Acquisition of Irregular Patterns in Spanish Verbal Morphology

LAIA MAYOL
University of Pennsylvania*
laia@ling.upenn.edu

Abstract. This paper examines acquisition data of two irregular patterns in Spanish verbal morphology: diphthongization and velar insertion. The data contains 345 instances of errors, which are not equally distributed among all conjugation classes or error types. A statistical analysis revealed that input from adults and whether a verb belongs to the first conjugation are the variables that best predict the verbs’ correct usage. This data poses challenges to theories of morphology that claim that all irregular verbs are individually stored in the lexicon, while it is compatible with theories that propose that rules apply to subclasses of irregular verbs.

1 Introduction

This paper examines the acquisition of some irregular patterns in Spanish verbal morphology, namely those involving diphthongization and velar insertion. I present data on the acquisition of these patterns, perform a statistical analysis and relate it to different models proposed to account for child acquisition of morphology, namely, the Words and Rules model (Pinker and Ullman 2002) and the Rules and Competition model (Yang 2002).

English past tense has been the object of a long controversy in the literature between connectionism and generative linguistics (see McClelland and Patterson (2002) and Pinker and Ullman (2002)). However, English past tense has a relatively simple morphology and phonology. Thus, the debate could benefit from looking at a morphologically more complex language, such as Spanish, in which there are clear patterns within the irregular verbs.

The paper is structured as follows: Section 2 describes the irregularities this paper is concerned with; Section 3 gives an overview of previous work on morphology acquisition; Section 4 presents the child data and the main results and Section 5 the analysis. Section 6 concludes.

*Many thanks to Charles Yang for guiding me in this project and making the automatic data extraction possible. Thanks also to the three anonymous reviewers for their very detailed and useful comments and to Josh Tauberer for proof-reading this paper.
2 Irregular patterns in Spanish verbal morphology

Spanish verbs are highly inflected: their morphology combines inflectional pieces (which contain information about tense, mood, number and person agreement) with the stem; the stem can be further subdivided into a root and a theme vowel, as shown in (1) for the second person plural of the past tense of the verb *hablar* (‘to talk’):

(1) hablo ba ais
    speak theme.vowel past 2p plural

There are three theme vowels in Spanish, which divide the verbs into three classes or conjugations: the first conjugation, marked with the vowel [a], is the open class conjugation and the one that contains most verb types; the second conjugation, marked with [e], and the third conjugation, marked with [i], have fewer members, an important number of which are irregular.

2.1 Diphthongization

Spanish verbs present a well-known morphophonemic alternation\(^1\). In certain verbs, mid vowels are diphthongized in stressed syllables. (2) shows the pattern for the present indicative of the verbs *comenzar* (‘begin’) and *contar* (‘count’).

(2) comienzo comienzas comienza comenzamos comenzáis comienzan
cuento cuentas cuenta contamos contáis cuéntan

This alternation is lexically arbitrary; that is, not all verbs containing mid vowels present a diphthong in stressed syllables. In fact, we find minimal pairs of verbs, such as *contar-montar*, in which the first verb presents the diphthong (cuento), while the second does not (monto).

In verbs from the first and second conjugation, this alternation is unpredictable. However, while in the first conjugation, diphthongization is a minority pattern and non-alternation is the default pattern for new verbs (Albright et al. 2000), in the second conjugation, diphthongization is more common. In contrast, all third conjugation verbs containing a mid vowel present an alternation. For such verbs, there are two different patterns: (i) diphthongization of the stressed syllables of the present indicative and (ii) raising of the stressed mid vowel (always to [i]). Both patterns are illustrated in (3), with the verbs *mentir* (‘lie’) and *pedir* (‘ask’), respectively.

(3) miento miéntes miénte mentimos mentís miénten
    pido pídes pide pedimos pedís piden

---

\(^1\)See Harris (1985) for an analysis in which a phonological rule with a morphological conditioning derives the diphthongized forms.

\(^2\)The graphic accent in (2), (3) and (4) indicates phonological stress.
The diphthong alternation is not restricted to verbal morphology. De-verbal nouns sometimes show the diphthong in the stressed syllable: *juego* (‘game’ from *jugar*). This pattern is also found in pairs of related words, such as *bueno* (‘good’) - *bondad* (‘goodness’) or *huevo* (‘egg’) - *oviparo* (‘oviparous’).

### 2.2 Velar insertion

A velar stop is inserted in the first person singular of the indicative and also in the subjunctive of some verbs from the second and third conjugation. Although this alternation affects a small number of verbs, some of them are very common, such as *tener* (‘to have’), *poner* (‘put’) or *salir* (‘go out’), whose first person singular forms are *tengo*, *pongo* and *salgo*, respectively.

