

Spanish diminutive formation without rules or constraints¹

David Eddington

Version: 11-28-2000

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 which 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 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, and the one which most discussion of the subject has focused on. For this reason, only the allomorphy of *-ito* will be considered.

Diminutive formation has been the topic of a number of investigations. For example, Jaeggli (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 Carreira (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' (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 prespecified, 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 which 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 a requires 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 dubious (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, Dijkstra, and Schreuder 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, Palmeri, Goldinger and Pisoni (1993) and Goldinger (1997) provide evidence which 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 (AML). 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, Kibler, and Albert 1991; Medin and Schaffer 1978; Riesbeck and Schank 1989; see Shanks 1995 for an overview of exemplar models; see Daelemans, Gillis, and Durieux 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 which share variables with the given context, and creates groups of database items called subcontexts. Of course, words which share more variables with the given context will appear in more subcontexts. Subcontexts are further

combined into more comprehensive groups called supracontexts. Upon inspection, some subcontexts will be homogenous, 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 heterogenous. By minimizing disagreements and eliminating members of heterogenous subcontexts, database items belonging to the most clear-cut areas of contextual space (homogenous 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 which are similar to the given context, each member is available as a potential analog. Database items which have a great deal in common with the given context will appear in many different subcontexts, and have a greater chance of affecting the behavior of the given context in comparison to those items which 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 heterogenous 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 homogenous 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 which 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 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.² 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 is the case with rule, constraint, and prototype approaches. This implies that the variables which 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 Skousen 1995: 223-226 for an example).

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 effect 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; Eddington 2000; Hintzman 1986, 1988; Hintzman and Ludlam 1980; Medin and Schaffer 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 which coincide with one's particular preconceived assumptions, and to overlook others. For example, Morin (1999) demonstrates that the criteria proposed to distinguish between Spanish words which 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 Cuetos frequency dictionary (1995; 5 million words), the LEXESP tagged frequency dictionary (Sebastián, Cuetos, and Carreiras, in preparation; 3 million words³), a corpus of spoken peninsular Spanish (Marcos Marín, no date a; 1 million words), a corpus of Argentine Spanish (Marcos Marín, no date 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.⁴ Therefore, the resulting diminutives were gleaned from a pool of 50.8 million words. Both written and spoken registers are included, although spoken sources comprise 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 to appear; Wang and Derwing 1986, 1994)

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 allomorphs were identified.⁵ However, 13 of the database items demonstrate unusual changes in the diminutives which are not found in any of the 13 groups, and which are found in three or fewer items. For example, the proper names *Antonio* and *Antonia* have diminutives with a palatalized nasal: *Antoñito*, *Antoñita*. The diminutives of *caliente*⁶ '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 which must be considered isolates since they do not fit into any of the 13 major categories described above: *fútbol* 'football' > *futbito*, *pie* 'foot' > *piececito*, *café* 'coffee' > *cafelito*, *cafetito*, *dos* 'two' > *dositos*, *José* 'Joseph' > *Joselito*, *azúcar* 'sugar' > *azucarlito*, *lejos* 'far away' > *lejecitos*, and *diagnosis* 'diagnosis' > *diagnosito*. These items were also categorized and included in the database. In addition, the base form of *valsecito* 'waltz' was included in two different categories since it was impossible to determine whether the base form was *vals* or *valse*. Once the database was completed, 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) -V ITO(S): *-ito(s)* is added to the singular base form, replacing the final vowel:
minuto 'minute' > *minutito*, *elefante* 'elephant' > *elefantito*.
- (2) -V ITA(S): *-ita(s)* is added to the singular base form, replacing the final vowel:
galleta 'cookie' > *galletita*, *Lupe* 'proper name' > *Lupita*.
- (3) -V ECITO(S): *-ecito(s)* is added to the singular base form, replacing the final vowel: *vidrio* 'glass' > *vidriecito*, *quieto* 'peaceful' > *quietecito*.
- (4) -V ECITA(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 base 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' > *Andresito*.
- (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*,
luz 'light' > *lucécita*.
- (11) - \mathfrak{N} \mathfrak{S} ⁷ITOS: *-itos* is added to the singular base form, replacing the vowel and
false plural morpheme: *lejos* 'far away' > *lejitos*, *Marcos* 'Mark' > *Marquitos*.
- (12) - \mathfrak{N} \mathfrak{S} ITAS: *-itas* is added to the singular base form, replacing the vowel and
false plural morpheme: *Lucas* > 'Luke' *Luquitas*, *garrapatas* 'tick' > *garrapatitas*.
- (13) - \mathfrak{N} 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.

