Modeling Substitution Errors in Spanish Morphology Learning

Libby Barak[1] **(libby.berk@gmail.com)** and **Nathalie Fernandez Echeverri**[2] **(nfern@umd.edu)**
and **Naomi H. Feldman**[2,3] **(nhf@umd.edu)** and **Patrick Shafto**[1,4] **(patrick.shafto@gmail.com)**
[1]Department of Mathematics and Computer Science, Rutgers University, Newark, NJ, 07102 USA
[2]Department of Linguistics, University of Maryland, College Park, MD, 20740 USA
[3] Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park, MD, 20740 USA
[4] School of Mathematics, Institute for Advanced Study, Princeton, NJ, 08540 USA

Abstract

In early stages of language acquisition, children often make inflectional errors on regular verbs, e.g., Spanish-speaking children produce *–a* (present-tense 3rd person singular) when other inflections are expected. Most previous models of morphology learning have focused on later stages of learning relating to production of irregular verbs. We propose a computational model of Spanish inflection learning to examine the earlier stages of learning and present a novel data set of gold-standard inflectional annotations for Spanish verbs. Our model replicates data from Spanish-learning children, capturing the acquisition order of different inflections and correctly predicting the substitution errors they make. Analyses show that the learning trajectory can be explained as a result of the gradual acquisition of inflection-meaning associations. Ours is the first computational model to provide an explanation for this acquisition trajectory in Spanish, and represents a theoretical advance more generally in explaining substitution errors in early morphology learning.

Computational models of morphology learning mostly focus on English, leaving more complex inflectional systems and their corresponding acquisition trajectories relatively neglected (Legate and Yang, 2007; Freudenthal et al., 2007; Engelmann et al., 2019). English-speaking children often make omission errors by producing the bare form when an inflectional *–ed* or *–s* is expected. However, substitution errors are more common for languages such as Spanish and French, in which the bare form is not as frequent (Wexler, 1994; Grinstead et al., 2009; Aguado-Orea and Pine, 2015). Spanish-learning children mostly start by using the 3rd person singular (3Sg) inflections for any present tense context. Within a few months, children start using the 1st person singular inflection correctly, but continue to use the 3rd-Singular where 2nd-Singular or 3rd-Plural are expected (Fernández Martínez, 1994). Some studies have referred to the 3Sg form as the Spanish bare form (Grinstead et al., 2009), implying that substitution errors are governed by a form's default status. Other theories refer to 3Sg substitution errors as 'one-off' errors based on the semantic distance of the obligatory inflection from 3Sg (replacing either the number or the person) (Aguado-Orea and Pine, 2015).

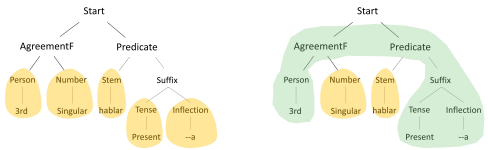

Figure 1: A tree illustration for learning and generating inflected regular verb forms in Spanish, (a) using the CFG rules, (b) using stored rules and a fragment that associates person, tense and inflection.

In this work, We propose a computational model of Spanish inflection learning that explains this trajectory of gradual acquisition (Barak et al., 2023). We simulate morphology learning using the Fragment Grammars (FG) model, a Bayesian nonparametric model that learns over a context-free grammar (CFG) (O'Donnell, 2015). FG has been shown to replicate various observation from English acquisition. Here, we design the CFG to represent the semantic and syntactic properties related to Spanish verb learning (i.e., person, number, tense, and regularity).

FG mimics human behavior in the ability to store frequently occurring fragments of the CFG to make computation more accurate and efficient. Thus, the model can generate an inflected verb using the CFG rules only or a combination of rules and fragments (See Figure 1). Following the methodology of O'Donnell (2015), we extracted all verbs in the child-directed portion of the Spanish CHILDES for ages 18-50 months using consecutive months to simulate progressive data (MacWhinney, 2000).

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 408-410.
Amherst, Massachusetts, June 15-17, 2023

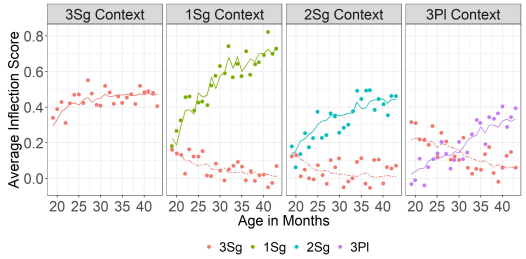

Figure 2: The average score for predicting the grammatical inflections or substituting with 3Sg inflections for the present tense showing a replication of substitution pattern and acquisition order.

