurals.

Tokenizers play an important role in the field of NLP, as they break text into smaller units called tokens. Assigning a unique token to each word can help achieve a precised semantic representation. However, this approach may lead to a less robust model for unseen words. While tokenization can sometimes align with morphological boundaries (e.g *cars* → [’car’, ’##s’]), this is not always guaranteed. A more effective practice for capturing morphosyntactic patterns is segmenting words into their lemmas and morphemes (Ullman, 2016).

The authors of the paper examine how different tokenization strategies influence model predictions for plural agreement in Spanish. They explore three approaches to tokenizing Spanish plurals: single-token representation, morphemic tokenization, and non-morphemic tokenization. Their focus is on cases where the plural is formed in the simplest way. In Spanish, plurals are typically marked by appending one of two common morphemes, -s or -es, to the singular form. The study looks at how the tokenization of Spanish nouns influences language model predictions for specific morphosyntactic rules, following (Batsuren et al., 2021).

2 Data

The data used in their experiments were extracted from the AnCora Treebanks (Martinez Alonso and Zeman, 2016). This corpus consists of 17,662 sentences, 547,558 tokens, and 560,137 syntactic words. The authors extracted the singular form lemmas along with their corresponding plural nouns. The plurals were categorized according to their affix: nouns ending in vowels take the plural suffix -s, while those ending in consonants use -es. Additionally, a native Spanish speaker annotated the grammatical gender of the plurals. Certain words were excluded, including irregular nouns, misspellings, and words not listed in the Real Academia Española (RAE) online dictionary.

There are 3 different datasets for each type of tokenization. These are *csv* files that contains different columns that have information of each word.

For the single-token dataset, there are these columns: whole_word (contains the word), root (the lemma of the word) and affix (the plural suffix applied). Here, they know that the word is tokenized as the whole_word.

Whole_Word	**Root**	**Affix**
reservas	reserva	s
sectores	sector	es

For the multi-token morphemic dataset, there are these columns: whole_word (contains the word), root (the lemma of the word) and affix (the plural suffix applied). Here, they know that the word is tokenized in the morpheme boundary, so it is like [root,affix].

Whole_Word	**Root**	**Affix**
remedios	remedio	s
gestores	gestor	es

For the multi-token non-morphemic dataset, there are four columns: whole_word (contains the word), root (the lemma of the word), affix (the plural suffix applied) and the tokenized (how the word was tokenized). Here, it is important to know exactly how the word was tokenized by the tokenizer.

Whole_W	Root	Af.	Tokenized
patronos	patrono	s	[patr, ##onos]
bebedores	bebedor	es	[bebe, ##dores]

3 Model

The authors used a Spanish pre-trained BERT model, called BETO (Cañete et al., 2023). It was trained similar in size to BERT-Base, with 12 self-attention layers, 12 attention heads each, and a hidden size of 768, totaling 110M parameters. The training corpus, containing about 3B words, was sourced from Wikipedia, the OPUS Project (Tiedemann, 2012), and other Spanish-language materials. A vocabulary of 32K tokens was created using the SentencePiece byte pair encoding algorithm (Kudo and Richardson, 2018). They used techniques like Dynamic Masking, Whole-Word Masking (WWM), and larger batch sizes compared to the original BERT. The training was done in two phases with a total of 2 million steps using Google's TPU v3-8.

4 Tokenization procedures

4.1 Tokenization type

The authors created three lists of plurals:

- **One-token plurals** (n=1247): These are stored as single tokens in the tokenizer's vocabulary, such as:
  - ”reservas” as [‘reservas’]
- **Multi-token morphemic plurals** (n=508): These plurals follow morpheme boundaries in tokenization, such as:
  - ”remedios” as [‘remedio’, ‘##s’]
- **Multi-token non-morphemic plurals** (n=627): These plurals do not follow morpheme boundaries in tokenization, such as:
  - ”patronos” as [‘patr’, ‘##onos’]

4.2 Relation of Tokenization and Frequency

The authors conducted an analysis of the relationship between word frequency and how they were tokenized. For this, they used oral frequency data of 2,071 spoken words from a corpus of over 3 million words. In other words, the plural words in this corpus were analyzed based on their frequency in speech (Alonso et al., 2011).

A linear model was built to predict the logarithmic frequency of a word based on its tokenization scheme. This means that the model aimed to explain the frequency of words based on how they were tokenized. The model explained a significant portion of the variability in the data ($R^2 = 0.33$, indicating that the model accounted for 33% of the variability in frequency based on tokenization). They used morphemic tokenization as the reference class for the analysis. They had reached the conclusion that the frequency of a wordform likely played a significant role in determining how it was tokenized.