++Insert Table 1 Here++

It is important to note that in some cases, diminutive formation would appear to produce sequences of [jí] in the rhyme of the penult syllable: [lím.pjo] 'clean' > *[lím.pjí.to], [ar.ma.rjo] 'closet' > *[ar.ma.rjí.to]. However, [ji] is a non-occurring rhyme in Spanish which is why the glide does not appear (Elordieta and Carreiras 1996:55; Harris 1994:182; Prieto1992:196).

Instead, the corresponding diminutives are [lím.pí.to] and [ar.ma.rí.to]. Sequences of [i+i] are also attested in one diminutive in the database (*tiíta* 'aunt'), but even this is unusual enough that the sequence results in a single high vowel in alternate diminutive forms: *tito*, 'uncle' *tita* 'aunt'.

++Insert Table 2 here++

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/⁸ are formed; 44% of the doublets have stems containing diphthongs of this sort.

Another thing which 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 Carreira 1996; Prieto 1992). Only 13 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 *-ito/a* (Category 1 and 2). The sole exception found is *retoque* 'retouch' > *retoquecito*. However *retoquito* 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 *retoquecito*, as possible variants of 12 of the 13 test words they were presented (e.g. *estuche* 'case' > *estuchecito*; *chocolate* 'chocolate' > *chocolatecito*).

Perhaps the oddest of all diminutive forms are those which appear to involve infixation of *-it-* before a word final *-or* or *-ar*: *Víctor* 'Victor' > *Victítor*, *azúcar* 'sugar' > *azuquítar*, *ámbar* 'amber' > *ambítar*. 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; Jaeggli 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 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 uninflected 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 antepenult 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 doublets (Table 2). For example, both *ratonito* 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. *ratonito*). 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 types 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 hold between *cuerno* ‘horn’ 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:

$$pueblo : pueblecito, cuerno : cuernecito :: cuervo : ?$$

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.

++Insert Figure 1 Here++

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*, *cuerno*, and *pueblo* 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 a 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 *cuervecito* 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 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 comprised 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 which 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: *carnal* fem. 'buddy' > **carnalito*, *sofá* 'couch' > **sofita*.

Other errors entailed applying a suffix such as *-ito/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).

++Insert Table 3 about here++

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 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' (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 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 orderings, 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 hampering the predictive power. For instance, Crowhurst distinguishes between word final *-e*'s 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' criticism are 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, to appear, 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, if not, the stressed/unstressed status of the penult 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, if not, the stressed/unstressed status of the antepenult syllable, 6) the phoneme(s) in the coda of the penult syllable, if any, 7) the phoneme(s) in the rhyme of the antepenult syllable, if any, 7) the phoneme(s) in the onset of the final syllable, 8) the phoneme(s) in the onset of the penult syllable.

This hierarchy coincides to a great deal 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 hierarchy 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; Jaeggli 1980;⁹ 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 which 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 *-V ito/a* categories (e.g. *comadre* 'godmother' > *comadrita*). Nevertheless, there are exceptions to this generalization. Crowhurst (1992:251) cites *sangre* 'blood' > *sangrita*, *mugre* 'filth' > *mugrita*, *leche* 'milk' > *lechita*, and *hambre* 'hunger' > *hambrita*. Harris (1994:183) cites other exceptions: *tigre* 'tiger' > *tigrito*, *chile* 'chili' > *chilito*, *nene* 'boy' > *nenito*. The database for the present study includes *Pepito* < *Pepe*, *Maitita* < *Maite*, *grandita* < *grande* 'big', and *chismito* < *chisme* 'gossip.' On the other hand, subjects in Prieto's study (1992:174) vacillated between *-V ito/a* and *-cito/a* type diminutives whose bases have three or more syllables: *chocolate* > *chocolatito/chocolatecito*, *estuche* > *estuchito/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 belong to separate syllables. 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 *-ito*, and all feminine words with the same characteristics were given diminutive forms ending in *-ita*. 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 *-ecito*, or *-ecita* 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 *-ito/a*,

and Dialect B with *-ʎecito/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 *-ecito/a* or *-ʎito/a* allomorphy in the two simulated dialects.¹⁰

