Spanish diminutive formation without rules or constraints

DAVID EDDINGTON

Abstract

Spanish diminutive formation is analyzed in terms of analogy, more precisely within the computationally explicit framework of analogical modeling of language (AML: Skousen 1989, 1992). Accordingly, all known diminutives are presumed to be stored in the mental lexicon as completely formed units with associative links to their base forms. 2460 diminutive types were identified in various corpora and served as both the analogical database and the test items. When memory is unhampered by noise in the system, the probability that a previously known form will be chosen as the diminutive of its base is 100%. However, a simulation was performed in which all of the 2460 diminutives were treated as if they were previously unknown. The analogical influence of other forms allowed the correct diminutive form to be chosen in 92% of the cases. What is more, roughly half of the errors were found to be actually attested forms, which raises the success rate to 96%. Another simulation was performed that demonstrates that individual and dialectal differences in diminutive formation arise due to differing contents of the mental lexicon across speakers, as well as to the influence of competing gangs of phonologically similar base forms.

1. Introduction

The formation of diminutive variants of nouns, adjectives, and certain adverbs is a highly productive process in Spanish. Diminutives express concepts such as familiarity, small size, and disdain (see Zuluaga 1993 for a discussion of the semantics of diminutives). However, the purpose of the present study is not to investigate their semantic traits, but rather to account for the allomorphy of the diminutive suffix. Several suffixes exist (-ito, -illo, -zuelo, -ico, -uco), but -ito is the most commonly occurring.

Linguistics 40-2 (2002), 395-419
and the one that most discussion of the subject has focused on. For this reason, only the allomorphy of -ito will be considered.

Duminative formation has been the topic of a number of investigations. For example, Jaeggi (1980) discusses diminutives in Uruguayan Spanish from a classical generative standpoint, while Crowhurst (1992) and Prieto (1992) argue that they are to be dealt with in terms of prosodic constraints. Elordieta and Carreiras (1996) provide an analysis within optimality theory, while Ambadiang (1996, 1997) makes the case that diminutive formation belongs to the realm of morphology instead of phonology. Harris's (1994) paper is primarily a critique leveled at Crowhurst’s account, although many of his points are equally applicable to Prieto’s work as well. He argues that the high degree of lexical idiosyncrasy between competing diminutive forms suggests that they are not predictable on a phonological basis. Instead, he proposes that each word is marked in the lexicon as to which diminutive form(s) it will take (1994: 185).

In one regard, the present study follows Harris. That is, if one assumes that a base specifies which diminutive(s) it will take, that is similar to saying that the base is associated with its diminutive form, both of which have individual representation in the mental lexicon. The difficulty with this position is that it cannot account for the productive aspect of diminutive formation. Since base forms cannot be learned with their diminutive allomorphy presupposed, how does one go about producing a diminutive form s/he has never before heard or read? There must also be some mechanism for production, and phonological factors appear to play an important part in determining the phonological shape a diminutive form will take.

The purpose of the present paper, then, is to demonstrate that diminutive formation may be accounted for without recourse to highly abstract underlying representations, rules, or constraints, but by analogy to other fully specified pairs of bases and their corresponding diminutives in the mental lexicon. The dialectal and individual variability that exist in regard to diminutive allomorphy will also be accounted for. The remainder of the paper is structured as follows. Sections 2 and 3 lay out the theoretical background and the framework upon which the present study is based. Sections 4 and 5 describe a database of diminutive forms that was compiled from a search of about 51 million words. The allomorphy displayed by each diminutive is also discussed. The resulting database is an essential part of the analogical simulation of diminutive formation that is described in section 6. An example of how analogy may account for dialectal differences in diminutive formation appears in section 7.

2. Theoretical background

In the traditional generative approach to language, the lexicon is presumed to contain morphemes, and only those aspects of words that are unpredictable. Morphologically complex words are assembled, and predictable features of a word are added by means of rules. This state of affairs is deemed necessary based on the assumption that the amount of storage space available in the brain is limited. Accordingly, language acquisition is a matter of tacitly deriving the rules of a language based on the linguistic input received, in combination with genetically inherent linguistic abilities. This view results in a very minimal lexicon and requires a great deal of computation. A difficulty with this theoretical stance is that much of the machinery required for computation, such as abstract underlying representations, the cycle, and underspecification have been called into question (Burzio 1996; Cole 1995; Cole and Hualde 1998; Steriade 1995). In addition, psychological correlations to such mechanisms is highly dubiosus (Eddington 1996b; Lamb 2000).

The opposing view is that the lexicon includes vast amounts of stored information that is redundant and predictable, including detailed phonetic information about individual word tokens (Brown and McNeill 1966; Bybee 1994; Pisoni 1997). This possibility was suggested at an early period in generative history (Halle 1973; Jackendoff 1975), and several more recent theoretical proposals assume that most known words are stored as wholes in the lexicon (Butterworth 1983; Bybee 1985, 1988, 1998; Stemberger 1994). The psychological literature also contains empirical evidence to support massive lexical storage (e.g. Alegre and Gordon 1999; Baayen et al. 1997; Bybee 1995; Manelis and Tharp 1977; Sereno and Jongman 1997). In fact, storage may go beyond individual words and encompass recurrent word combinations as well as entire phrases (Bod 1998; Bybee 1998; Pawley and Syder 1983). It also appears that storage limitations on memory are not as problematic as previously supposed. For instance, Palmer et al. (1993) and Goldinger (1997) provide evidence that suggests that individual word tokens are stored in long-term memory.

This position, nevertheless, is not without difficulties of its own. Language is characterized by its productivity. If all forms are merely listed, how are new and previously unknown forms processed? Generally, those who maintain massive storage suggest an analogical process of some sort to account for productivity, but the exact nature of the analogical process invoked is, more often than not, left unspecified. The present study incorporates an explicit model of analogy that fills this void. It entails storage of fully specified pairs of bases and their corresponding
 diminutives, and a precise procedural algorithm for choosing the correct diminutive allomorph when the diminutive form is novel or temporarily inaccessible from memory.

