



DOI: 10.58734/plc-2025-0014



Mary Rosa Espinosa-Ochoa 

Center of Hispanic Linguistics, National Autonomous University of Mexico

Balanced Frequency Among Rivals in Child Spanish Morphology: A Corpus-based Study Of One Child

The study of overgeneralization provides evidence for understanding the factors that influence the application of previously acquired linguistic patterns in a novel way by children. According to the literature, frequency plays an important role in overgeneralization. For instance, overgeneralization of morphemes is less likely to occur to high frequency verbs. The aim of this study is to determine whether the elements the child retrieves to create a verbal morphological overgeneralization are of greater frequency than the target elements of their previous speech and the input. It consists of a high-density longitudinal study of a middle-class child (1;11,24-2;02,24), whose utterances were recorded daily for three months. Elements retrieved by the child to create a verbal overgeneralization were compared against the target elements of their previous speech and the input using non-parametric tests. According to the results obtained, it can be said that overgeneralized forms are associated with equal frequency patterns in the child's speech.

Key words: Spanish, child language acquisition, verb morphology, overgeneralization

Address for correspondence: Mary Rosa Espinosa-Ochoa

Center of Hispanic Linguistics, National Autonomous University of Mexico

E-mail: mrespinosao@gmail.com

This is an open access article licensed under the CC BY NC ND 4.0 License.

Explaining how children produce sentences they have not heard before has been a key problem in language acquisition studies. The study of overgeneralization in inflectional morphology provides evidence to understand how children apply linguistic patterns in a novel way. According to the literature on child acquisition, frequency effects play an important role in overgeneralization (Ambridge et al., 2015). For example, the low-frequency irregular verb *drank* is more likely to be overgeneralized (e.g., **drinked*) than the high-frequency irregular verb *blew* (e.g., **blowed*). In general terms, it can be said that if two forms compete, the most frequent will be the dominant.

These phenomena have also been tested in languages with complex morphological systems. It is known that overgeneralization errors are less common for verbs that score high in phonological density and type frequency, e.g., third person singular verb forms in Spanish (Aguado-Orea & Pine, 2015), and that type effects are smaller if the token frequency is higher, e.g., the verb form *said* is frequent and therefore stored as an unanalyzed phrase which do not contribute to generalization (Ambridge et al. 2015), because high token frequency in the input allows a form to become entrenched, and high type frequency allows for pattern detection as more frequent forms become entrenched. On the other hand, token frequency in complex morphological systems has yielded contradictory results. For instance, in Croatian, a language with a verb system containing several stems for each verb, a negative correlation was found between token frequency and class size in the input and the production of overgeneralized forms (Hrzica et al., 2023). In Finnish past tense verbs, children performed better with low frequency tokens than with high frequency ones (Kirjavainen et al., 2012).

Inflectional morphology of Spanish is very complex. There are three different paradigms of verbs, with irregular verbs in each. Irregular verbs are classified by their phonological variation in the root or stem. In addition, Spanish has pronominal clitics that share characteristics of inflectional morphemes.

The Spanish Inflectional System

Spanish is a language with a complex inflectional system. In the dialectal variety of Spanish analyzed in this study, there are 53 inflectional forms (Rojas Nieto, 2003). Spanish verbs have been classified according to their infinitive ending as belonging to one of three classes. The first classification, with the largest number of verbs, almost all of them regular, includes those that end with *-ar*, for example, *cantar* ('to sing'). The second classification, of those ending in *-er*, includes a considerable number of irregular verbs, although many are regular, for example, *beber*. The third classification, of those ending in *-ir*, has the smallest number of verbs, and most of them are irregular: for example, *ir* ('to go').

In the morphology of Spanish verbs, inflectional affixes are combined with stems, which are in turn a combination of theme vowels (TVs) and roots, as in (1):

Cantar ('to sing'), first classification:

- (1) cant á ba mos ('we were singing')
 sing tv PST.IND.IPFV 3PL

Inflectional morphemes express person, number, tense, and aspect. Roots must always contain these morphological markings. Although most verbs are regular, there are a handful of irregular verbs. According to Alcoba (1999), verbs are irregular when: (i) their root is altered, as in the verb *contar* ('to count'): *cont* + *ar* (infinitive) > *cuént* + *o* (1SG Indicative Present); or (ii) in a few cases, when the stem is altered by suppression of the TV:

- (2) verb pon-er ('to put,' infinitive) > pone ('put' IMP.2SG)
 pone(d)ré (FUT.IND.1SG, 'I will put')
 verb cab-er ('to fit', INF) > caberé (FUT.1SG 'I will fit')
 verb hac-er ('to make', infinitive) > ha(ce)ré (FUT.1SG 'I will make')

(3) Suppletive verbs *ir* ('to go') and *ser* ('to be') use distinct roots from different etymological origins in their conjugations, in complementary distribution (see Table 1).