++Insert Table 4 here++

Modifying the database for Dialect A entailed eliminating the majority of the words that demonstrate *-ʎecito* and *-ʎecita* allomorphy. It is 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 *-ʎecito/a*, with leakage toward *-ʎito/a*, and three favor *-ʎito/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 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 which 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 unretrievable 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 which 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. Several examples of this may be gleaned from the simulation presented in Section 6.

++Insert Table 5 here++

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 which 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.

References

- Aha, David W., Dennis Kibler, and Marc K. Albert. 1991. Instance-based learning algorithms. *Machine Learning* 6.37-66.
- Alameda, José Ramón, and Fernando Cuetos. 1995. *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Oviedo, Spain: University of Oviedo Press.
- Alegre, Maria, and Peter Gordon. 1999. Frequency effects and the representational status of regular inflections. *Journal of Memory and Language* 40.41-61.
- Ambadiang, Théophile. 1996. La formación de diminutivos en español: ¿Fonología o morfología? *Lingüística Española Actual* 18.175-211.
- _____. 1997. Las bases morfológicas de la formación de diminutivos en español. *Verba* 24.99-132.
- Baayen, Harald R., Ton Dijkstra, and Robert Schreuder. 1997. Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* 37.94-117.
- Bod, Rens. 1998. *Beyond grammar*. Stanford, CA: CSLI.
- Brown, R. and D. Mc Neill. 1966. The 'tip of the tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior* 5.325-337.
- Burzio, Luigi. 1996. Surface constraints versus underlying representations. *Current trends in phonology: Models and methods*, ed. by Jacques Durand and Bernard Laks, 97-112. Paris X and University of Salford: University of Salford Publications.
- Butterworth, B. 1983. Lexical representation. *Language production*, vol. 2, ed. by B. Butterworth, 257-294. London: Academic Press.
- Bybee, Joan. 1985. *Morphology*. Amsterdam: John Benjamins.

- _____. 1988. Morphology as lexical organization. *Theoretical approaches to morphology*, ed. by Michael Hammond, and Michael Noonan, 119-41. San Diego: Academic Press.
- _____. 1994. A view of phonology from a cognitive and functional perspective. *Cognitive Linguistics* 5.285-305.
- _____. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10.425-55.
- _____. 1998. The emergent lexicon. *Proceedings of the Chicago Linguistic Society*, vol. 34, ed. by M. Gruber, C. Derrick Higgins, K. S. Olson, and T. Wysocki, 421-435. Chicago: Chicago Linguistic Society.
- _____. Phonology and Language Use. To appear. Stanford, CA: Cambridge University Press.
- Bybee, Joan L., and Dan I. Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58.265-289.
- Chandler, Steve. 1995. Non-declarative linguistics: Some neuropsychological perspectives. *Rivista di Linguistica* 7. 233-247.
- Cole, Jennifer 1995. The cycle in phonology. *The handbook of phonological theory*, ed by John A. Goldsmith. 70-113. Cambridge, MA: Blackwell.
- Cole, Jennifer S., and José I. Hualde. 1998. The object of lexical acquisition: A UR-free model. *Proceedings of the Chicago Linguistic Society*, vol. 34, ed. by M. Gruber, C. Derrick Higgins, K. S. Olson, and T. Wysocki, 447-458. Chicago: Chicago Linguistic Society.
- Crowhurst, Megan J. 1992. Diminutives and augmentatives in Mexican Spanish: A prosodic analysis. *Phonology* 9.221-253.