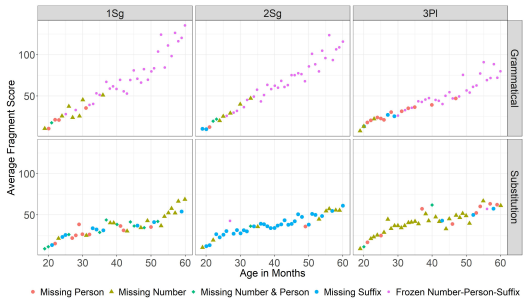

Figure 3: The average scores for stored fragments the model uses to predict 3Sg or the grammatical inflections for either 1Sg, 2Sg, or 3Pl. Each fragment includes partial set of features. Grammatical forms are enabled by the frozen number-person-suffix fragment for all three, but a different cause for mistakes is observed for each inflection context.

We annotated a new data set of Spanish verbs to identify regular and irregular verbs and their corresponding inflection for each possible context.[1]

We compared the conditional probability of the model generating the grammatical forms for the 1st-Singular (1Sg), 2nd-Singular (2Sg), and 3rd-Plural (3Pl) to their production with 3rd-Singular (3Sg) present tense inflection for each of the verbs in the data. Our model replicates the preference for 3Sg inflections for present tense both in its earlier prediction accuracy and also in varying degree of substitution pattern with other suffixes (Figure 2). Replicating psycholinguistic findings (Fernández Martínez, 1994; Grinstead et al., 2009), the model captures 1Sg as the next accurately predicted inflection, while 2Sg and 3Pl take longer. Such psycholinguistic theories referred to such errors as near-misses caused by replacing a single feature (number or person). Our analysis of the model's grammar suggests only one of the three errors stems from a single-feature miss. The model stores several fragments that associate *–a* and *–e* suffixes with 3Sg early on. Partial fragments that include partial context may lead the model to predict 3Sg for a context that requires 1Sg, 2Sg, or 3Pl (Figure 3). Errors for 1Sg mostly result from fragments that recognize *–o* as a frequent inflection without number or person attributes. 2Sg substitutions originate from fragments without the suffix, i.e., difficulty in identifying which inflections should be associated with 2Sg. Finally, mistakes for 3Pl are indeed explained in a one-off prediction stemming from fragments with missing number. Our work provides an important extension of an existing model to a highly-inflected language. Our results replicate observations from Spanish learning monolinguals while providing a novel explanation to the source for the children's erroneous productions.

[1] https://github.com/CoDaS-Lab/SVMorph

Acknowledgments

We thank Tim O'Donnell for sharing his code and Jan Edwards and Zara Harmon for helpful discussion. This research was supported by NIH R21 DC017217. For additional details please refer to Barak et al. (2023).

References

Javier Aguado-Orea and Julian M Pine. 2015. Comparing different models of the development of verb inflection in early child spanish. *PloS one*, 10(3):e0119613.

Libby Barak, Nathalie Fernandez, Naomi H Feldman, and Patrick Shafto. 2023. Modeling substitution errors in spanish morphology learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Felix Engelmann, Sonia Granlund, Joanna Kolak, Marta Szreder, Ben Ambridge, Julian Pine, Anna Theakston, and Elena Lieven. 2019. How the input shapes the acquisition of verb morphology: Elicited production and computational modelling in two highly inflected languages. *Cognitive Psychology*, 110:30–69.

Almudena Fernández Martínez. 1994. El aprendizaje de los morfemas verbales: Datos de un estudio longitudinal. *La adquisición de la lengua española*, pages 29–46.

Daniel Freudenthal, Julian M Pine, Javier Aguado-Orea, and Fernand Gobet. 2007. Modeling the developmental patterning of finiteness marking in english, dutch, german, and spanish using mosaic. *Cognitive Science*, 31(2):311–341.

John Grinstead, Juliana De la Mora, Mariana Vega-Mendoza, Blanca Flores, et al. 2009. An elicited production test of the optional infinitive stage in child spanish. In *Proceedings of the 3rd Conference on Generative Approaches to Language acquisition North America (GALANA 2008)*, pages 36–45. Erlbaum Hillsdale, NJ.

Julie Anne Legate and Charles Yang. 2007. Morphosyntactic learning and the development of tense. *Language Acquisition*, 14(3):315–344.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Timothy J O'Donnell. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Ken Wexler. 1994. 14 optional infinitives, head movement and the economy of derivations. *Verb movement*, page 305.