Pronominal Clitics

Pronominal clitics are not typically considered within verbal morphology, nor have they been the focus of overgeneralization research in language acquisition studies (Clahsen et al., 2002; Radford & Ploenning-Pacheco, 1995). They occupy a fixed position in the sentence, usually defined with reference to the verb on which they depend as an argument. They are pronominal forms of unstressed objects that appear attached to the verb, either in front of (proclisis) or behind (enclisis) in a relationship of strict adjacency, where only another clitic can intervene between the two (Fernández Soriano, 1999).

Fernández Soriano (1999) argues that these clitics are affixes, not full words. They are distinguished from independent words because they depend phonologically

Table 1. *Present, Past, and Future Paradigm of Suppletive Verb ir ('to go')*

	Present indicative	Past perfect	Future
1SG	voy ('I go')	fui ('I went')	iré ('I will go')
2SG	vas ('you go')	fuiste ('you went')	irás ('you will go')
3SG	va ('s/he/it goes')	fue ('she/he/it went')	irá ('she/he/it will go')
1PL	vamos ('we go')	fuimos ('we went')	iremos ('we will go')
2PL	van ('you go')	fueron ('you went')	irán ('you will go')
3PL	van ('they go')	fueron ('they went')	irán ('they will go')

on the verb with which they appear, they cannot occur in isolation, they cannot be part of a coordination, and they cannot be elided by identity. They can be said to have characteristics analogous to those of inflectional morphemes (Fernández Soriano, 1999, p. 1255).

Overgeneralization in the Verbal System in Child Spanish

Most researchers agree that non-target-like verb forms in Child Spanish reflect the morphological asymmetry of the Spanish verb system, children overgeneralize regular verb morphology onto irregular verbs, both in root formation and inflectional affixation (Clahsen et al., 2002; Pérez Pereira & Singer, 1984; Radford & Ploenning-Pacheco, 1995). The study that documents the largest number of types of overgeneralizations made by Spanish-speaking children in spontaneous speech is perhaps that of Clahsen et al. (2002), who classify the errors into two main categories: stem/root errors and suffixation errors.

Stem/root errors. Clahsen et al. (2002) identify the incorrect use of roots in three modalities. In the first, incorrect roots are used together with a suffix that does not belong to the verb, as in **sabo* instead of *sé* ('I know'). Although the suffix *-o* expresses the regular paradigm of the first-person singular in the present indicative, the irregular form *sé* is the conventional one for this verb. In the second modality, the verb uses the regular suffix but the incorrect root, as in **cayí* for *caí* ('I fell'). In the third modality, children produce a non-existent root with a correct suffix, as in **punieron* instead of *pusieron* ('they put'). In all of these cases the anomaly corresponds mainly to the root.

Suffixation errors. In the case of suffixation errors, Clahsen et al. (2002) report the use of irregular affixes replaced by regular ones, as in the case of **queri-ba* for *quer-ía* ('I wanted'). This case illustrates the incorrect use of the conjugation for the first classification instead of the second or third. A second type they describe consists of the incorrect use of a regular conjugation suffix for the second or third classification instead of the required one, as in *pusí* instead of *puse* ('I put') and **ensucí*, instead of *ensució* ('I soiled'). In this case, *ensuciar* ('to soil') is a verb of the regular conjugation for the first classification, which requires an *-e* ending for the first-person singular of the past perfect, but the child overapplies the non-regular *-í* suffix for the conjugation of the second and third classifications.

Clahsen et al. (2002) studied whether the onset of overgeneralizations is determined by frequency of use, and they found no correlation. However, they do not investigate how frequency factors play a role in overgeneralization.

The Trace-Back Method

The trace-back method involves taking as a basis the last two transcripts of a longitudinal study of children and identifying how many of the constructions uttered by the children can be traced back to their previous speech. These last

two transcripts are called TEST. Target utterances are traced back to the MAIN corpus, that is, the remaining transcripts, to verify whether the TEST sentences can be derived directly from the previous speech, using operations such as ADD or JUXTAPOSE (e.g., *what's + that*; where *what's* is the fixed string and *that* is the addition or juxtaposition). The method seeks to verify whether children construct language using a small set of patterns that “can either be fixed strings, exact repetitions of previous material or partially schematic units with an open slot like [What's X], [or] so-called frame-and-slot patterns” (Hartmann et al., 2021, p. 478). The goal is to check whether the elements of children's speech have been incorporated in a piecemeal fashion and constructed with an element already existing in their linguistic production, or if they are generated spontaneously without any previous basis, which would show abstract knowledge. This method covers only a certain period of a child's acquisition process.

This method has been shown to be highly effective in tracing back a relatively high percentage of utterances. Koch et al. (2022) have concluded that it is effective in the way it was originally tested by Lieven et al. (2003) and Dabrowska and Lieven (2005), but that it has greater effectiveness with dense corpora, tested on the same child, and if lower thresholds are taken into account (two occurrences instead of three or more). The method has been questioned for not taking into account the frequency with which utterances are presented to the child. According to Kol et al. (2014), the trace-back method does not consider the effect of recurring strings on the entrenchment of linguistic structures (Bybee, 1995, 2006). They suggest a model that is more sensitive to frequency.

