# Stanza Extravaganza: Improving Latin American Spanish Morphological Classifiers

Kiara Gonzalez, Daniel Helo

## INTRODUCTION

**Language Acquisition** is the study of how humans acquire language [2]. The field has long relied on corpus data (child speech and child-directed speech) for testing different hypotheses on how this process is carried out. Such experiments require robust data processing on **lengthy transcriptions.** [3] Information on pertinent parts of speech, dependencies, and other tags is difficult to parse manually and can increasingly be automated through **Natural Language Processing** (NLP) tools. Nonetheless, widely available statistical NLP algorithms, such as Stanford's **Stanza** [1], struggle with **non-English** languages due to a lack of training data.
We propose a new data **pre-processing** algorithm for Spanish dialects that augments transcription data and improves parsing accuracy, by leveraging properties of pronouns called **clitics** .

## METHODS

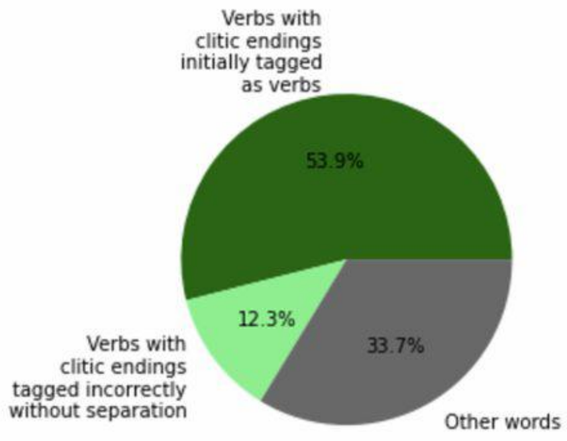
IWe utilized the Stanford **Stanza** library [1], a Python package designed for detailed natural language processing to analyze a comprehensive dataset of Spanish from Paraguay and Argentina. We identified a recurring issue with misclassifying verbs conjugated with **clitic pronouns**. To improve accuracy, we developed a preprocessing script that separates verbs from their **clitic endings** ('me', 'nos', 'te', 'se', 'la(s)', 'lo(s)', 'le(s)'), allowing for more precise identification by Stanza. This **preprocessing** ensures that if the initial part of a word is recognized as a verb, it is processed separately from its clitic, significantly enhancing the effectiveness of linguistic analysis within our dataset.

## RESULTS

To evaluate Stanza's performance on the data that was preprocessed using the Python script, we compared the tagging results of both raw and preprocessed files. We discovered that before preprocessing, 18.6% of verbs with clitic endings were not correctly identified as verbs by the morphological tagger. However, after separating the clitic endings from the verbs before tagging, these words were accurately classified as verbs by the NLP model. This demonstrates the effectiveness of preprocessing in improving Stanza's tagging accuracy.

**CHILDES** — Child Language Data Exchange System

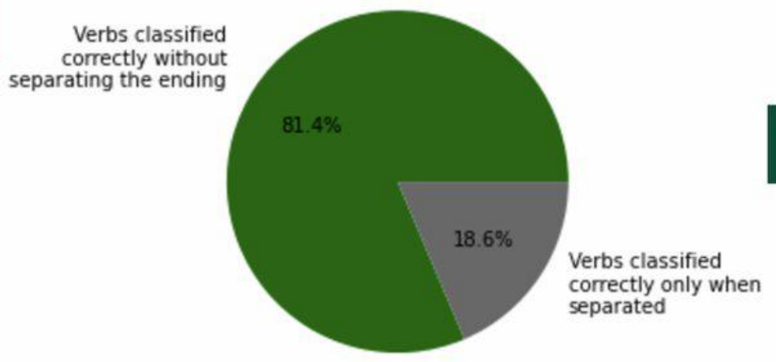
Pie chart of all words that were separated by the python script before checking if the left part was a verb




**Spanish Clitics**

| Person | Sing | Plur |
|---|---|---|
| 1st | *me* | *nos* |
| 2nd | *te* | *os*[15] |
| 3rd Masc | *lo* | *los* |
| 3rd Fem | *la* | *las* |

Verbs with clitic endings classified as verbs




## DISCUSSION

The proposed classifier has helped elucidate linguistic rules in Spanish corpora of acquisition data. Nonetheless, the clitic patterns used to develop it are subject to regional differences. For example, **Argentinian** Spanish clitics are based on **gender**, as opposed to **Paraguayan** clitics based on **animacy** [4]. Some regions draw from both systems. This reduces **quality training data** for a statistical (Stanza-like) approach. Thus, it remains to be seen how a more robust solution can be applied to any general clitic ruleset independent of regional differences.
Further research in clitic ruleset acquisition will also require more accessible and **customizable interfaces** to tools like Stanza. While our approach works for Spanish, it is **language-restricted** since it uses domain knowledge. Posterior tools like ours must allow for custom data augmentation technique implementation. This is an open area of development.

## CONCLUSIONS

The proposed approach for linguistic pre-processing of Spanish corpora data based on clitic rulesets provides a substantial **quantitative improvement** for statistical morphological classifier accuracy. Implementing language-specific pre-processing can be an effective way to **compensate** for the lack of training data in statistical NLP algorithms. Nonetheless, such implementations are dependent on various **linguistic factors**, and they can even vary from region to region.
As such, we highlight the need to develop more pre-processing techniques like the proposed one for different languages, to increase the **efficiency** of tools like Stanza. Additionally, work in this area will elucidate how general approaches to compensating lack of data can enrich the parsing of niche corpora, such as the one analyzed in Argentine and Paraguayan Spanish. Given our positive results, developing such **tooling** in accessible and customizable ways can tangibly **improve** linguistic acquisition research.

## REFERENCES

[1] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020. [pdf][bib]
https://www.researchgate.net/figure/Accusative-Object-Clitics-and-Pronouns-in-BP-and-Spanish_tbl3_314234582
[2] De Villiers, Jill G., and Peter A. De Villiers. Language acquisition. Harvard University Press, 1978.
[3] Ambridge, B. and Rowland, C.F. (2013), Experimental methods in studying child language acquisition. WIREs Cogn Sci, 4: 149-168. https://doi.org/10.1002/wcs.1215
[4] Requena, Pablo E. "Direct object clitic placement preferences in Argentine child Spanish." (2015).