3. Analogical modeling of language

Analogical modeling of language (AML) is a model designed to predict linguistic behavior on the basis of stored memory tokens (Skousen 1989, 1992, 1995, 1998). In this regard, it is similar to other exemplar-based models (Aha et al. 1991; Medin and Schaffer 1978; Riesbeck and Schank 1989; see Shanks 1995 for an overview of exemplar models; see Daelemans et al. 1994 for a comparison of AML and Aha et al.). AML makes its predictions on the basis of a given context. A given context is a set of variables that represents linguistic information about the entity whose behavior is being predicted. These variables may represent a phoneme in a certain position in a word, a part of speech, or a sociolinguistic or morphological variable. The reader is referred to Skousen (1989, 1992) for a detailed treatment of the AML algorithm, but a brief sketch of the model is in order.

For the sake of simplicity, let us assume that the given context contains information about a single word whose behavior we want to predict. AML searches the database (which represents the mental lexicon) for words that share variables with the given context, and creates groups of database items called subcontexts. Of course, words that share more variables with the given context will appear in more subcontexts. Subcontexts are further combined into more comprehensive groups called supercontexts. Upon inspection, some subcontexts will be homogeneous, that is, the members “agree” or exhibit the same behavior. (Behavior in this sense could mean that the members are all of the same syntactic class, take the same suffix, undergo the same phonological process, etc.) Other subcontexts will have “disagreements” in that they contain members with differing behaviors; these subcontexts are said to be heterogeneous. By minimizing disagreements and eliminating members of heterogeneous subcontexts, database items belonging to the most clear-cut areas of contextual space (homogeneous subcontexts) are available to exert their influence on the behavior of the given context.

Three important effects result from the application of AML’s algorithm (Skousen 1995: 217). The gang effect is obtained because when there is a large group of items that are similar to the given context, each member is available as a potential analog. Database items that have a great deal in common with the given context will appear in many different subcon- texts, and have a greater chance of affecting the behavior of the given context in comparison to those items that have less in common. This is called proximity. Finally, heterogeneity occurs when an item in the database is eliminated from consideration as an analog because there is another item, with a different behavior, that is closer to the given context.

The analogical set is arrived at once all members of heterogeneous subcontexts have been eliminated. AML uses the items in the analogical set to calculate the probability that the given context will be assigned one of the behaviors manifest in the items in the database. In general, what AML calculates is that the behavior of the words most similar to the given context predicts the behavior of the given context, although the behavior of less similar words has a small chance of applying, as long as those words appear in homogeneous subcontexts. It is important to note that AML predicts the behavior of one given context on the basis of the behavior of lexical items in the analogical set. All predictions are made locally, and no global generalizations are abstracted from the data.

There are two ways in which the analogical set may be used (Skousen 1989: 82). The first, called selection by plurality, is used to determine the “winner.” Accordingly, the most commonly occurring behavior in the analogical set is applied to the given context. This is similar to the way in which a connectionist model overcomes competing influences and settles on a single output. In a nearest-neighbor approach that identifies more than one neighbor, the behavior demonstrated by the majority of the nearest neighbors is declared the winner.

Of course, not all research questions involve deciding which behavior ultimately beats its competitors. Measuring leakage between behaviors is often of theoretical interest as well. Extreme cases of leakage occur when one word may exhibit two or more behaviors. For example, Wulf (1998) demonstrates how AML is able to predict leakage between alternating plural forms of certain low-frequency German nouns. In less extreme cases, leakage indicates the direction in which slips-of-the-tongue and/or language change may progress. Random selection allows one to determine if an item is mostly surrounded by other items with the same behavior, or the degree to which there are other items with different behaviors bearing similarities to the item in question. Random selection uses the probabilities calculated by the algorithm, that a specific behavior will apply to a given item. It essentially involves randomly selecting one of the members of the analogical set and applying the behavior of that member to the given context. Behaviors that are more frequent in the analogical set have a higher probability of applying.

These two types of selection may also apply to nearest-neighbor models in a similar fashion. Assume, for example, that five neighbors have been
chosen for a given context, one of which has behavior A and four
behavior B. Nearest-neighbor models most often select by plurality
and would declare B to be the winner. However, random selection shows
that the given context is not completely surrounded by other items with
behavior B. It also demonstrates some leakage toward behavior A.

Since behavior is determined in terms of a given context, no global
characterization of the data is made, such as the case with rule,
constraint, and prototype approaches. This implies that the variables
that may be important in determining the behavior of one given context
may be not be important in determining the behavior of a different one (see

Perhaps the most attractive part of an analogical approach is its
simplicity. It is based on the fairly uncontroversial idea that words are
stored in the mind and retrieved as necessary. That groups of similar
words can affect the behavior of other words with similar characteristics
is well attested in the psycholinguistic literature (e.g. Bybee and Slobin
1982; Stemberger and MacWhinney 1988). There is also ample evidence
that behavior may be based on stored exemplars (Chandler 1995;
Medin and Schaffler 1978; Nosofsky 1988).

4. Selection of the database

In an analogical approach to language, a database of linguistic forms is
needed from which analogs may be selected. For this reason, a database
of existing diminutives was compiled. There are, however, other reasons
for considering a large number of instances when attempting to account
for a linguistic phenomenon. Basing an analysis on a limited number of
examples is always risky since one is often predisposed to find examples
that coincide with one’s particular preconceived assumptions and to
overlook others. For example, Morin (1999) demonstrates that the cri-
teria proposed to distinguish between Spanish words that end in word
markers and those that do not are not supported when a much larger
number of examples is considered. In a similar vein, Eddington (1996a)
finds that when a large number of instances is considered, the relationship
between certain derivational suffixes and diphthongization in Spanish
word stems is far from binary, as previous investigation had considered
it to be.