The characteristics of this method that have been proved to be effective (Hartmann et al. (2021) are considered in this study, as well as the suggestion made by Kol et al. (2014).

The Present Study

The present study is a dense corpus-based analysis, unique among Spanish corpora for its frequency of collection, in which a trace-back method is used to verify whether there is a frequency pattern among overgeneralized and target verb elements, roots and stems, inflectional morphemes, and pronominal clitics in the child's speech. The specific research goal is to determine whether the elements that the child retrieves to create an overgeneralization are of higher frequency than the target elements in her previous speech and the input. Based on the literature (Ambridge et al., 2015), a plausible hypothesis would be that the most frequent forms would be the ones the child retrieves when producing overgeneralizations. The study verifies whether this holds true for Spanish inflectional morphology, which presents a higher degree of complexity than English nominal and verbal inflectional morphology. The Trace-back method was chosen to conduct this study since it has proven to be effective in tracing the linguistic items that a child has previously produced to form new items. It was adapted so that it is useful for considering frequency, in the way it will be described in the following section.

Method

This is a longitudinal corpus-based case study of a monolingual Spanish-speaking child (1;11,25-2;03-09; MLU: 3.03), with an analysis of a dense database video recorded for one hour every day over three months, in Mexico City. The child was videorecorded by a member of her family during daily activities at home with parents, aunts, and grandparents. The family is middle-class and lives in an urban environment, the parents had ongoing university studies at the time of data collection. The transcription of the video recordings was made by native speakers of Mexican Spanish, residents of Mexico City, using a word processor. This is a dense sample, unique among the other Hispanic corpora belonging to the Corpus ETAL-UNAM (Rojas Nieto, 2007). The number of turns analyzed per participant is shown in Table 2.

Coding procedure

Each child's utterance containing a verb was first manually identified in the transcription and organized into a numbered list, from the earliest to the most recent utterance. All verbs were maintained in the chronological order in which the child uttered them. Two forms were tagged using overgeneralized verbs as parameter: target exemplars and overgeneralized exemplars. An overgeneralized exemplar was defined as a form the child used to express a function that conventionally corresponded to another form, for example, for the verb *caber* 'to fit' the child produced **cabo* 'I fit' instead of *quepo* 'I fit'. In this case, *cab-* was counted as an overgeneralized stem, and *quep-* was counted as the target stem. Target exemplars were marked, and overgeneralized exemplars were traced back manually using the list. Discourse repetitions and dubious cases were eliminated. The procedure was repeated independently by the author six months later, to check the whole process again. Cases in which a conventional form was used with another function were not considered, even when they were overgeneralizations, for example, *ábrola* ('I open it') instead of the imperative *ábrela* ('open it'), using the first person present indicative of the verb *abrir* as an imperative, or *pónela* ('he/she/it puts it'), using the third person singular indicative verb form as an imperative.

It should be emphasized that the corpus was not divided into TEST and MAIN corpora; rather, every overgeneralization was the test. In order to count the token frequency, overgeneralized and target exemplars were also numbered and tagged with a specific color to be able to locate and count them. A total of 343 verb types were identified in the child's speech. A type of verbal form means a

Table 2. *Number of Turns per Participant in the Corpus Analyzed*

Child	Mother	Father	Other Family Members
9720	5870	6846	3414

verb stem with a given affix. For example, if the child used *caer* ('to fall') in two forms, for example, *cayó* ('it fell') and *cae* ('it falls'), this would count as two verbal form types. The 23 overgeneralizations found are shown in Tables 3, 4, and 5.

Three types of overgeneralizations were identified: root/stem overgeneralizations (Tables 3), inflection overgeneralization (Table 4), and confusion of the formal imperative inflection with informal clitic pronominals (Table 5). The right-hand column of the tables shows 'rival stems,' those that belong to the paradigm of each overgeneralized verb, to indicate the possible sources of overgeneralization children can draw from. Verbs marked with a plus sign appear in more than one table because they contain more than one type of overgeneralization; these were counted twice.

Results

The frequencies of the rival forms can be seen in Tables 6, 7, 8, and 9. To determine whether there was a statistical difference among them, we applied a nonparametric test, because this type of test does not assume an a priori probability distribution for the data. We considered also that token frequencies may have differed depending on when we started tracking in the corpus. The later we began tracking, the greater the probability a form would appear, as pointed out by Hartmann et al. (2021). The Mann-Whitney *U* test was chosen because the data points were independent and few, and the study variable was categorical. In none of the cases analyzed was there sufficient evidence to conclude that the difference between the population medians was statistically significant (*p*-values > .01 in most cases and *p*-values > .05 .1 in most cases).