- Daelemans, Walter, Steven Gillis, and Gert Durieux. 1994. Skousen's analogical modeling algorithm: A comparison with lazy learning. *Proceedings of the International Conference on New Methods in Language Processing*, ed. by D. Jones, 1-7. Manchester: UMIST.
- Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 1999. TiMBL: Tilburg memory based learner, version 2.0, reference guide. *Induction of Linguistic Knowledge Technical Report*. Tilburg, Netherlands: ILK Research Group, Tilburg University. ([Http://ilk.kub.nl/~ilk/papers/ilk9901.ps.gz](http://ilk.kub.nl/~ilk/papers/ilk9901.ps.gz)).
- Eddington, David. 1996a. Diphthongization in Spanish derivational morphology: An empirical investigation. *Hispanic Linguistics* 8.1-35.
- Eddington, David. 1996b. The psychological status of phonological analyses. *Linguistica* 36.17-37.
- _____. 2000. Spanish stress assignment within the analogical modeling of language. *Language* 76.92-109.
- _____. To appear. A comparison of two analogical models: Tilburg memory-based learner versus analogical modeling of language. To appear in an introductory text on *Analogical Modeling of Language*, ed. by Deryle Lonsdale, and Royal Skousen.
- Elordieta, Gorka, and María M. Carreiras. 1996. An optimality theoretic analysis of Spanish diminutives. *Proceedings from the main session of the Chicago Linguistic Society's 32nd meeting*, ed. by Lise M. Dobrin, Kora Singer, and Lisa McNair. Chicago: Chicago Linguistic Society.
- Goldinger, Stephen D. 1997. Words and Voices: Perception and production in an episodic lexicon. *Talker variability in speech processing*, ed. by Keith Johnson and John W.

- Mullennix, 33-65. San Diego: Academic.
- Halle, Morris. 1973. Prolegomena to a theory of word formation. *Linguistics Inquiry* 4.3-16.
- Harris, James. 1994. The OCP, prosodic morphology and Sonoran Spanish diminutives; A reply to Crowhurst. *Phonology* 11.179-190.
- Hintzman, Douglas L. 1986. Schema abstraction in a multiple-trace memory model. *Psychological Review* 93.411-428.
- _____. 1988. Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review* 95.528-551.
- Hintzman, Douglas L., and Genevieve Ludlam. 1980. Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory and Cognition* 8.378-382.
- Jackendoff, Ray. 1975. Morphological and semantic regularities in the lexicon. *Language* 51.639-671.
- Jaeggli, Osvaldo A. 1980. Contemporary studies in Romance languages: Eighth annual linguistics symposium on Romance Languages, ed by Frank Nuessel Jr., 145-158. Bloomington, IN: Indiana University Linguistics Club.
- Lamb, Sydney. 2000. Bidirectional processing in language and related cognitive systems. *Usage-based models of Language*, ed. by Michael barlow and Suzanne Kemmer, 87-119. Stanford, CA: CSLI Publications.
- Manelis, Leon and David A. Tharp. 1977. The processing of affixed words. *Memory and Cognition* 5.690-695.
- Marcos Marín, Francisco, (director). No date a. Corpus oral de referencia del español

contemporáneo. Textual corpus, Universidad Autónoma de Madrid.

[Http://elvira.llf.uam.es/docs_es/corpus/corpus.html](http://elvira.llf.uam.es/docs_es/corpus/corpus.html).

Marcos Marín, Francisco, (director). No date b. Corpus lingüístico de referencia de la lengua española en Argentina. Textual corpus, Universidad Autónoma de Madrid.

<http://www.llf.uam.es/~fmarcos/informes/corpus/coarginl.html>.

Morin, Regina. 1999. Spanish substantives: How many classes? *Advances in Hispanic Linguistics*, ed. by Javier Gutiérrez-Rexach and Fernando Martínez-Gil, 214-230. Somerville, MA: Cascadilla Press.

Medin, Douglas L. and Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85.207-238.

Nosofsky, Robert M. Exemplar-based accounts of relations between classification, recognition, and typicality. 1988. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14.700-708.

Palmeri, Thomas J., Stephen D. Goldinger, and David B. Pisoni. 1993. Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19.309-28.

Pawley, Andrew, and Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*, ed. by Jack C. Richards and Richard W. Smith, 191-225. London: Longman.

Pisoni, David. 1997. Some thoughts on 'normalization' in speech perception. *Talker variability in speech processing*, ed. by Keith Johnson and John W. Mullennix, 9-32. San Diego: Academic.