In the literature on diminutives, it is often unclear from what source
the authors derive the diminutives on which they base their analyses;
most dictionaries include few citations for diminutive forms, and those
that do appear often have a lexicalized meaning apart from that of the
diminutive. In Prieto’s study (1992), some diminutive forms were elicited
from native Spanish speakers by means of a survey. In the present study,
diminutives were extracted from several corpora: the Alameda and Cueto’s
frequency dictionary (1995: 5 million words), the LEXESP tagged fre-
quency dictionary (Sebastían, Cueto, and Carreiras n.d.: 3 million
words), a corpus of spoken peninsular Spanish (Marcos Marín, n.d. a;
1 million words), a corpus of Argentine Spanish (Marcos Marín, n.d. b;
2 million words). In addition to these sources, Mark Davies of Illinois
State University graciously provided me with the diminutive forms from
his corpus project totaling 39.8 million words. Therefore, the resulting
diminutives were gleaned from a pool of 50.8 million words. Both written
and spoken registers are included, although spoken sources compose only
about 7% of the sample. Samples from every Spanish-speaking country
(with the exception of Equatorial Guinea) are included, but no effort was
made to balance each country’s representation in the database. A total
of 2466 diminutive types were identified. Type frequency was used in the
present study since research suggests that type frequency is more relevant
to the analogical extension of a pattern than token frequency (Bybee

5. Diminutive allomorphy in the database

Each diminutive was grouped by hand according to the allomorphic
relationship that it has with its base form. In this way, 13 major allo-
morphs were identified. However, 13 of the database items demonstrate
unusual changes in the diminutives that are not found in any of the 13
groups and are found in three or fewer items. For example, the proper
names Antonio and Antonia have diminutives with a palatalized nasal:
Antonito, Antonita. The diminutives of caliente6 ‘hot’ and independiente
‘independent’ are odd in that they lose their diphthongs, yielding calentito
and independentitas. This is an unusual outcome given the fact that
every other word with a diphthong maintains it in the diminutive:
prieto ‘tight’ > prietito, cuento ‘story’ > cuentito. Additionally, other
diminutives were found that must be considered isolates since they do
not fit into any of the 13 major categories described above: futbol
‘football’ > futbito, pie ‘foot’ > piecicito, café ‘coffee’ > cafelito, cafetito,
dos ‘two’ > dositos, José ‘Joseph’ > Joselito, azucar ‘sugar’ > azucarito,
lejos ‘far away’ > lejeccitos, and diagnoso ‘diagnosis’ > diagnostito. These
items were also categorized and included in the database. In addition,
the base form of valsicito ‘waltz’ was included in two different categories
since it was impossible to determine whether the base form was val or vale. Once the database was computed, six items were chosen at random and deleted, in order to yield a number of items divisible by ten. The result was a database containing 2460 different diminutives.

With the exception of the odd items just discussed, the remaining 99.5% of the database items fall into one of 13 major categories. A circled V or S indicates that that particular element of the base form does not appear in the diminutive form:

(1) -VITO(S): -ito(s) is added to the singular base form, replacing the final vowel: minuto ‘minute’ > minútilo, elefante ‘elephant’ > elefántito.

(2) -VITA(S): -ita(s) is added to the singular base form, replacing the final vowel: galleta ‘cookie’ > galletita, Lupe ‘proper name’ > Lupita.

(3) -VECITO(S): -ecito(s) is added to the singular base form, replacing the final vowel: vidrio ‘glass’ > vidriecito, quieta ‘peaceful’ > quietecito.

(4) -VECITA(S): -ecita(s) is added to the singular base form, replacing the final vowel: yerba ‘grass’ > yerbecita, piedra ‘stone’ > piedrecita.

(5) -CITO(S): -cito(s) is added to the singular form: traje ‘suit’ > trajecito, pastor ‘shepherd’ > pastorcito.

(6) -CITA(S): -cita(s) is added to the singular base form joven ‘young girl’ > jovencita, llave ‘key’ > llavecita.

(7) -ITO(S): -ito(s) is added to the singular base form: normal ‘normal’ > normalito, Andrés ‘Andrew’ > Andrestito.

(8) -ITA(S): -ita(s) is added to the singular base form: nariz ‘nose’ > naricita, Isabel ‘Isabella’ > Isabelita.

(9) -ECITO(S): -ecito(s) is added to the singular base form: pez ‘fish’ > pececito, rey ‘king’ > reyecito.

(10) -ECITA(S): -ecita(s) is added to the singular base form: flor ‘flower’ > florecita, hel ‘light’ > lucecita.

(11) -SITOS: -itos is added to the singular base form, replacing the vowel and false plural morpheme: lejos ‘far away’ > lejitos, Marcos ‘Mark’ > Marquitos.

(12) -SITAS: -itas is added to the singular base form, replacing the vowel and false plural morpheme: Lucas > ‘Luke’ Luquitas, garapatas ‘tick’ > garapatitas.

(13) -CITA(S): -cita(s) is added to the singular base form, replacing the final vowel: jamona ‘fat woman’ > jamoncita, patrona ‘patron saint’ > patroncita.

Table 1 categorizes the contents of the database in terms of a number of important features.

It is important to note that in some cases, diminutive formation would appear to produce sequences of [ji] in the rhyme of the penult syllable: [lim pij] ‘clean’ > [lim pij to], [ar ma rijo] ‘closet’ > [ar ma rijo]. However, [ji] is a nonoccurring rhyme in Spanish, which is why the glide does not appear (Elordieta and Carreiras 1996: 55; Harris 1994: 182; Prieto 1992: 196). Instead, the corresponding diminutives are [lim pi to] and [ar ma ri to]. Sequences of [i+i] are also attested in one diminutive in the database (titita ‘aunt’), but even this is unusual enough that the sequence results in a single high vowel in alternate diminutive forms: tito ‘uncle’, tita ‘aunt’.

An analysis of the resulting database reveals a number of interesting facts. First, it contains quite a few doublets, that is, different diminutive forms of the same base form (Table 2). This is not unusual given the extensive corpora from which the diminutives were gleaned, along with the fact that the database cuts across many dialects. As Prieto (1992: 170) indicates, one of the major dialectal differences involves how the diminutives of bisyllabic words containing one of the diphthongs [je]/ or [we]/ are formed; 44% of the doublets have stems containing diphthongs of this sort.