If we focus on the cause of overgeneralization, case by case, we can see that in the case of *sabo*, the child applied the conjugation of the first-person singular in the present indicative tense on the stem of the regular pattern of the present conjugation. What is striking about the case of *vio* is that in order to construct a first-person present tense expression, the child used the stem of the preterit. In the case of the verb *caber*, there is no regular pattern in the Spanish system, since it has three stems, *cab-*, *quep-*, and *cup-*, and we see that in fact it overgeneralizes using *cab-* and *quep-* and crosses the suffixes. Obviously, these are in all cases verbs that have competing stems, but the data do not show a tie-breaker in competition, since the populations are statistically identical.

The Mann-Whitney test was also applied to all the tokens of the morphemes whose form was equal to the overgeneralized exemplar, despite their meaning. Table 7 shows the target and rival morphemes which express the same meaning. For example, the suffix *-é* and the suffix *-e* express the second singular preterit (*llevé* 'I took', 'I /*acabé* 'I finished' vs. *dije* 'I said' /*traje* 'I bought'), but third person singular subjunctive verb forms ending in *-e*, such as *sane* ('he heals') were also counted and tested in a previous analysis. In both cases the

Table 3. *Stem Overgeneralization*

Overgeneralization	Target	Rival stems
Stems + inflection		
(+ pronoun)		
<i>*sabo</i>	<i>sé</i> ('I know')	<i>sab-</i>
	know-PRS.IND.1SG	<i>Sé</i> (PRS.IND.1SG)
		(infinitive: <i>saber</i>)
<i>*vio</i>	<i>veo</i>	<i>vi-</i>
	<i>ve-o</i> ('I see')	<i>v(e)-</i>
	see-PRS.IND.1SG	(infinitive: <i>ver</i>)
<i>*vias+</i>	<i>ves</i> ('you see')	<i>vi-</i>
	<i>v(e)-es</i>	<i>v(e)-</i>
	see-PRS.IND.2SG	(Infinitive: <i>ver</i>)
<i>*cabo</i>	<i>quepo</i>	<i>cab-</i>
	<i>quep-o</i> ('I fit')	<i>quep-</i>
	fit-PRS.IND.1SG	(infinitive: <i>caber</i>)
<i>*quepes</i>	<i>cabes</i>	<i>cab-</i>
	<i>cab-es</i> ('you fit')	<i>quep-</i>
	<i>fit-PRS.IND.2SG</i>	(infinitive: <i>caber</i>)
<i>*váyate</i>	<i>vete</i> ('go')	<i>ve</i> (Imperative)
	<i>v-e-te</i>	<i>vay-</i> (Subjunctive)
	go-IMP.SG-PRO2SG	(infinitive: <i>ir</i>)
<i>*séntete+</i>	<i>siént-a-te</i> ('sit down')	<i>sent-</i>
	<i>sit-IMP-PRO2SG</i>	<i>sient-</i>
		(infinitive: <i>sentar</i>)
<i>*se dormieron</i>	<i>se durmieron</i>	<i>dorm-</i>
	<i>se durm-ieron</i> ('they fell asleep')	<i>durm-</i>
	PRO sleep-PST.PRF3PL	(infinitive: <i>dormir</i>)
<i>*cáigate+</i>	<i>cáete</i>	<i>ca-</i>
	<i>cá-e-te</i> ('fall down')	<i>caig-</i>
	fall-IMP2SG-PRO	(infinitive: <i>caer</i>)
<i>*cáyate+</i>	<i>cáete</i>	<i>ca-</i>
	<i>cá-e-te</i> ('fall down')	<i>cay-</i>
	fall-IMP2SG-PRO	(infinitive: <i>caer</i>)

Table 4. *Inflectional Overgeneralization*

Overgeneralization	Target	Rival Inflections
*trajé	traj-e ('I brought')	-é (REG 1st conjugation 1SG Past Perfect)
	bring-PST.PRF1SG	-e (IRR 1SG.Past Perfect)
*caigues	caig-as ('you fall')	-es (1st conjugation 2SG Present Subjunctive)
	fall-PRS.SBJV2SG	-as (2nd conjugation 2SG Present Subjunctive)
*vengue	veng-a ('she/he comes')	-e (1st conjugation 1 & 3SG Present Subjunctive)
	come-PRS.SBJV3SG	-a (3rd & 2nd conjugation 1 & 3SG Present Subjunctive)
*vias+	v-es ('you see')	-as (1st conjugation 2SG Present)
	See-2SG.PRESENT	-es (2nd conjugation 2SG Present)
*echastes	ech-aste ('you threw')	-s (2SG conjugations in Spanish end in -s)
	throw-PST.PRF2SG	-ste (2SG Past perfect)
*manchastes	manch-aste ('you stained')	-s (2SG conjugations in Spanish end in -s)
	stain-PST.PRF2SG	-ste (2SG Past perfect)
*me pegués	me peg-ué ('I hit')	-s (2SG conjugations in Spanish end in -s)
	PRO hit-PST.PRF1SG	-é (1SG Past perfect)