- Prieto, Pilar. 1992. Morphophonology of the Spanish diminutive formation: A case for prosodic sensitivity. *Hispanic Linguistics* 5.169-205.
- Riesbeck, Chris K., and Roger S. Schank. 1989. *Inside case-based reasoning*. Hillsdale, N.J.: Erlbaum.
- Sebastián, Núria, Fernando Cuetos, and Manuel Carreiras. In preparation. LEXESP: Creación de una base de datos informatizada de español. Report, Universitat de Barcelona.
- Sereno, Joan A., and Allard Jongman. 1997. Processing of English inflectional morphology. *Memory and Cognition* 25.425-37.
- Shanks, David R. 1995. *The psychology of associative learning*. Cambridge: Cambridge UP.
- Skousen, Royal. 1989. *Analogical modeling of language*. Dordrecht: Kluwer Academic.
- _____. 1992. *Analogy and structure*. Dordrecht: Kluwer Academic.
- _____. 1995. Analogy: A non-rule alternative to neural networks. *Rivista di Linguistica* 7.213-232.
- _____. 1998. Natural statistics in language modelling. *Journal of Quantitative Linguistics* 5.246-255.
- Stemberger, Joseph Paul. 1994. Rule-less morphology at the phonology-lexicon interface. *The reality of linguistic rules*, ed. by Susan D. Lima, Roberta L. Corrigan, and Gregory K. Iverson, 147-169. Amsterdam: Benjamins.
- Stemberger, Joseph Paul and Brian MacWhinney. 1988. Are inflected forms stored in the lexicon? *Theoretical approaches to morphology*, ed. by Michael Hammond and Michael Noonan, 101-16. San Diego: Academic Press.
- Steriade, Donca. Underspecification and markedness. *The handbook of phonological theory*, ed

- by John A. Goldsmith. 114-174. Cambridge, MA: Blackwell.
- Wang, H. S. and B. L. Derwing. 1986. More on English vowel shift: The back vowel question. *Phonology Yearbook* 3, 99-116.
- _____. 1994. Some vowel schemas in three English morphological classes: Experimental evidence. In: *In honor of William S-Y. Wang: Interdisciplinary studies on language and language change*, 561-575. Taipei: Pyramid Press.
- Wulf, Douglas. 1998. An account of German plural formation using an analogical computer model. Manuscript, University of Washington.
- Zuluaga O., Alberto. 1993. La función del diminutivo en español. *Thesaurus: Boletín del Instituto Caro y Cuervo: Muestra antológica 1945-1985*, vol. 1, ed. by Rubén Paez Patino, 305-330. Santafe de Bogotá: Instituto Caro y Cuervo.

1. This study was carried out with the help of a grant from the National Science Foundation (#00821950).
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://mdavies.for.ilstu.edu/personal/texts.htm>.
5. Some diminutives, such as *clasesitas*, and *tanquesito* were found with the suffix **-sito/a(s)*. These are obviously due to spelling errors in dialects that do not distinguish /s/ and /ʃ/, and do not indicate an additional suffix which contrasts with *-cito/a(s)*. In these cases, the spelling was regularized.
6. *Calientito* 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 other words, such as *cumpleaños*, the word ends in the plural morpheme derivationally speaking, (*cumple* + *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+dad* 'goodness'). However, the database also contains some words with non-alternating diphthongs such as *nieta* and *hueco*.
9. Jaeggli (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 than one form is ever allowed.'
10. In the case of words ending in *-e*, either *-cito/a* or *-ecito/a* may be applied to yield diminutives in *-ecito/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 towards other suffix types.