Another thing that is supported by the database is the tendency for bisyllabic -e-final words to form diminutives with the addition of -cito/a (category 5 and 6; Crowhurst 1992; Elordieta and Carreiras 1996; Prieto 1992). Only 15 of the 90 base words of this type have diminutives that run counter to this tendency (e.g. leche ‘milk’ > lechita). On the other hand, base words with three or more syllables generally take diminutives with the addition of -rito/a (category 1 and 2). The sole exception found is retoque ‘retouch’ > retoquecitio. However retoquecitio is also an attested form (see Table 3). In contrast to the evidence from the corpora, Prieto’s (1992: 174) 12 informants produced diminutives, such as retroquecitio, as possible variants of 12 of the 13 test words they were presented (e.g. estuche ‘case’ > estuchecito; chocolate ‘chocolate’ > chocolutecito).

Perhaps the oddest of all diminutive forms are those that appear to involve inflexion of -i- before a word final -or or -ar: Victor ‘Victor’ > Victito, azúcar ‘sugar’ > azúzucar, ambor ‘amber’ > ambirtar. Their unusual status is evident in that rule accounts of these forms require modifications in order to yield the correct outcome (Crowhurst 1992; Prieto 1992; Jaeggi 1980). Furthermore, exactly what Spanish dialect one may find such diminutives in is unclear in the literature. The fact that not one instance of this kind of diminutive was found in 51 million words of text.
### Table 1. Summary of database items by category

<table>
<thead>
<tr>
<th>Number</th>
<th>Category:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>988</td>
</tr>
<tr>
<td>Masculine</td>
<td>979</td>
</tr>
<tr>
<td>Feminine</td>
<td>2^*</td>
</tr>
<tr>
<td>No gender</td>
<td>7^b</td>
</tr>
<tr>
<td>-o final</td>
<td>910</td>
</tr>
<tr>
<td>-a final</td>
<td>1^e</td>
</tr>
<tr>
<td>-e final</td>
<td>73</td>
</tr>
<tr>
<td>-r final</td>
<td>-</td>
</tr>
<tr>
<td>-n final</td>
<td>-</td>
</tr>
<tr>
<td>-l final</td>
<td>-</td>
</tr>
<tr>
<td>-d final</td>
<td>-</td>
</tr>
<tr>
<td>-s final</td>
<td>-</td>
</tr>
<tr>
<td>Other final</td>
<td>4^v</td>
</tr>
<tr>
<td>/je/ or /we/ in root</td>
<td>46</td>
</tr>
<tr>
<td>Bislavvlic -e final</td>
<td>10^g</td>
</tr>
<tr>
<td>3+ syllables -e final</td>
<td>63</td>
</tr>
</tbody>
</table>

---

a. foto, mano.
b. E.g. abajo, adelante, callando.
c. papá.
d. bikini, güisqui, punqui, Itaki.
e. E.g. chisme, Pepe, tigre.
f. E.g. mapa, problema, sistema.
g. E.g. ahora, arriba, encima.
h. grande, lesche, Maite.
i. All are bisyllabic.
j. Ten of the remaining words end in [jo], leaving iento, lleno, paso, sayo.
k. mano.
l. allá.
m. carro, rio.
n. opel.
o. David.
p. bucy, cuij, gojey, güisqui, papú.
q. bucy, diente, mueble.
r. retoque.
s. Four of the remaining words end in [ja], leaving hecha, lengua, mano, seca.
t. pared.
u. suerte.
w. Adrián, Juan, pompón, ratón.
x. papá.
y. arruz, bistec, chalet, coñac, copey, desiz, lápiz, maiz, reloj.
z. Pilar.
A. Inversión.
B. verdad.
C. Glaidy.
D. cruz, nariz.
E. All monosyllabic.
F. tren.
G. sol.
H. flux, vale.
I. per, rey.
J. tos.
K. flor.
L. cruz, luz, voz.
M. apenas.
N. antojos, Carlos, cumpleaños, lejos, Marcos.
O. Lucas.
P. allá, papá.
Q. huevona.
R. gafas, garrapatas, Mercedes.
S. apenas.
T. vals.
suggests that if they exist at all, they are extremely uncommon, or possibly low-prestige forms that would not be found outside the familiar spoken register.

6. The role of the database in analogy

The critical part of this study is determining the extent to which the AML algorithm is able to account for diminutive allomorphy. To this end, information about the base form of each diminutive was converted into a series of variables. The base form is the uninfluenced noun, adjective, gerund, or adverb from which the diminutive is derived. For example, the base form of *ratoncito* 'little mouse' is *ratón*. The variables were chosen in accordance with the principles of distinguishability and proximity (Skousen 1989: 52). Proximity involves choosing those variables that are closest to the phenomenon that is being predicted. Since diminutive formation occurs word-finally, the most relevant features are those that appear toward the end of the word. Therefore, the variables included the following information about each base form: (1) the existence, and stressed or unstressed status, of the final three syllables; (2) the gender of the word: masculine, feminine, or none in the case of adverbs and gerunds; (3) the word's final phoneme; (4) the phonological content of the antepenultimate rhyme and final two syllables of the word.

The criterion of distinguishability suggests that each word should be represented with enough variables that it is unique from every other word in the database. One thing that makes it impossible for each database item to have a unique set of variables is the existence of doubles (Table 2). For example, both *ratoncito* and *ratoncito* are attested diminutives of the same base form *ratón*. Therefore, both forms are represented with the same set of variables. Of course, one entry for *ratón* specifies that its diminutive is of the sort found in category 5 (i.e. *ratoncito*, see section 5), while the other entry indicates that its diminutive is of the type found in category 7 (i.e. *ratoncito*). However, category markers are not treated as variables when analogies are made; instead, they specify the kind of relationship that is found between a base and its diminutive.