After doing additional analyses they found two key points: first, BETO made more accurate predictions for more frequent wordforms; second, BETO still performed better with some tokenization schemes than others, even after accounting for word frequency.

4.3 Artificial Tokenization

The authors artificially tokenize the singular forms of the nouns at the morpheme boundary for both single-token and multi-token non-morphemic plurals. To achieve this, they concatenate the suffix with the singular form of the noun.

For example, for the single-word type, the word *reservas* is originally tokenized as ['reservas'] and artifically tokenized in the morpheme boundary as ['reserva', '##s']. Also for the multi-token non-morphemic type, for example, *patronos* originally tokenized ['patr', '##onos'] is artifially tokenized as ['patr', '##ono', '##s'].

5 Method

The primary research question focused on the impact of the original tokenization scheme on article agreement in Spanish, specifically how it affects a language model's ability to predict the correct article. The study also explored whether the artificial tokenization scheme provides sufficient information for successful agreement and how its performance compares to the original tokenization scheme.

First, the authors processed three types of plural noun representations (single-token, morphemic multi-token, non-morphemic multi-token). Then, they created sentence templates with masked articles. For the task, in each input sequence, the article (e.g., ”la”, ”las”, etc.) was hidden by masking it ([MASK]) and the special tokens were added ([CLS], [SEP]). For example:

Single-token: “[CLS] [MASK] mujeres [SEP]”
Non-morphemic: “[CLS] [MASK] patr ##onos [SEP]”
Morphemic: “[CLS] [MASK] remedio ##s [SEP]”
Artificial tok.: “[CLS] [MASK] patr ##ono ##s [SEP]”

After masking the article, the language model (BETO) outputs logits (raw scores) for each of the four possible article types:

1) Definite singular (e.g., ”la”)
2) Indefinite singular (e.g., ”una”)
3) Definite plural (e.g., ”las”)
4) Indefinite plural (e.g., ”unas”).

To convert the logits into probabilities, the softmax function is applied. This function turns the raw output into a set of probabilities, showing how likely each article is to fill the masked position in the sentence.

This code calculates surprisal values for the predicted probabilities of articles. Surprisal, defined as *-log(p)*, quantifies how unexpected or surprising an event is, with higher values indicating lower probabilities and greater unexpectedness. This allows for a detailed analysis of how well the model aligns with grammatical expectations, providing insights into the predictability of definite and indefinite articles in different linguistic contexts.

Agreement was measured by computing the log of the ratio of the predicted probabilities for plural vs singular articles for each noun. A positive log-odds meant the plural article was more likely, while a negative log-odds meant the singular article was more likely. Singular nouns were expected to have more negative log-odds, and plural nouns more positive log-odds. This calculation was done separately for both definite and indefinite articles.

Taking into account the different versions of each wordform, the final dataset consisted of 13,276 observations, each with a corresponding log-odds ratio. All data and visualizations were analyzed in R.

6 Results

The way plural nouns were tokenized had an impact on how well agreement worked. To figure this out, a mixed-effects model was used with Log Odds as the main variable. This model showed that it explained more variability ($X^2 = 6.54$), meaning that different tokenization approaches had varying levels of success. Overall, all tokenization types performed well, with Log Odds being greater than 0 for plurals and less than 0 for singulars.

Two approaches were evaluated: the original tokenization scheme specific to plural nouns and an artificially-induced morphemic tokenization scheme.

For morphemic tokenization, the accuracy scores are as follows: 0.97 for the original class and 0.98 for the non-morphemic class. In the case of non-morphemic tokenization, the scores are 0.98 for the original class and 0.96 for artificially morphemic schemes. Finally, for the single-token method, the accuracy achieved was 0.98 for the original scheme and 0.97 for the artificially morphemic scheme. These results highlight the comparative performance of the tokenization approaches for handling plural nouns. This shows that different tokenization methods were pretty accurate at predicting the right article. However, the results suggest that aligning tokenization with morphology is not really necessary to get good agreement performance.

To measure the success of the artificial tokenized procedure, the authors used a linear mixed-effects model to predict Log Odds, incorporating fixed effects such as Article Type, Word Number, Tokenization Scheme, and Affix (e.g., ”##s” or ”##es”). The model explained a significant amount of variance, showing that these factors influenced the outcomes. Despite relying on an artificial tokenization procedure, the analysis revealed strong performance in achieving accurate article-number agreement.