A few common verbs show both alternations (diphthongization and velar insertion) at the same time. Both *tener* (‘to have’) and *venir* (‘to come’) show velar insertion in the present subjunctive and 1st person singular of the present indicative and [ie] diphthongization in 2nd singular and 3rd person of the present indicative, as shown in (4) for the verb *tener*. Also, *poner* shows velar insertion in the present subjunctive and first person singular of the present indicative and [ue] diphthongization in the participle (i.e. *puesto*).

(4) téengo tiénies tiénie tenémos tenéis tiénen

### 3 Background

#### 3.1 Morphology acquisition

Several models have been proposed to explain how children acquire regular and irregular inflected forms. Here I am comparing two models which make explicit use of rules:

1. The Words and Rules model (Pinker and Ullman 2002) proposes that all regular forms are derived by a rule, while irregular forms are stored in the lexicon. A stored form blocks the application of the rule (*brought blocks bringed*). According to this approach, children produce overregularizations, such as *goed*, when the irregular form has not been stored in the lexicon and, therefore, nothing prevents the rule from applying. In contrast, overirregularizations are predicted to be impossible and indeed, in English, they are extremely rare (Xu and Pinker 1995). This approach denies that there are subclasses within the irregular verbs, since every irregular verb is individually stored in memory.

---

3 Other models, such as connectionism, propose that there are no rules, but a network which maps base forms to past-tense forms (Rumelhart and McClelland 1986). Since I am mostly interested in models that make use of rules, I will not be discussing this model further in this paper.
2. The Rules and Competition model (Yang 2002) assumes that there are rules both for the regular and for the irregular forms and that these rules compete against each other. Every irregular rule \( R \) which applies to a certain verb class is associated with a probability \( P_R \) and each assignment of a verb \( x \) to an irregular class of verbs \( S \) carries another probability \( P(x \in S) \). The acquisition task consists of updating both probabilities and learning is successful when \( \forall x P(x \in S) = P_R = 1 \), that is when the learner can reliably associate the irregular verb with the irregular rule and the irregular rule can reliably be applied over the default rule. Overregularization is predicted to occur when either \( P(x \in S) < 1 \) (the child cannot reliably associate a verb to its irregular class \( S \)) or when \( P_R < 1 \) (the default rule and not the irregular rule \( R \) wins the competition). This model also predicts that (a) for two verbs in the same class, the most frequent verb will be used more correctly and (b) for two verbs equally frequent from different classes, the one in the most frequent class will show a more correct use.

In both models, children need to learn which rule(s) are productive and which are not and need to memorize. However, what exactly is memorized is different in each model, as explained above.

3.2 Morphology acquisition in Spanish

Clahsen et al. (2002) analyzed verb inflection produced by 15 Spanish-speaking children taken from longitudinal and cross-sectional samples of spontaneous speech and narratives. They found that regular suffixes were sometimes applied to irregular verbs; in contrast, there were no cases of irregular suffixes applied to regular verbs. They report 168 irregular verb tokens in which there were errors, against 3446 correct irregular forms, yielding an overall error rate of 4.6%. The two most common cases of overregularizations were stem overregularizations (116) (sabo instead of the correct form sé) and conjugation-internal regularizations (124) (pusí instead of puse). In contrast, they found only 2 cases in which errors occurred in regular verb forms (against 2071 correct verbs). Thus, they found a strong difference between regular and irregular verbs in Spanish children’s errors.

Clahsen et al. (2002) did not include the errors regarding diphthongs in these counts. They report 107 tokens in which a non-diphthongized form was produced in a context that required a diphthongized one. There were no cases in which the children produced a diphthongized form in a context that required a non-diphthongized one. They analyzed overregularization rates by grouping verbs according to their sample frequencies. Overregularization rates were higher for verbs with low sample frequencies. They explain these findings through the Words and Rules model, since it postulates the difference between regular and irregular verbs that they found in the data.
3.3 Psycholinguistic and computational studies

Bybee and Pardo (1981) carried out a nonce-probe experiment with Spanish verbs and found an interesting asymmetry between first and third conjugation verbs. When presented with ‘wug’ first conjugation verbs in third person singular, like bierca and duenta, subjects mostly produced biercó (73%) and duentó (86%), instead of bercó or dontó. That is, for first conjugation verbs, subjects assumed that the diphthong was part of the root. In contrast, when subjects were presented with nonce infinitives in third conjugation, subjects mostly introduced an alternation, either diphthongization or raising, depending on the phonological shape of the verb.