In section 5, 13 major categories of diminutive type are described. For example, both *pueblo* and *cuerno* are specified as members of category 3 in the database. This means that the morphophonemic relationship that holds between *pueblo* 'town' and its diminutive *pueblecito* is the same one that holds between *cuervo* 'crow' and *cuernecito*. Therefore, if these two words are chosen as analogs for the word *cuervo* 'crow', diminutivization by analogy is assumed to take the form of a proportional analogy:

<table>
<thead>
<tr>
<th>Base form</th>
<th>Diminutive A</th>
<th>Diminutive B</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>altar</td>
<td>altarito</td>
<td>altar</td>
<td></td>
</tr>
<tr>
<td>Antonia</td>
<td>Antonitita</td>
<td>Antonio</td>
<td></td>
</tr>
<tr>
<td>Antonio</td>
<td>Antontito</td>
<td>Antonio</td>
<td></td>
</tr>
<tr>
<td>bote</td>
<td>botito</td>
<td>jar</td>
<td></td>
</tr>
<tr>
<td>café</td>
<td>cafetito</td>
<td>coffee</td>
<td></td>
</tr>
<tr>
<td>caliente</td>
<td>calentito</td>
<td>hot</td>
<td></td>
</tr>
<tr>
<td>carne</td>
<td>carnita</td>
<td>meat</td>
<td></td>
</tr>
<tr>
<td>carro</td>
<td>carrito</td>
<td>car</td>
<td>chauffeur</td>
</tr>
<tr>
<td>chófer</td>
<td>choferito</td>
<td>cross</td>
<td></td>
</tr>
<tr>
<td>cruz</td>
<td>crucita</td>
<td>neck</td>
<td></td>
</tr>
<tr>
<td>cuello</td>
<td>cuellito</td>
<td>bill</td>
<td></td>
</tr>
<tr>
<td>cuenta</td>
<td>cuentita</td>
<td>story</td>
<td></td>
</tr>
<tr>
<td>cuento</td>
<td>cuentito</td>
<td>horn</td>
<td></td>
</tr>
<tr>
<td>cuerno</td>
<td>cuernito</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cuerpo</td>
<td>cuerpicito</td>
<td>large</td>
<td></td>
</tr>
<tr>
<td>grande</td>
<td>grandecita</td>
<td>whisky</td>
<td></td>
</tr>
<tr>
<td>gáisqui</td>
<td>gáisquicito</td>
<td>gáisqui</td>
<td></td>
</tr>
<tr>
<td>hierba</td>
<td>hierbicita</td>
<td>grass</td>
<td></td>
</tr>
<tr>
<td>hierro</td>
<td>hierrito</td>
<td>iron</td>
<td></td>
</tr>
<tr>
<td>hombre</td>
<td>hombruito</td>
<td>man</td>
<td></td>
</tr>
<tr>
<td>huevito</td>
<td>huevito</td>
<td>egg</td>
<td></td>
</tr>
<tr>
<td>indio/a</td>
<td>indio/a</td>
<td>Indian</td>
<td></td>
</tr>
<tr>
<td>Jorge</td>
<td>Jorigito</td>
<td>George</td>
<td></td>
</tr>
<tr>
<td>José</td>
<td>Joselito</td>
<td>Joseph</td>
<td></td>
</tr>
<tr>
<td>Juan</td>
<td>Juantito</td>
<td>John</td>
<td></td>
</tr>
<tr>
<td>juego</td>
<td>jueguito</td>
<td>game</td>
<td></td>
</tr>
<tr>
<td>lento</td>
<td>lentito</td>
<td>slow</td>
<td></td>
</tr>
<tr>
<td>lleno</td>
<td>llenito</td>
<td>full</td>
<td></td>
</tr>
<tr>
<td>manueto</td>
<td>manueta, manueto</td>
<td></td>
<td></td>
</tr>
<tr>
<td>muerto</td>
<td>muерcto</td>
<td>dead</td>
<td></td>
</tr>
<tr>
<td>nieta</td>
<td>niétita</td>
<td>granddaughter</td>
<td></td>
</tr>
<tr>
<td>papá</td>
<td>papito, papaito</td>
<td></td>
<td></td>
</tr>
<tr>
<td>paso</td>
<td>pasito</td>
<td>step</td>
<td></td>
</tr>
<tr>
<td>piedra</td>
<td>piedrita</td>
<td>stone</td>
<td></td>
</tr>
<tr>
<td>pieza</td>
<td>piecita</td>
<td>piece</td>
<td></td>
</tr>
<tr>
<td>pueblo</td>
<td>pueblito</td>
<td>town</td>
<td></td>
</tr>
<tr>
<td>puerta</td>
<td>puertita</td>
<td>door</td>
<td></td>
</tr>
<tr>
<td>quiebo/a</td>
<td>quiebeto/a</td>
<td>calm</td>
<td></td>
</tr>
<tr>
<td>ratón</td>
<td>ratoncito</td>
<td>mouse</td>
<td></td>
</tr>
<tr>
<td>río</td>
<td>ricto</td>
<td>river</td>
<td></td>
</tr>
<tr>
<td>rubio</td>
<td>rubito</td>
<td>blond</td>
<td></td>
</tr>
<tr>
<td>sueño</td>
<td>sueñito</td>
<td>sleep</td>
<td></td>
</tr>
<tr>
<td>tambor</td>
<td>tamboarto</td>
<td>drum</td>
<td></td>
</tr>
<tr>
<td>tiempo</td>
<td>tiempito</td>
<td>time</td>
<td></td>
</tr>
<tr>
<td>tren</td>
<td>trenito</td>
<td>train</td>
<td></td>
</tr>
<tr>
<td>viejo/a</td>
<td>viejito/a</td>
<td>old</td>
<td></td>
</tr>
<tr>
<td>viento</td>
<td>vientito</td>
<td>wind</td>
<td></td>
</tr>
<tr>
<td>vuelo</td>
<td>vuelito</td>
<td>flight</td>
<td></td>
</tr>
<tr>
<td>vuelta</td>
<td>vuelcito</td>
<td>revolution</td>
<td></td>
</tr>
</tbody>
</table>
(14) pueblo : pueblecito, cuerno : cuernecito : : cuervo : ??