Number \ Category Y	1	2	3 ⁱ	4 ⁱ	5	6	7	8	9 ^E	10 ^E	11	12	13
Total	988	1011	35	28	201	36	100	14	6	5	5	5	13
Masculine	979	11 ^f	35	-	200	-	99	-	5	-	4	1 ^O	-
Feminine	2 ^a	990	-	28	-	36	-	14	-	5	-	3 ^R	13
No Gender	7 ^b	10 ^g	-	-	1 ^l	-	1 ^v	-	-	-	1 ^M	1 ^S	-
-o Final	910	-	35	1 ^k	2 ^m	-	-	-	-	-	-	-	-
-a Final	1 ^c	1000	-	27	2 ^P	-	1 ^x	-	-	-	-	-	13
-e Final	73	11	-	-	60	17	-	-	-	-	-	-	-
-r Final	-	-	-	-	43	5	8	1 ^z	-	1 ^K	-	-	-
-n Final	-	-	-	-	85	13	4 ^w	1 ^A	1 ^F	-	-	-	-
-l Final	-	-	-	-	1 ⁿ	-	65	8	1 ^G	-	-	-	-
-d Final	-	-	-	-	1 ^o	1 ^t	-	1 ^B	-	-	-	-	-
-s Final	-	-	-	-	1 ^T	-	13	1 ^C	2 ^H	1 ^J	5 ^N	5	-
Other Final	4 ^d	-	-	-	5 ^p	-	9 ^y	2 ^D	2 ^I	3 ^L	-	-	-
/je/ or /we/ in Root	46	29	21 ^j	20 ^s	3 ^q	1 ^u	-	-	-	-	-	-	1 ^Q
Bisyllabic -e Final	10 ^e	3 ^h	-	-	60	17	-	-	-	-	-	-	-
3+ Syllables -e Final	63	8	-	-	1 ^r	-	-	-	-	-	-	-	-

a-foto, mano; b-e.g. abajo, adelante, callando; c-papá; d-bikini, güisqui, punqui, Iñaki; e-e.g. chisme, Pepe, tigre; f-e.g. mapa,

problema, sistema; g-e.g. ahora, arriba, encima; h-grande, leche, Maite; i-all are bisyllabic; j-10 of the remaining words end in [jo],

leaving *lento, lleno, paso, sayo*; k-*mano*; l-*allá*; m-*carro, río*; n-*opel*; o-*David*; p-*buey, cují, godoy, güisqui, pupú*; q-*buey, diente, mueble*; r-*retoque*; s-j-4 of the remaining words end in [ja], leaving *hecha, lengua, mano, seca*; t-*pared*; u-*fuelle*; v-*nomás*; w-*Adrián, Juan, pompón, ratón*; x-*papá*; y-*arroz, bistec, chalet, coñac, copey, deslíz, lápiz, maíz, reloj*; z-*Pilar*; A-*inversión*; B-*verdad*; C-*Gladys*; D-*cruz, nariz*; E-all monosyllabic; F-*tren*; G-*sol*; H-*flux, valse*; I-*pez, rey*; J-*tos*; K-*flor*; L-*cruz, luz, voz*; M-*apenas*; N-*anteojos, Carlos, cumpleaños, lejos, Marcos*; O-*Lucas*; P-*allá, papá*; Q-*huevo*; R-*gafas, garrapatas, Mercedes*; S-*apenas*; T-*vals*.

Table 1. Summary of Database Items by Category.

Base Form	Diminutive A	Diminutive B	Gloss
<i>altar</i>	<i>altarcito</i>	<i>altarito</i>	altar
<i>Antonia</i>	<i>Antonita</i>	<i>Antoñita</i>	Antonia
<i>Antonio</i>	<i>Antonito</i>	<i>Antoñito</i>	Antonio
<i>bote</i>	<i>botecito</i>	<i>botito</i>	jar
<i>café</i>	<i>cafecito, cafetito</i>	<i>cafelito</i>	coffee
<i>caliente</i>	<i>calientito</i>	<i>calentito</i>	hot
<i>carne</i>	<i>carnecita</i>	<i>carnita</i>	meat
<i>carro</i>	<i>carrocito</i>	<i>carrito</i>	car
<i>chófer</i>	<i>chofercito</i>	<i>choferito</i>	chauffeur
<i>cruz</i>	<i>crucecita</i>	<i>crucita</i>	cross
<i>cuello</i>	<i>cuellecito</i>	<i>cuellito</i>	neck
<i>cuenta</i>	<i>cuentecita</i>	<i>cuentita</i>	bill
<i>cuento</i>	<i>cuentecito</i>	<i>cuentito</i>	story
<i>cuerno</i>	<i>cuernecito</i>	<i>cuernito</i>	horn
<i>cuervo</i>	<i>cuerpecito</i>	<i>cuerpito</i>	body
<i>grande</i>	<i>grandecita</i>	<i>grandita</i>	large
<i>güisqui</i>	<i>güisquicito</i>	<i>güisquito</i>	whisky
<i>hierba</i>	<i>hierbecita</i>	<i>hierbita</i>	grass
<i>hierro</i>	<i>hierrecito</i>	<i>hierrito</i>	iron
<i>hombre</i>	<i>hombrecito</i>	<i>hombrito</i>	man
<i>huevo</i>	<i>huevecito</i>	<i>huevo</i>	egg