Table 5. *Confusing Formal Imperative Inflection with Informal Clitic Pronominals and Vice Versa*

Overgeneralization	Target Informal	Target Formal
*cálm-e-te	cálmate	cálmese
	cálm-a-te ('calm down')	cálm-e-se
	relax-IMP2SG-PRO	relax-IMP2SG-PRO
*súb-a-te	súbete	súbese
	súb-e-te ('climb')	súb-a-se
	climb-IMP2SG-PRO	climb-IMP2SG-PRO
*mét-a-te	métete	métase
	mét-e-te ('get in')	mét-a-se
	get.in-IMP2SG-PRO	get.in-IMP2SG-PRO
*mét-e-se	métete	métase
	mét-e-te ('get in')	mét-a-se
	get.in-IMP2SG-PRO	get.in-IMP2SG-PRO
*s(i)ént-ete+	siéntate	siéntese
	siént-a-te ('sit down')	siént-e-se
	sit-IMP2SG-PRO	sit-IMP2SG-PRO
*cáy-a-te+	cáete	cáigase
	cá-e-te ('fall down')	cáig-a-se
	fall-IMP2SG-PRO	fall-IMP2SG-PRO

same results were obtained with the Mann-Whitney test. In the case of the verbs *manchaste* ('you stained') and *echaste* ('you threw'), the frequency of the tokens corresponding only to the verbs of the first conjugation are shown in column 5, but the Mann-Whitney test was also performed taking into account the verbs in the second-person singular of the second and third conjugations, and again no statistical difference was found.

The overgeneralizations of imperative plus clitic pronoun are a particular case, since the child mixed the morphemes of the formal imperative conjugation with those of the informal pronoun and vice versa. One possible explanation is that the child perceived the inflection and the pronoun as a 'chunk.' We thus took into account this target combination of morpheme forms plus formal pronominal clitic and informal morpheme plus informal pronominal clitic (Table 8) and applied the Mann-Whitney test between the overgeneralized exemplary and target 1, and then the overgeneralized exemplary and target 2. Assuming that any of the morphemes (formal and informal) were more frequent, we also performed the analysis shown in Table 9, obtaining the same statistical result. Note that the overgeneralized verb form *cáigate* was not considered in Table 11 because the target form is *cáigase*, so there was no problem with the combination of inflection and pronominal clitic. In this case the only element that was overgeneralized was the pronominal clitic.

Dąbrowska and Lieven (2005) established a frequency threshold of two occurrences in the main corpus for identifying component units, "assuming that children have stored all units that occur with a frequency of 2 or more in the corpus". Using this criterion, the verbal forms that did not appear at least twice were discarded from the analysis. The Mann-Whitney test was run again, this time taking in account all data in Tables 6 to 9 as a list (group overgeneralized vs. group target), and the same results were obtained ($p = .394148$; $p > .01, .05$, and $.1$). A Wilcoxon Signed Rank Test, another non-parametric test, was also performed with this list (group overgeneralized vs. group target obtaining the same results again (Wilcoxon Signed Rank $Z = p = .342$, $p > .01, .05$ and $.1$).

Changing the Paradigm of Analysis

The trace-back analysis paradigm was changed to verify whether there was a statistical difference between the overgeneralized forms and the target forms, this time using the input of the whole corpus, not just the exemplars prior to overgeneralization. Both input and output were considered to be closely associated, as their distribution frequencies are "extremely similar" (Ambridge et al., 2015, p. 242). By measuring the input in the whole corpus, we eliminated the factor limiting the frequency of the earliest forms, since for these we had fewer exemplars. In this way the number of exemplars that could be analyzed was increased, since in adult speech exemplars are always more numerous than in child speech. Finally, a change in analysis from infant to adult speech might show a difference, because the most important frequency

Table 6. *Traced-Back Frequency of Verbal Roots or Stems*Mann-Whitney $U = 45.5$, $p = .757$, $p > .1$

Overgeneralization	Overgeneralized Exemplar		Target	
	Stem	Traced-Back Frequency	Stem	Frequency
*sabo	sab-	3	sé	6
*vío	ví	1	ve-	4
*vías	ví	4	ve-	43
*cabo	cab-	0	quep-	0
*quepes	quep-	0	cab-	4
*váyate+	vay-	4	ve-	3
*séntete+	sent-	2	sient-	2
*se dormieron	dorm-	12	durm-	2
*cáigate	caig-	2	ca-	2
*cáyate+	cay-	7	ca-	2

Table 7. *Traced-back Frequency of the Morphemes*Mann-Whitney $U = 14.5$, $p = 0.224$, $p > .1$

Overgeneralization	Overgeneralized Exemplar		Target	
	Morpheme	Traced-back frequency	Morpheme	Traced-back frequency
*trajé	-é	35	-e	12
*caigues	-es	15	-as	13
*vengue	-e	1	-a	2
*vías	-as	38	-es	87
*echastes	-s	145	-ste	8
*manchastes	-s	146	-ste	10
*me pegués	-es	8	-é	3