They focused on plural nouns and compared the effectiveness of artificial tokenization versus the default method. A mixed-effects model showed that including the tokenization scheme explained more variance. However, the artificial tokenization had lower Log Odds (M = 3.38) compared to the original tokenization (M = 3.95), indicating that while the artificial tokenization was successful, it was less effective than the default scheme.

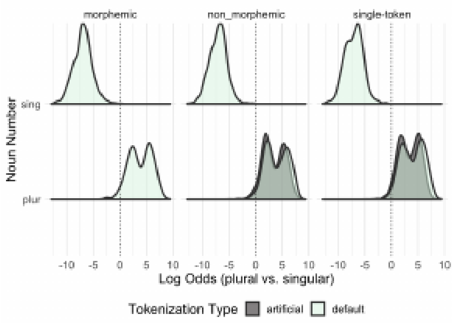

Figure 1: Log Odds for each type of tokenization.

7 LDA

They use embeddings to identify patterns in noun types by taking the mean embedding of the last four layers of the transformer model, BETO. For their analysis, they focus only on two-token plurals.

First, they created a plot to separate the linear axis of the single-token singular and plural forms. To do this, they applied Linear Discriminant Analysis (LDA) with two classes of embeddings (single-token singular and plural). They projected the embeddings into a single value and observed that distinct clusters formed for singular and plural nouns. However, all plural types (single-token, morphemic, non-morphemic, and artificial) clustered together and could not be separated along a single axis.

Therefore, they used LDA in 3D to identify three linear axes that successfully separated the different plural types. The figures show that single-token and non-morphemic plurals are separable from the other plural types. Artificial and default morphemic plurals are also separable, but they cannot be distinguished from each other. Although artificial tokenization was not seen during training, the representations are similar due to the plural marker tokens (-s, -es).

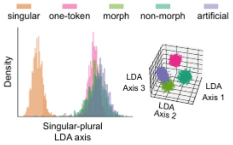

Figure 2: LDA applied to singular and plural embeddings shows overlapping patterns (left) and distinctions (right) for plural forms with different tokenizations.

8 Conclusions

This study explored the impact of different tokenization schemes on BETO, a Spanish language model, in predicting appropriate articles for plural nouns. The results showed that single-token representations slightly improved predictions, but artificial re-tokenization along morpheme boundaries also allowed the model to generalize and predict articles, despite never having seen those token sequences in training. However, this approach was slightly less accurate than the original tokenization scheme. The similar performance across single-token, morphological, non-morphological, and artificially-tokenized plurals suggests that multiple mechanisms may be at play in the model's agreement process. Interestingly, tokenization along morpheme boundaries did not necessarily improve agreement performance, contrary to findings in other tasks like machine translation and similarity judgments.

One limitation of this work is its scope, as future research could examine additional morphological phenomena, languages, and tokenization schemes. Another limitation is that the model's performance was near-ceiling, making it difficult to detect the impact of tokenization strategies in more challenging tasks. Additionally, this study does not explore the internal mechanisms involved in agreement for different tokenizations, which remains a valuable direction for future research.

9 Implementation

The data and code of the authors experiment is uploaded in github (`https://github.com/catherinearnett/spanish-plural-agreement`), so I was able to reproduce the experiments.

First, I ran the Python script *run_task_article-agreement.ipynb* to generate the model's article predictions and create a new *CSV* file containing the data and the probabilities for each article associated with each word.

Next, I reviewed the R code used for analyzing these results (*article_agreement_analysis.Rmd*). The experiments and analyses conducted by the authors were well executed and provided valuable insights.

I then decided to focus on the LDA section, as it was an area where I could make modifications and gain a deeper understanding of the process. To achieve this, I extracted the relevant code from *lda_plots.py* and created a new Python notebook to run it step by step. It is worth noting that I had to make several corrections and improvements to the original code, as it contained some errors.

As a result, I was able to successfully reproduce their experiments and generate the plots presented in the original paper.

9.1 New experiments

However, I wanted to contribute something original and create my own experiments. After reviewing the paper, I noticed that the authors assumed the gender of the article was correctly annotated by experts. This led me to a new idea: testing whether different tokenization strategies might influence how the model predicts the gender of the article.

For this purpose, I created a new Python notebook to code and explain each step. These all experiments are available on Github: `https://github.com/AdrianCuadron/gender_and_tokenization`

First, two different mini-datasets were created: one for feminine words and another for masculine words. In each dataset, three types of tokenization were generated, as in the original paper (single-token, morphemic, and non-morphemic) and we put 10 words for each one.

Then, we tokenize the sentence with the [MASK] token. The tokenizer converts the input sentence into tokens and returns a tensor that can be fed into the model. Then, we get the logits (raw scores) from the model. The model processes the input sentence and outputs logits, which are raw scores for every token in the vocabulary, indicating how likely each token is to fill the [MASK] position.