Albright et al. (2000) tried to determine whether there are superficial cues that distinguish diphthongizing from non-diphthongizing contexts. Their algorithm derived rules and probabilities for first conjugation verbs. Overall, diphthongization tended to be disfavored in the data, appearing with probability 0.09. Higher probabilities were found for more specific phonological contexts. These probabilities were then tested against Spanish speaker’s intuitions and they found a significant correlation.

4 Data, methods and results

The present study is based on an analysis of the transcriptions of the speech of six monolingual Spanish-speaking children, drawn from the CHILDES database (MacWhinney and Snow 1985): María (1;7-3;10), Magín (1;7-2;7), Irene (0;11-3;2), Juan (2-4), Koki (1;7- 2;11) and Eduard (1;4-3;10).

One of the goals of this paper is to analyze in which circumstances children make morphological mistakes and whether they are correlated with other factors. To that end, all child utterances containing the relevant alternations or failing to present the correct alternation were automatically extracted and manually corrected. That is, we extracted both correct forms, such as tengo, and cases of overregularization, in which the child failed to produce the alternation and followed the regular pattern (i.e. teno). The following types of verbs were extracted: (1) Verbs with [ie] alternation (IE): sentar vs. siento, (2) Verbs with [i] alternation (I): pedir vs. pido, (3) Verbs with [ue] alternation (UE): lllover vs. llueve, (4) Verbs with velar insertion (VI): salir vs. salgo and (5) Verbs with velar insertion and ie diphthongization (VI + IE): tener vs. tengo vs. tiene. Additionally, an exhaustive manual search was performed to find cases of overirregularization: i.e. cases in which a child introduced one of the alternations in a regular verb (for example, cuemo instead of como, first person singular of comer, ‘eat’). No such cases were found, in accordance with what Clahsen et al. (2002) reported.

Tables 1.1, 1.2 and 1.3 summarize the data and its main results. Table 1.1 summarizes the errors that each child produced: the number of incorrect tokens (Inc), the number of correct tokens (Corr) and the infinitive of the
verbs in which they produced some errors (the first number in parenthesis indicates the number of tokens of that verb with an error and the second the total count of tokens of that verb that the child produced\textsuperscript{4}). The incorrect verbs have been sorted by the type of irregularity they failed to show\textsuperscript{5}.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|l|}
\hline
Child & Inc & Corr & Incorrect verbs \\
\hline
Edu. & 7 & 15 & UE: colgar (1/1), poder (4/4) \\
& & & VI: tener (2/3) \\
Mag. & 98 & 618 & UE: jugar (3/7), poder (3/41) \\
& & & IE: querer (45/365), tener (16/90) \\
& & & VI: tener (10/34), poner (21/43) \\
Kok. & 134 & 132 & UE: colgar (9/9), llover (2/3), poder (14/26), dormir (2/4) \\
& & & IE: cerrar (3/5), sentar (15/15), querer (29/86), tener (20/37), venir (13/16) \\
& & & VI: parecer (1/1), poner (15/25), tener (11/31) \\
Ire. & 44 & 355 & UE: contar (2/37), jugar (1/1), sonar (2/2), volar (6/6), llover (2/3), poder (1/6) \\
& & & IE: querer (17/77), tener (10/123), venir (1/19), merendar (1/1) \\
& & & VI: tener (1/58) \\
Jua. & 17 & 73 & UE: morder (6/8), dormir (3/4) \\
& & & IE: querer (3/24), tener (2/10), venir (1/1) \\
& & & VI: tener (2/6) \\
Mar. & 45 & 284 & UE: acordar (1/1), colar (1/1), jugar (5/6), llover (3/5), poder (2/14), poner (1/4), volver (7/7), sonar (2/7), volar (3/4) \\
& & & IE: cerrar (3/8), sentar (1/15), hacer (1/25), querer (3/24), tener (1/21) \\
& & & VI: tener (11/110) \\
\hline
\end{tabular}
\caption{Error data for every child in the sample}
\end{table}

Tables 1.2 and 1.3 summarize the data: the former sorts it by conjugation and the latter by type of irregularity. Each table includes the count of incorrect, correct and total tokens used by children, the correct usage rate (CUR) and the total count of tokens used by adults. The adult counts estimate the input that children received: for each verb with the relevant alternation that the children uttered in the transcriptions, we extracted, from the same transcriptions, the tokens of those verbs that adults uttered. The CUR measures the children’s knowledge of irregular verbs and it is calculated by dividing the number of correct tokens by the total number of tokens, thus it is equivalent to the percentage of correct verbs.