Table 8. *Traced-back Imperative Morpheme + Clitic Pronoun*Column 2 against column 3: Mann-Whitney $U = 23.5$, $p = 0.948$, $p > .1$ Column 2 against column 4: Mann-Whitney $U = 11$, $p = 0.088$, $p > .05$

Overgeneralization	Overgeneralized Exemplary	Target 1	Traced-Back Frequency	Target 2	Traced-Back Frequency
*mét-e-se	-ese	0	-ase	2	38
*cálm-e-te	-ete	3	-ate	17	0
*s(i)ént-e-te	-ete	3	-ate	20	0
*mét-a-te	-ate	4	-ete	9	2
*súb-a-te	-ate	34	-ete	31	2
*cáyate	-ate	9	-ete	1	0
*cáigate	-ate	10	-ete	1	0

Table 9. *Traced-Back Morphemes*Mann-Whitney $U = 10.5, p = .262, p > .1$

Overgeneralization	Imperative (-e) Traced-back frequency	Imperative (-a) Trace-back frequency
*cálmete	8	90
*súbate	211	47
*métate	26	4
*métese	56	248
*s(i)é(nt)ete	13	129
*cáyate	4	62

Table 10. *Total Tokens Input Frequency for Stems*Mann-Whitney $U = 27, p = 0.636, p > .1$

Child Overgeneraliza- tion	Stem	Input Frequency Overgeneralized Exemplar	Stem	Input Frequency Target Exemplar
*sabo	sab-	143	sé	102
*vío/vías	ví-	227	ve-	1275
*cabo/quepes	cab-	32	quep-	8
*váyate	vay-	74	ve-	50
*séntete	sent-	36	sient-	38
*dormieron	dorm-	107	durm-	14
*cáyate	cay-	58	cae-	71
*cáigate	caig-	28	cae-	71

Table 11. *Total Tokens Input Frequency for Inflectional Morphemes*The Mann-Whitney $U = 19.5, p = .563, p > .1$

Child Overge- neralization	Overgeneralized Morpheme	Input Frequency	Target Morpheme	Input Frequency
trajé	-é	109	-e	57
(no te) caigues	-es	244	-as	203
vengue	-e	166	-a	304
vías	-as	961	-es	1273
echastes	-s	2554	-ste	510
manchastes	-s	2554	-ste	510
(me) pegués	-es	244	-é	109

effects in early acquisition have been measured from the frequency of the input (Ambridge et al., 2015). All verbs from the overall input were extracted using WordStatt and identified manually using the same criteria described in Section 4.1. The Mann-Whitney Test was also used to verify the statistical difference in the input frequency, since data points were again independent and few. In most of the cases analyzed, there was sufficient statistical evidence to reject the alternative hypothesis ($p > .01$, $.05$, and $.1$), so we concluded that the populations were identical in all cases.

Neither in the child's previous inventory nor in the general input data was there a dominant frequency pattern leading to the choice of the overgeneralized exemplar. Although the child's data were numerically a much smaller sample, the frequencies did not follow the same trend. For example, if we compare the data in Table 6 and Table 10, *sab-* was found less frequently than *sé* in the child's speech. However, the opposite is true of *dorm-* and *durm-*.

If we compare the overgeneralized combinations with respect to the combinations of the formal variant of the child's previous discourse (Table 9), we see that, as in Table 13, it cannot be proved that $p < .1$. This is the only case in which p -values were not greater than $.05$). Thus, we cannot rule out that chunks (suffix + clitic pronoun) were of higher frequency and may have influenced overgeneralization.

The frequencies of the Tables 10 to 13 were used as a list to perform the Mann-Whitney Test and the Wilcoxon Signed Test. No statistically significant differences were found again (Mann-Whitney $U = 10.5$, $p = .262$, $p > .1$; Wilcoxon Signed $Z = -1.72$, $p = .085$, $p > .05$).

Discussion

The results of this study showed that the elements this child retrieved to create an overgeneralization were not necessarily of higher frequency than the target elements, either in her previous speech or in the input. Our analysis showed that there is a period in the child's language development in which the elements compete, and during this period there is no statistically significant information to help her identify the target forms. That is, the child was at a stage in which she was organizing her system, but the associations between the target form and the competitors were unstable, so she could not operate on these associations, and they became entrenched. According to Hržica et al. (2024, p. 252): "All linguistic knowledge is available in the format of associations, and entrenchment processes operate over these associations so that they become entrenched. The greater the number of items displaying similar patterns, the more entrenched the schema". These results only showed another perspective, since we can say that when there were similar frequencies of items displaying similar patterns showing an association, the function or meaning expressed is subject to overgeneralization.