Exactly how speakers derive the diminutive (e.g. infixation of -ecit-, or deletion of -o and suffixation of -ecito) is largely unimportant.

However, Bybee's (1988) conception of morphology as networks of links between stored lexical items suggests another way of viewing the analogical process. Consider Figure 1, which represents a very simplified state of affairs. The solid lines conjoining base forms and diminutives indicate phonological similarities between the stored bases and their diminutives, which have already formed links due to their semantic similarity. In like manner, the diminutive suffixes of each word are linked to each other, as well as to other diminutives. It is these relationships that are assumed when pueblo and cuerno are marked as taking category 3 diminutives. The dotted lines between cuervo, cuernecito, and pueblecito represent phonological similarities that are activated when pueblo and cuerno are chosen as analogs for cuervo.

The next step builds on Skousen's (1992) proposition that, once constructed, analogical sets may be stored. Or perhaps it is better to assume that the set is not stored per se, but that the members of the set come to form links with each other based on their shared similarities. Therefore, pueblo and cuerno have presumably cooccurred in other analogical sets in the past, which is why there is a link between the diphthong they have in common. Once cuernecito has been chosen as the diminutive of cuervo, it will form new links with pueblo and cuerno on the basis of their phonological similarities. It is in this way that connections are formed between words that are semantically and phonologically similar.

7. The AML simulation

A ten-fold cross-validation simulation was performed. This entailed dividing the database into ten groups of equal size. One group was then removed, and its members served as the test cases. The members of the remaining nine groups composed the training set from which analogs were sought. Each group served as the test set only once. If a member of the training set matched a test item exactly, it was not considered as a possible analog. In this way, the influence of one member of a doublet on another was eliminated. Selection by plurality (see section 3) was assumed, since the goal in this simulation was to predict a winner from among the possible outcomes.

Under these conditions, the AML algorithm assigned the wrong diminutive suffix to only 198 items, resulting in a success rate of 92%. What this indicates is that there is a great deal of analogical consistency: base forms that take the same diminutive suffix have many features in common, enough that the large majority of them can serve as analogs for each other under conditions of imperfect memory, or if the items are treated as novel. Some errors involved incorrect diminutive allomorphy, but correct gender markers: parte 'part' > *partita. Others erred in terms of gender assignment: carnel fem. 'buddy' > *carnaldito, sofa 'couch' > *sofita. Other errors entailed applying a suffix such as - pretendido/a to words without final vowels, such as verdad 'truth'.

Nevertheless, many of the errors appeared to be plausible diminutive forms. The doublets demonstrated this, in that errors on one member of the doublet almost always entailed assigning it the diminutive suffix of the other member. In addition, in 65% of the cases of misassignment of a doublet, the second most probable behavior was the correct one. These results occurred in spite of the fact that when tested, both members of a doublet were excluded from the database and were unable to serve as analogs for each other. In order to determine if other erroneous diminutive forms predicted by AML were actually well-formed diminutives in some dialect of Spanish, the World-Wide Web was consulted. All erroneous forms (with the exception of errors of the type found in verdad) were sought on Spanish-language pages. Of the 198 errors, attested forms of 104 were found, either as an attested doublet in the database (see Table 2) or on a Spanish-language web page (Table 3).

What this demonstrates is that 52% of the errors calculated by the model are not true errors, but merely alternative diminutive forms. This is a clear indication that the model has captured the essence of diminutive formation. When the unattested diminutive forms that the model predicts are subtracted from the total number of database items, the overall success rate of the model reaches 96.2%.

It would be desirable to be able to compare the results of the analogical simulation with the success rate of one of the generative approaches already cited. Regrettably, a straightforward and fair comparison of this
sort is not possible for a number of reasons. In none of the studies were
the rules and constraints designed to account for the full range of data
found in the database. Elordieta and Carreiras’s (1996) study is arguably
the extreme case in this regard; it only includes diminutives demonstrating
eight of the 13 major categories found in the database, it does not include
a discussion of bisyllabic bases containing the diphthongs /je/ and /we/,
and it makes no mention of the existence of alternative diminutive forms
of the same base. Another difficulty with making such a comparison is
that some analyses (Crowhurst 1992; Jaeggli 1980) only cover diminutive
allomorphy in a specific dialect.

The use of abstract formal mechanisms that are not surface-apparent
is also troublesome. There is no doubt that by means of formalisms such
as diacritics, underlying representations, and rule and constraint order-
ings, any of these analyses could easily be modified to account for all of
the diminutives in the database, but it would be of interest to determine
what predictive value these analyses have if their abstract aspects are
eliminated. Unfortunately, formal mechanisms are such an integral part
of these analyses that they may not be eliminated without severely ham-
pering the predictive power. For instance, Crowhurst distinguishes
between word-final -es that are epenthetic and those that are terminal
elements. Each one is associated with a different type of diminutive. Even
Harris (1994), who argues from a generative standpoint, feels that
Crowhurst’s use of a number of formal mechanisms in her study is
ad hoc. Many of Harris’s criticism is equally applicable to Prieto’s study
as well. As a result, he maintains that it is not possible to generate a
diminutive on the basis of the phonological shape of the base form.
Nevertheless, the present study indicates that an analogical approach is
able to achieve this goal.