\textsuperscript{4}For some of the verbs, children did not make any errors. That’s why the sum of the second numbers in parenthesis is smaller than the number in Corr.

\textsuperscript{5}Tokens of verbs with VI+IE irregularity have been sorted depending on whether they failed to show VI (\textit{tengo} instead of \textit{tengo}) or IE (\textit{tene} instead of \textit{tiene}) This is why the verb \textit{tener}, which shows the two irregularities, appears twice for Mária.
Laia Mayol

<table>
<thead>
<tr>
<th>Conjugation</th>
<th>Inc</th>
<th>Correct</th>
<th>Total</th>
<th>CUR</th>
<th>Adult Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>60</td>
<td>154</td>
<td>214</td>
<td>0.72</td>
<td>740</td>
</tr>
<tr>
<td>2nd</td>
<td>265</td>
<td>1191</td>
<td>1456</td>
<td>0.82</td>
<td>4711</td>
</tr>
<tr>
<td>3rd</td>
<td>20</td>
<td>132</td>
<td>152</td>
<td>0.87</td>
<td>213</td>
</tr>
</tbody>
</table>

Table 1.2: Data sorted by conjugation

<table>
<thead>
<tr>
<th>Conjugation</th>
<th>Inc</th>
<th>Correct</th>
<th>Total</th>
<th>CUR</th>
<th>Adult Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0</td>
<td>13</td>
<td>13</td>
<td>1</td>
<td>39</td>
</tr>
<tr>
<td>IE</td>
<td>186</td>
<td>904</td>
<td>1090</td>
<td>0.83</td>
<td>2373</td>
</tr>
<tr>
<td>UE</td>
<td>85</td>
<td>272</td>
<td>357</td>
<td>0.76</td>
<td>1508</td>
</tr>
<tr>
<td>VI</td>
<td>74</td>
<td>288</td>
<td>362</td>
<td>0.80</td>
<td>1744</td>
</tr>
</tbody>
</table>

Table 1.3: Data sorted by type or irregularity

5 Analysis: not all irregularities are equal

5.1 General comments

Tables 1.2 and 1.3 show that the children’s errors are not equally distributed among all classes of verbs. A chi-square test was performed on both tables and confirmed that the differences in both tables are significant (p<0.05 for both distributions)\(^6\). Thus, there is a correlation both between conjugation class and verbs’ CUR and between type of irregularity and verb’s CUR.

As for conjugations, the first conjugation had the lowest global CUR and the third conjugation the highest. The good results for the third conjugation can be easily explained considering that, as mentioned in Section 2, all verbs in this conjugation undergo some alternation, so it is fully predictable that they will undergo either raising or diphthongization. Interestingly, no errors were found regarding raising and the only errors in the third conjugation verbs were in the verbs dormir (‘sleep’, 5 errors) and venir (‘come’, 15 errors), the latter showing both [ie] diphthongization and velar insertion. First and second conjugation verbs are much more unpredictable and, thus, are expected to show lower CUR. However, the counts for both child and adult production for the 2nd conjugation are much higher, since this conjugation includes very common verbs, such as tener (‘to have’), querer (‘to want, to love’) or poner (‘put’). In contrast, first conjugation is the default class and contains more types with fewer tokens.

As for type of irregularity, as mentioned, raising verbs have a global CUR of 1: no child produced any errors on these verbs, in spite of being the class for which the children received the least input from the adults\(^7\).

\(^6\)The test was performed on 3x2 and 4x2 tables, where the row categories were either conjugation class or type of irregularity and the columns categories the incorrect and correct tokens.

\(^7\)As a reviewer notes, there are much fewer tokens for this class and, thus, it would be suitable to have more data. However, the table does reflect the real distribution of verbs.
The irregularity in which the children produced more errors was [ue] diph- tongization, while [ie] diphotingization and VI stay in the middle. In these last three classes, there seems to be a correlation between input from adults and CUR.

There are several cases in our data in which the children have a perfect CUR of 1 for a certain verb, despite the fact that it did not occur at all in the input they received from adults in the transcriptions. That was the case for the following children and verbs: Irene and apretar (‘press’), María and colgar (‘hang’), Juan and nevar (‘snow’) and servir (‘serve’), Koki and vestir (‘dress’) and Eduard and temblar (‘shake’).

5.2 Statistical analysis

A linear regression analysis was performed on all the verb types which occurred in the overall children speech at least ten times, which was the case for 20 out of the 50 verb types in our data.