Table 12. *Total Tokens Input Frequency for Imperative Morphemes (IM) + Pronominal Clitic (PC)*Column 3 against column 5: Mann-Whitney $U = 4.5, p = 0.031; p < 0.01$ Column 3 against column 7: Mann-Whitney $U = 17.5, p = 1.00; p > .1$

Child Overgeneralization	Overgeneralized IM + PC	Input Frequency	Formal IM + PC	Input Frequency	Informal IM + Formal PC	Input Frequency
Cálmete	-ete	129	-e + se	52	-a + te	459
Súbate	-ate	459	-a + se	72	-e + te	129
Métate	-ate	459	-a + se	72	-e + te	129
Métese	-ese	52	-a + se	72	-e + te	129
Siéntete	-ete	129	-e + se	52	-a + te	459
Cáyate	-ete	129	-e + se	52	-a + se	72

Table 13. *The Same Verb in the Imperative with an Unstressed Pronoun*Mann-Whitney $U = 8.5, p = .460, p < .1$

Child Overgeneralization	Overgeneralized Suffix Based on the Pronoun Form: STEM-suffix-X		Counterpart	
Cálmete	-e-	0	-a-	5
Súbate	-e-	29	-a-	0
Métate	-e-	25	-a-	2
Métese	-a-	2	-e-	25
Siéntete	-a-	40	-e-	16

The literature on child English acquisition has shown that if two forms compete, the more frequent form will be the dominant one (Ambridge et al., 2015). According to these results, if two forms compete that do not differ significantly in terms of token frequency, she might retrieve either one. Both competitors appeared in the child speech, and the overgeneralized form was not necessarily the most frequent one, yet the frequency was not significantly higher or lower than that of the competitors. In an overgeneralization such as **sabo*, for example, it is clear that the child retrieved the present indicative first-person singular suffix to perform the overgeneralization. The target form *sé* had not reached a level of frequency sufficient to become entrenched and block the target form, even though it is already present in the discourse. The overgeneralization of **cabo* and **quepes* is a prototypical example: we cannot ask whether the child used the more frequent stem to perform an overgeneralization, since she overgeneralized with both stems. This suggests that the frequency data are not consistent, that they do not show sufficient frequency to entrench either of the forms. Both stems were available in the child's memory, just as they were for the suffixes of first-person -o and second person singular -es, but the child needed more experience with the target forms *quepo* and *cabes* so that they become entrenched.

This observation invites similar questions about other languages. The negative correlation between token frequency and class size in the input and the production of overgeneralized forms found for Croatian (Hrzica et al., 2023) may result from a lack of sufficient difference between competing forms. In Finnish, if children perform better with low than with high frequency tokens, it may be the result of statistically significant differences between the competitors among the universe of low frequency tokens.

This study suggests that, in some cases, the child overgeneralizes competing morphemes whose frequency is not significantly different. If we compare this with a case such as the English plural noun **mouses* versus *mice*, it could be said that children block the form **mouses*, while the present study shows that they do not block the form **sabe* even though *sé* exists. The inflectional suffix of the first person is set as the base and the child can retrieve the stem *sab-* to obtain **sabo*, but the target form *sé* is also available. *Sabo* is not blocked, since *sé* has not reached sufficient frequency to block an overgeneralization, so the forms can alternate.

Although these results showed that, generally speaking, the child overgeneralized morphemes whose rivals were not statistically significantly different from each other, it is clear that there is a story to tell behind each particular verb form. For example, Baker (2022) has studied the acquisition of the stigmatized variant of the second person singular preterit indicative, for example *manchastes*, ('you stained'), as opposed to the standard variant *manchaste*. This form is used in less prestigious variants which adults regularize following the dominant pattern of Spanish second person preterit conjugation.

Baker finds that this stigmatized use is prevalent in older children (4;00-12;00) when this variant is used in their environment, but the frequency of use is the same in younger children whether it is used in their environment or not. In general, children tend to overgeneralize *-s* in low-frequency verbs, such as *pintastes* ('you painted') and *abristes* ('you opened'). It is interesting to note that Baker (2022, p. 11) did not find overgeneralizations in lexically bound collocations like *tuviste miedo* ('you were afraid'), or *dijiste que* ('you said that'). This same verbal form *dijiste* ('you said') was overregularized when it did not appear as a lexically bound collocation, adding evidence to confirm that highly frequent strings are learned as chunks.

Limitations and Future Research

It is important to consider that the results obtained here are limited to a study made on one child, it would be necessary to repeat the analysis with dense corpora of a larger number of children to validate the results.

The analysis in the present study does not allow us to determine how children overcome this period of overgeneralization. We can only suggest that experience with the stem + suffixation (PRO) target combination in the language will lead them to do so. It has been suggested that children know how elements can be combined syntactically because meaning guides them as to the use of transitive and intransitive verbs. This argument is not applicable to type of constructions examined here, insofar as the meaning of the overgeneralized forms is identical. It has also been argued that children learn by statistical preemption, as they are exposed to the conventional forms. A larger dense longitudinal study would be necessary to determine whether forms accumulate and whether exposure to conventional forms blocks overgeneralizations in specific constructions.