One question that the simulation does not address is, which features
of the base form are most important in determining the form of the
diminutive. AML only makes predictions on the basis of a given context.
Therefore, an inspection of the analogical set for a given context allows
one to find the most relevant variables for that given context alone. That
is to say, an overall characterization of the data is not readily obtainable
in AML. Nevertheless, it may be computed using a different analogical
algorithm called TiMBL (Daelemans et al. 1999; see Eddington forthcoming
for a comparison of AML and TiMBL on a similar diminutivization
simulation). Accordingly, the most relevant variables, ordered from most
to least relevant, are (1) the stressed or stressless status of the final
syllable, (2) the gender of the base, (3) whether the base is monosyllabic
or not, and if not, the stressed/unstressed status of the penultimate
syllable, (4) what phoneme appears word-finally, as the nucleus or coda of
the final syllable, (5) whether the base has two or fewer syllables, and if not,
the stressed/unstressed status of the antepenultimate syllable, (6) the pho-
nome(s) in the coda of the penultimate syllable, if any, (7) the phoneme(s) in
the rhyme of the antepenultimate syllable, if any, (8) the phoneme(s) in the
onset of the final syllable, (9) the phoneme(s) in the onset of the penul-
timate syllable.

This hierarchy coincides to a great extent with the findings of other
studies on Spanish diminutives; the most relevant variables in the base
form are the number of syllables it contains, its stress pattern and gender,
and its final phoneme. However, one must keep in mind that this global
characterization does not preclude the possibility that a different hier-
archy may hold when predicting the diminutive form of an individual
base. For instance, in bisyllabic words containing /je/ or /we/ in the
penult nucleus (e.g. cuenta and hierba), the contents of the penult nucleus
is a much more important factor than it is for other words without these
diphthongs.
8. Variability between diminutive forms

One thing noted in most previous studies on Spanish diminutives is that there is some variability in choosing diminutive forms with one suffix or another (Crowhurst 1992; Harris 1994; Jaeggl 1980; Prieto 1992). Elordieta and Carreiras (1996), on the other hand, make no mention of diminutives that differ from those their analysis accounts for. Nevertheless, variation exists, both between dialects and within individual speakers. Consider the diminutives of words ending in -e, for example. In general, bisyllabic words of this type take diminutives that are formed by adding -cito/a to the base form (e.g. madre ‘mother’ > madrecita). However, bases with three or more syllables generally belong to the -cito/a category (e.g. comadre ‘godmother’ > comadrita). Nevertheless, there are exceptions to this generalization. Crowhurst (1992: 251) cites sangre ‘blood’ > sangrinita, mugre ‘filth’ > mugrinita, leche ‘milk’ > lechita, and hambre ‘hunger’ > hambrina. Harris (1994: 183) cites other exceptions: tigre ‘tiger’ > tigrinito, chile ‘chili’ > chilito, nene ‘boy’ > nenito. The database for the present study includes Pepito < Pepe, Matita < Maite, grandita < grande ‘big’, and chismito < chisme ‘gossip’. On the other hand, subjects in Prieto’s study (1992: 174) vacillated between -cito/a and -cito/a type diminutives whose bases have three or more syllables: chocolate > chocolatecito/estuche > estuchecito, comadre > comadrita/comadrecita.

Prieto (1992) and Crowhurst (1992) employ a number of different generative devices to account for alternating diminutive forms. For example, to account for some of the variation, Crowhurst (1992) suggests that some speakers have a minimal word template composed of two bisyllabic feet, while others do not. This is similar to Prieto’s (1992) position. Crowhurst explains the alternation between diminutives such as dientito ‘tooth’ and dientecito by proposing that in the former, the diphthong of the stem is resyllabified in the course of the derivation, in such a way that each of its components belongs to a separate syllable. In the case of dientecito, no such resyllabification occurs.

How is such variation accounted for in an analogical model? According to analogy, it is due to differences in the lexicon. Dialectal differences arise because, in the course of acquiring a language, a person adopts the diminutives that are commonly used in the surrounding speech community, and the form of these diminutives varies from dialect to dialect. To this point, this is a fairly tautological statement; dialectal variations exist because they do. However, the differing contents of the mental lexicon from dialect to dialect means that there is a different set of possible analogs on which to determine the diminutive form of new and previously unknown diminutives. An example should clarify this position.

Prieto (1992: 170) notes that one of the major dialectal differences has to do with the treatment of bisyllabic words containing the diphthongs /je/ and /we/ in the stem (e.g. diente > dientito or dientecito). The present database does not purport to represent any particular dialect; however, it may be employed to simulate dialectal differences. To this end, the database was modified. For dialect A, all masculine bisyllabic words with /je/ or /we/ in the stem were marked as taking diminutive forms ending in -cito/a, and all feminine words with the same characteristics were given diminutive forms ending in -cita. For example, the database entry for cuento was modified so that its diminutive would be cuentito. The entry for vieja ‘old’ was marked as having the diminutive viejita. For dialect B, these same words were considered to take -decito or -decita, depending on their gender (cuentecito, viejecita).

According to AML, if the diminutive forms of all of the items in the database are remembered with 100% accuracy, dialect A will produce all of these diminutives with -cito/a, and dialect B with -decito/a. In and of itself, this is hardly an interesting outcome. However, the way each dialect processes novel items is of interest. Table 4 contains the calculated probabilities that the diminutive form of several words (that do not appear in the database) will appear with either -decito/a or -cito/a allomorphy in the two simulated dialects.

Modifying the database for dialect A entailed eliminating the majority of the words that demonstrate -decito and -cito/a allomorphy. It is

<table>
<thead>
<tr>
<th>Base word</th>
<th>Dialect A</th>
<th>Dialect B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rejo</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Ruego</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Siervo</td>
<td>99.98</td>
<td>0</td>
</tr>
<tr>
<td>Trueño</td>
<td>99.99</td>
<td>0</td>
</tr>
<tr>
<td>Cuelga</td>
<td>100.0</td>
<td>0</td>
</tr>
<tr>
<td>Cuerda</td>
<td>99.95</td>
<td>0.05</td>
</tr>
<tr>
<td>Fiebre</td>
<td>95.48</td>
<td>3.74</td>
</tr>
<tr>
<td>Friega</td>
<td>99.95</td>
<td>0.05</td>
</tr>
<tr>
<td>Niebla</td>
<td>100.0</td>
<td>0</td>
</tr>
<tr>
<td>Nieve</td>
<td>87.76</td>
<td>11.39</td>
</tr>
<tr>
<td>Prueba</td>
<td>100.0</td>
<td>0</td>
</tr>
<tr>
<td>Suerte</td>
<td>72.28</td>
<td>7.92</td>
</tr>
<tr>
<td>Sierra</td>
<td>99.97</td>
<td>0</td>
</tr>
</tbody>
</table>
not surprising, then, that almost no analogical pressure is exerted by words of this type in the dialect A simulation. The diminutives produced by dialect B, in contrast, demonstrate more variability. Nine of the novel diminutives are favored to appear with -\(\grave{e}\)cito/a, with leakage toward -\(\grave{e}\)rito/a, and three favor -\(\grave{e}\)rito/a. The probability of a diminutive with either type of allomorphy is about equal for siervo 'servant'.