<i>indio/a</i>	<i>indiecito/a</i>	<i>indito/a</i>	Indian
<i>Jorge</i>	<i>Jorgecito</i>	<i>Jorgito</i>	George
<i>José</i>	<i>Josecito</i>	<i>Joselito</i>	Joseph
<i>Juan</i>	<i>Juancito</i>	<i>Juanito</i>	John
<i>juego</i>	<i>juegucito</i>	<i>jueguito</i>	game
<i>lento</i>	<i>lentecito</i>	<i>lentito</i>	slow
<i>lleno</i>	<i>llenecito</i>	<i>llenito</i>	full
<i>mano</i>	<i>manecita</i>	<i>manita, manito</i>	hand
<i>muerto</i>	<i>muertecito</i>	<i>muertito</i>	dead
<i>nieta</i>	<i>nietecita</i>	<i>nietita</i>	granddaughter
<i>papá</i>	<i>papacito</i>	<i>papito, papaíto</i>	daddy
<i>paso</i>	<i>pasecito</i>	<i>pasito</i>	step
<i>pedra</i>	<i>pedrecita</i>	<i>pedrita</i>	stone
<i>pieza</i>	<i>piececita</i>	<i>piecita</i>	piece
<i>pueblo</i>	<i>pueblecito</i>	<i>pueblito</i>	town
<i>puerta</i>	<i>puertecita</i>	<i>puertita</i>	door
<i>quieto/a</i>	<i>quietecito/a</i>	<i>quietito/a</i>	calm
<i>ratona</i>	<i>ratoncita</i>	<i>ratonita</i>	mouse
<i>ratón</i>	<i>ratoncito</i>	<i>ratonito</i>	mouse
<i>río</i>	<i>riecito</i>	<i>riocito</i>	river
<i>rubio</i>	<i>rubiecito</i>	<i>rubito</i>	blond
<i>sueño</i>	<i>sueñecito</i>	<i>sueñito</i>	sleep

<i>tambor</i>	<i>tamborcito</i>	<i>tamborito</i>	drum
<i>tiempo</i>	<i>tiempecito</i>	<i>tiempito</i>	time
<i>tren</i>	<i>trenecito</i>	<i>trencito</i>	train
<i>viejo/a</i>	<i>viejecito/a</i>	<i>viejito/a</i>	old
<i>viento</i>	<i>vientecito</i>	<i>vientito</i>	wind
<i>vuelo</i>	<i>vuelecito</i>	<i>vuelito</i>	flight
<i>vuelta</i>	<i>vueltecita</i>	<i>vuelcita</i>	revolution

Table 2. Doublets in the Database.

Word	Gloss
<i>airito</i>	air
<i>alfarcito</i>	pottery
<i>Adriancito</i>	Adrian
<i>barrigonita</i>	big-bellied
<i>bebecito</i>	baby
<i>buchecito</i>	crop of birds
<i>buenito</i>	good
<i>bueyecito</i>	ox
<i>callita</i>	street
<i>chaletcito</i>	chalet
<i>chilito</i>	chili pepper
<i>cuatecito</i>	buddy
<i>cuervecito</i>	crow
<i>cuestita</i>	incline
<i>Davidito</i>	David
<i>dosito</i>	two
<i>dulcito/a</i>	sweet
<i>fuentita</i>	fountain
<i>fuercita</i>	strength
<i>hambrita</i>	hunger
<i>hechita</i>	done