Another possible option is through interaction, where adults provide the exemplar through reformulations that allow the child to identify the appropriate form. The present study demonstrates that the child had access to the target forms in the input, and also had the target exemplar available in her repertoire, because she was already producing it. An interaction that would block overgeneralization would have to be significant, such as one that caused adults to laugh, leaving an impact on infant memory.

The case of clitic overregularization is also interesting; for the moment, we can only suggest that these are target and overgeneralized forms that appear with balanced frequency, which does not allow the child's mind to anchor the target combination. A study focused on this phenomenon remains to be carried out in the future.

Conflict of Interest Disclosure

The Author reports no conflicts of interest.

Funding

This study was funded by PAPIIT (IN401319).

Research Ethics Statement

This is a corpus-based analysis of one child. The parents of the child themselves video recorded the corpus and authorized its use for academic purposes. One of them transcribed the corpus. All information that could reveal the identity of participants was anonymized to protect the identity of participants.

Data Availability Statement

Data associated with the current study will be made available upon reasonable request to the corresponding author.

References

- Aguado-Orea, J., & Pine, J. M. (2015). Comparing different models of the development of verb inflection in early child Spanish. *PLoS ONE*, 10(3), 1–21. <https://doi.org/https://doi.org/10.1371/journal.pone.0119613>
- Alcoba, S. (1999). La flexión verbal [The verbal inflexion]. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (Vol. 3, pp. 4915–4992). Espasa.
- Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(02), 239–273. <https://doi.org/10.1017/S030500091400049X>
- Baker, E. (2022). ¡Casi te caíste!: Variation in second person singular preterit forms in Spanish Children. *Journal of Child Language*, 49(6), 1256–1267. <https://doi.org/10.1017/S0305000921000507>
- Clahsen, H., Aveledo, F., & Roca, I. (2002). The development of regular and irregular verb inflection in Spanish child language. *Journal of Child Language*, 29(3), 591–622. <https://doi.org/10.1017/S03050009020005172>
- Dabrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of childrens' question constructions. *Cognitive Linguistics*, 16(3), 437–474. <https://doi.org/10.1515/cogl.2005.16.3.437>
- Fernández Soriano, O. (1999). El pronombre personal. Formas y distribuciones. Pronombres átonos y tónicos [The personal pronouns. Forms and distributions]. In I. Bosque & V. Demonte (Eds.), *Gramática descriptiva de la lengua española* (pp. 1209–1274). Espasa.
- Hartmann, S., Koch, N., & Quick, A. E. (2021). The traceback method in child language acquisition research: Identifying patterns in early speech. *Language and Cognition*, 13(2), 227–253. <https://doi.org/10.1017/lang-cog.2021.1>
- Hržica, G., Botica, T. B., & Kosutar, S. (2023). Stem overgeneralizations in the acquisition of Croatian verbal morphology: Evidence from parental questionnaires. *Word Structure*, 16(2–3). <https://doi.org/10.3366/word.2023.0228>
- Hržica, G., Košutar, S., Botica, T. B., & Milin, P. (2024). The role of entrenchment and schematisation in the acquisition of rich verbal morphology. *Cognitive Linguistics*. <https://doi.org/10.1515/cog-2023-0022>
- Kirjavainen, M., Nikolaev, A., & Kidd, E. (2012). The effect of frequency and phonological neighbourhood density on the acquisition of past tense verbs by Finnish children. *Cognitive Linguistics*, 23(2), 273–315. <https://doi.org/10.1515/cog-2012-0009>
- Koch, N., Hartmann, S., & Quick, A. E. (2022). The traceback method and the

- early construction: theoretical and methodological considerations. *Corpus Linguistics and Linguistic Theory*, 18(3), 477–504. <https://doi.org/10.1515/cilt-2020-0045>
- Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, 30(2), 333–370. <https://doi.org/10.1017/S0305000903005592>
- Maslen, R., Theakston, A. L., Lieven, V. M., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language and Hearing Research*, 47, 1399–1333. [https://doi.org/10.1044/1092-4388\(2004/099\)](https://doi.org/10.1044/1092-4388(2004/099))
- Pérez Pereira, M., & Singer, D. (1984). Adquisición de morfemas del español [Acquisition of morphemes in Spanish]. *Infancia y Aprendizaje*, 27/28, 205–221. <https://doi.org/10.1080/02103702.1984.10822051>
- Radford, A., & Ploenning-Pacheco, I. (1995). The morphosyntax of subjects and verbs in child Spanish: A case study. *Essex Reports in Linguistics*, 5, 23–67.
- Rojas Nieto, C. (2003). Early acquisition of Spanish verb inflexion. A Usage-based account. *Psychology of Language and Communication*, 7(2), 17–36.
- Rojas Nieto, C. (2007). La base de datos ETAL. Origen, descripción y metas de un proyecto [ETAL Database. Origin, description and goals of a project]. *Jornadas Filológicas 2005*, 571–601.