The results of this simulation suggest an empirically testable hypothesis. The diminutives formed from bisyllabic bases containing the diphthongs /je/ and /we/ will demonstrate much more variation in dialect B than in dialect A. Prieto's study (1992) suggests that Peninsular Spanish may approximate more closely dialect B, while Bolivian Spanish may reflect dialect A. Of course, any attempt to test this hypothesis should focus on the production of diminutives that are most likely to be novel and less likely to be forms that are known.

Not only does the phonological shape of the diminutives vary from one dialect to another, but individuals also demonstrate some degree of uncertainty regarding the diminutive form of certain words. From an analogical standpoint, this may be due to two sets of circumstances. In the first, the speaker has heard and/or produced two or more diminutives of the same base word, with different suffixes. In this case, the probability that one of the diminutive forms will be chosen is proportional to the number of times it appears in the lexicon, in comparison with the other form(s).

If the diminutive form of a base word is completely novel, or if it is temporarily un retrievable from memory, analogy will calculate the base's similarity to others that exist in the lexicon. If the word is completely surrounded by similar items that all form diminutives in the same manner, only one diminutive form will be produced. However, in some cases, gangs of similar items with different behaviors may compete with each other, resulting in variability or uncertainty (Table 5). Several examples of this may be gleaned from the simulation presented in section 6.

### 9. Conclusions

The present study assumes that all known diminutive forms are stored in the mental lexicon as completely formed entities. Under conditions of perfect memory, the probability that a known form will be chosen as the diminutive of its base is 100%. However, base forms that take diminutives with the same allomorphy demonstrate a great deal of phonological similarity. This allows the allomorphy of novel diminutives to be predicted with a high degree of accuracy as well. The AML algorithm is able to correctly predict the shape of most items tested. In addition, about half of the errors it does produce are actually attested forms, which further supports the notion that diminutive formation may be explained as an analogical process. According to this analysis, individual and dialectal differences arise due to differences in the diminutive forms that exist in the mental lexicon, and the influence of competing gangs of phonologically similar items.

#### Table 5. Examples of competing gangs on selected base forms

<table>
<thead>
<tr>
<th>Base form</th>
<th>Variant A</th>
<th>Prob. of variant A</th>
<th>Variant B</th>
<th>Prob. of variant B</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeque</td>
<td>yequito</td>
<td>67.22</td>
<td>yequecito</td>
<td>32.78</td>
</tr>
<tr>
<td>Jorge</td>
<td>Jorgecito</td>
<td>23.33</td>
<td>Jorgecito</td>
<td>74.95</td>
</tr>
<tr>
<td>pierna</td>
<td>piernita</td>
<td>40.81</td>
<td>piernita</td>
<td>59.19</td>
</tr>
<tr>
<td>cuervo</td>
<td>cuervito</td>
<td>33.38</td>
<td>cuervito</td>
<td>66.31</td>
</tr>
<tr>
<td>David</td>
<td>Davidito</td>
<td>58.06</td>
<td>Davidcito</td>
<td>43.14</td>
</tr>
<tr>
<td>Chevrolé</td>
<td>Chevrolito</td>
<td>54.79</td>
<td>Chevrcito</td>
<td>40.43</td>
</tr>
<tr>
<td>nieta</td>
<td>nietita</td>
<td>67.68</td>
<td>nietica</td>
<td>31.21</td>
</tr>
<tr>
<td>corte</td>
<td>cortito</td>
<td>45.03</td>
<td>cortecito</td>
<td>51.53</td>
</tr>
</tbody>
</table>

#### Notes

1. This study was carried out with the help of a grant from the National Science Foundation (NSF821929). Correspondence address: Department of Spanish and Portuguese, University of New Mexico, Ortega Hall, Room 255, Albuquerque, NM 87131, USA. E-mail: davie@unm.edu.
2. Actually, one of the pointers in the analogical set is chosen, but the role of pointers in the algorithm has not been discussed in this summary description.
3. A newer version of LEXESP exists that contains 5.5 million words.
4. Details about these corpora are available at http://mdaves.for.ilstu.edu/personal/texts.htm.
5. Some diminutives, such as clasesitas and tanquesito, were found with the suffix *sitos/a/s. These are obviously due to errors in dialects that do not distinguish [s] and [θ] and do not indicate an additional suffix. In these cases, the spelling was regularized.
6. Colmito is also attested.
7. In some words from groups 11 and 12, s represents what seems to be the plural morpheme, since it appears word-finally and follows a stressless vowel. In words such as compleånos, the word ends in the plural morpheme derivationally speaking (compleÅ + años 'complete + years') but is used to denote both the plural and singular.
8. Prieto (1992) considers only instances of /je/ and /we/ that alternate with /e/ and /o/ in morphemic relatives (e.g. buen + o 'good', bon + da'd 'goodness'). However, the database also contains some words with nonalternating diphthongs such as nieta and hueso.

9. Jæger (1980: 157, note 5) notes the variation but dismisses it: "In some cases, native speakers may be unsure as to which form is the grammatical one, but saying this is very different from saying that more of that one form is ever allowed."

10. In the case of words ending in -e, either -cito/a or -te/cito/a may be applied to yield diminutives in -cito/a. For this reason, the probabilities that these suffixes apply are summed together. The fact that the probabilities for some items do not total 100% is due to small amounts of leakage toward other suffix types.

References