After extracting the logits corresponding to the [MASK] token's position, I apply softmax to convert them into probabilities. The softmax function converts the logits into a probability distribution, where each value represents the likelihood of a specific token filling the [MASK] position. Here, we extract probabilities for the expected articles. The code selects only the probabilities for the specific articles we are interested in (e.g., ”la”, ”el”, ”una”, ”un”). Finally, we calculate surprisal for each article.

In each instance of each dataset, the model predicted the token [MASK], and we recorded the probabilities for each gender based on the corresponding articles. Then, we calculated the Log Odds to determine whether the word was feminine or masculine. The Log Odds is calulated as log(prob_fem / prob_masc). A positive Log Odds indicates a feminine prediction, while a negative Log Odds indicates a masculine prediction.

With all the collected data, we checked if there were any words for which the model incorrectly predicted the article. Additionally, a box plot and a histogram were created to visualize the data distribution. This plots can be seen in A.

In conclusion, the single-token tokenization is clearly the most effective for correctly predicting the gender, as it consistently shows high and positive Log Odds values for feminine and negative values for masculine, reflecting accurate and reliable predictions. In contrast, the morphemic and non-morphemic tokenizations exhibit much lower and more scattered values, with a significant number of incorrect predictions, possibly due to the loss of important contextual information when words are split into subunits.

References

María Angeles Alonso, Angel Fernandez, and Emiliano Díez. 2011. Oral frequency norms for 67,979 spanish words. *Behavior Research Methods*, 43(2):449–458.

Catherine Arnett, Tyler Chang, and Sean Trott. 2024. Different tokenization schemes lead to comparable performance in Spanish number agreement. In *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–38, Mexico City, Mexico. Association for Computational Linguistics.

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.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data.

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.

Hector Martinez Alonso and Daniel Zeman. 2016. Universal Dependencies for the AnCora treebanks . *Procesamiento del Lenguaje Natural*, (57).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Michael T. Ullman. 2016. Chapter 76 - the declarative/procedural model: A neurobiological model of language learning, knowledge, and use. In Gregory Hickok and Steven L. Small, editors, *Neurobiology of Language*, pages 953–968. Academic Press, San Diego.

A Appendices

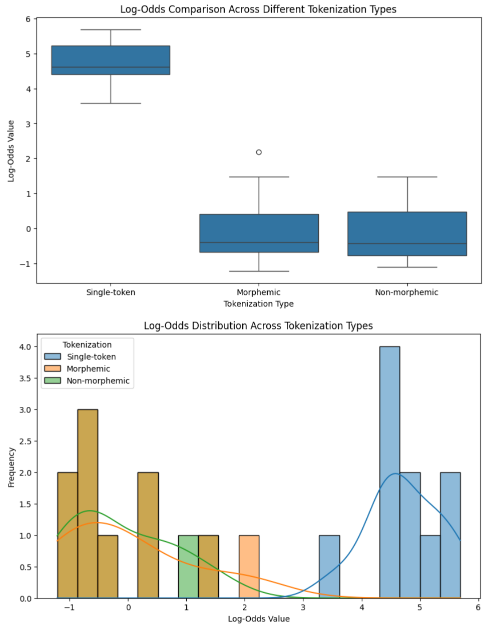

Figure 3: Boxplot and histogram for the feminine dataset Log Odds.

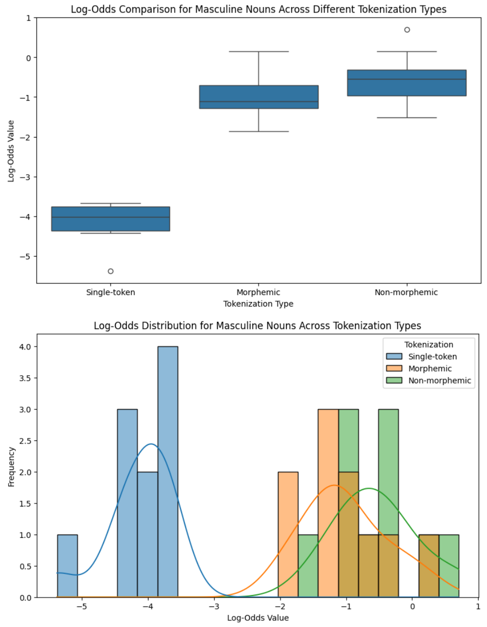

Figure 4: Boxplot and histogram for the masculine dataset Log Odds.