The goal of the regression was to predict each verb’s CUR from the following predictor variables: conjugation, type of irregularity and input from adults. Best subsets regression was performed and the best model according to Mallow’s Cp statistics contained the following two variables: (1) input from adults and (2) whether the verb belongs to the first conjugation or not. With that model, highly significant results were obtained (F=6.1, p = 0.01007).

Figure 1.1 shows a plot of CURs and adult input for every verb included in the analysis. The two lines represent the fitted lines predicted by our model: the higher line corresponds to non-first conjugation verbs and the lower line to first-conjugation verbs. That is, our model predicts that, for a particular verb, the more a child hears the verb, the higher its CUR. Moreover, it also predicts an asymmetry between first conjugation verbs and the rest of verbs, the former having lower CURs. This is consistent with the observations above: third conjugation verbs are predictable, since they always undergo some change, second conjugation verbs are very frequent in the adult and children production. First conjugation verbs have neither of these properties and, thus, children make more errors with these verbs.

5.3 Theoretical implications

The main findings that arise from the data and the statistical analysis are the following:

*Although this might be due to a sampling effect, the fact that the verb did not occur in the adult speech in our sample means that it occurred rarely.

*Since the adult count data did not seem to be normal, it was transformed using a base 2 logarithm. After the transformation, the Shapiro-Wilk test yields a 0.902 probability that the data is normal.
1. Many instances of overregularizations were found (345 tokens out of 1477, global CUR = 0.81). No instances of overirregularizations were found.

2. The two main predictors of a verb’s CUR are (a) adult input and (b) membership to the first conjugation.

3. There are significant differences in the verbs’ CUR according both to the conjugation they belong to and the type of irregularity they present. It is particularly striking that verbs with the raising alternation have a global CUR of 1.

4. Some verbs have a perfect CUR of 1, although there was no adult input for that verb in our data.

While findings (1) and (2a) are compatible with both the Words and Rules and the Rules and Competition model, findings (2b), (3) and (4) pose a challenge to the Words and Rules approach. According to this model, all irregularities are individually stored in memory. Therefore, they predict that overirregularizations should not occur, consistently with what the data is showing. It is also not unexpected that CURs correlate with adult input, since memory storage might be dependent on frequency of exposure. In contrast, it is unexpected under this model that belonging to a class or
showing a particular irregularity should affect the verb’s CUR. Also, it is
not expected that children make no mistake at all with irregular verbs rarely
seen in the input. The Words and Rules model might be useful to explain
past tense acquisition in English since the irregular patterns are not so clear
or common (although see McClelland and Patterson (2002) and Yang (2002)
for criticism). However, in Spanish, patterns and classes clearly exist within
the irregular verbs and, thus, it seems completely inadequate to assume
children store each verb individually in their mental lexicon, without taking
into account the verb’s conjugation or particular irregularity. A reviewer
correctly points out that this data shows that the regular-irregular distinc-
tion is not enough and that more fine-grained distinctions are necessary,
such as a ‘semi-regular’ category. It is precisely the ‘semi-regular’ verbs or
the regular processes within irregular verbs that the Rules and Competition
approach can capture, while the Words and Rules cannot.

All the findings are compatible with the Rules and Competition model,
which predicts that CURs should be affected by membership to a class and
by input frequency. The perfect CUR of third conjugation verbs with a
raising alternation is particularly interesting. This finding is fully consistent
with Bybee and Pardo’s (1981) study in which subjects mostly introduced
an alternation for nonce third conjugation verbs. That is, both children and
adults encountering a new verb in the third conjugation seem to be able to
reliably determine both the probability that the verb belongs to a particular
class (i.e. \( P(x \in S) = 1 \)) and that a particular irregular rule needs to apply
\( (P_R = 1) \), possibly depending on the phonological form of the verb. I leave
for future work to test whether the two specific predictions of the Rules and
Competition model hold for the data I presented. However, since classes
within irregularities play such an important role in this model, it is a much
more promising way of thinking about how children can learn the persistent
irregularities they encounter in Spanish verbal morphology.

6 Conclusion

This paper has analyzed data from acquisition of irregular patterns in Span-
ish verbal morphology. Three-hundred and forty-five tokens of overreg-
ularization in child speech were gathered and a statistical analysis was
performed. The statistical analysis revealed significant differences in the
verbs’ Correct Usage Rate depending on conjugation and type of irregu-
larity. Moreover, the best statistical model proposed adult input and first
conjugation membership as predictors for a verb’s CUR. The data presented
here presents challenges for the Words and Rules model, while it is consist-
tent with a model that takes into account classes to explain irregularities,
such as the Rules and Competition model.
Bibliography