<i>huequito</i>	hollow
<i>huellita</i>	track
<i>jamonita</i>	chunky woman
<i>juerguita</i>	binge
<i>lenguita</i>	tongue
<i>llavita</i>	key
<i>mierdita</i>	shit
<i>muellito</i>	dock
<i>nenecito</i>	child
<i>nietito</i>	grandson
<i>nubita</i>	cloud
<i>nuevito/a</i>	new
<i>patronita</i>	boss
<i>patronita</i>	owner
<i>Pepecito</i>	<i>Pepe</i>
<i>piecito</i>	foot
<i>pomponcito</i>	pompom
<i>prietito</i>	tight
<i>puestito</i>	stand
<i>retoquito</i>	touch-up
<i>reycito</i>	king
<i>ruedita</i>	wheel

<i>señorcito</i>	sir
<i>sequita</i>	dry
<i>suavita</i>	smooth
<i>tiendita</i>	store
<i>valsito</i>	waltz
<i>verdito</i>	green
<i>viajito</i>	trip

Table 3. Erroneous diminutives predicted by AML which are attested on the WWW.

BASE WORD	DIALECT A		DIALECT B	
	Prob. of - <i>ŷito</i>	Prob. of - <i>ŷecito</i>	Prob. of - <i>ŷito</i>	Prob. of - <i>ŷecito</i>
<i>riego</i>	100	0	26.9	72.7
<i>ruego</i>	100	0	22.3	77.8
<i>siervo</i>	99.98	0	48.66	50.17
<i>trueno</i>	99.99	0	0	100
	Prob. of - <i>ŷita</i>	Prob. of - <i>ŷecita</i>	Prob. of - <i>ŷita</i>	Prob. of - <i>ŷecita</i>
<i>cuelga</i>	100.00	0	23.12	76.88
<i>cuerda</i>	99.98	0	53.71	45.19
<i>fiebre</i>	95.48	3.74	29.34	60.12
<i>friega</i>	99.95	0.05	19.58	80.34
<i>niebla</i>	100	0	60.00	40.00
<i>nieve</i>	87.76	11.39	13.21	79.56
<i>prueba</i>	100	0	59.15	40.85
<i>suerte</i>	72.28	7.92	16.97	67.68
<i>sierva</i>	99.97	0	30.85	67.73

Table 4. Probabilities of Variant Forms in Two Simulated Dialects.

Base Form	Variant A	Prob. of Variant A	Variant B	Prob. of Variant B
<i>yegua</i>	<i>yeguita</i>	67.22	<i>yeguecita</i>	32.78
<i>Jorge</i>	<i>Jorgito</i>	23.33	<i>Jorgecito</i>	74.95
<i>pierna</i>	<i>piernita</i>	40.81	<i>piernecita</i>	59.19
<i>cuervo</i>	<i>cuervito</i>	33.38	<i>cuervercito</i>	66.31
<i>David</i>	<i>Davidito</i>	58.06	<i>Davidcito</i>	36.10
<i>Chevolé</i>	<i>Chevolito</i>	54.79	<i>Chevolecito</i>	40.43
<i>nieta</i>	<i>nietita</i>	67.68	<i>nietecita</i>	31.21
<i>corte</i>	<i>cortito</i>	45.03	<i>cortecito</i>	51.53

Table 5. Examples of Competing Gangs on Selected Base Forms.

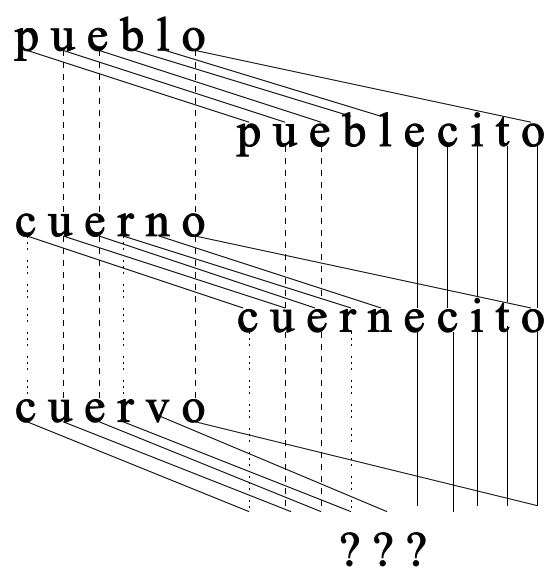


Figure1. Network Representation of How Analogy May Work.