

Spanish Gender Assignment in an Analogical Framework

David Eddington

RUNNING HEAD: Gender Assignment

KEY WORDS: Spanish, gender, analogy, exemplar model

This study was carried out with the help of a grant from the National Science Foundation (#00821950). Any correspondence may be directed to the author at Department of Spanish and Portuguese, University of New Mexico, Albuquerque, NM 87131-1146.

Abstract

The focus of the study is two-fold: 1) to determine how much phonemic material must be considered in predicting nominal gender in Spanish; 2) to determine if gender assignment can be considered an analogical process. A database of Spanish nouns extracted from a frequency dictionary serves as the database for the present study, while the phonemic make up and syllable structure of the penultimate rime and final syllable serve as the variables. A number of analogical simulations using the Analogical Modeling of Language algorithm demonstrate that phonemic material from as far back as the penultimate rime may aid in gender assignment. Analogy is also shown to closely mirror Spanish speakers' intuitions regarding the gender of unknown words. Gender assignment errors made by children also fall out of the analogical architecture.

1. INTRODUCTION Previous studies of Spanish gender fall into one of three major categories: 1) In the pedagogical approach (e.g. Bergen, 1978; Bull 1965; Teschner & Russell, 1984), the emphasis is placed on dividing Spanish nouns into masculine and feminine groups based mainly on their final phonemes, and listing the most common exceptions to those groups. This categorization is made to facilitate the acquisition of Spanish as a second language. 2) In the descriptive approach (Teschner, 1983; Rosenblat, 1952), on the other hand, the interest is in finding systematic correspondences between nominal gender and phonological patterns in Spanish words. 3) Generative analyses (e.g. Harris, 1985, 1991; Klein, 1983, 1989) strive to describe gender in terms of a rule system that derives a word's final phoneme(s) given the word's inherent gender and a set of abstract assumptions about the word's underlying structure.

Although each of these analyses may be valid in its own realm of inquiry, none of them have as a goal to determine how native speakers may go about assigning gender when syntactic clues such as gender-marked determiners and adjectives are absent. Given the mentalistic vocabulary often employed in the generative literature (i.e. *processes, derive, language acquisition, production, etc.*), one could gain the impression that such analyses do explain how native speakers assign gender. (This contrasts with the clearly descriptive goals of the pedagogical and descriptive approaches.) Generative accounts may be elegant representations of linguistic structure, however, their relationship to the psychological mechanisms that play a role in actual language usage

is dubious (Chandler, 1993; Derwing, 1973; Eddington, 1996; Lamb, 2000; Skousen, 1975). In addition, some researchers have clearly stated that generative accounts are simply not intended to be taken as models of linguistic performance (c.f. Kiparsky, 1975, p. 198; Chomsky & Halle 1968, p.117; Bradley. 1980, p. 38)

Even if a generative model of gender assignment did reflect actual language processing, an additional limitation would be that the rules are designed to assign word-final phonemes to words whose gender is previously known. That is, they are not designed to apply in reverse and assign gender based on the phonemic make-up of an unknown word. Consider the feminine word *cama*, and the masculine word *drama*. In spite of their gender differences, they are both *-a* final. According to Harris (1991), this is explained by the fact that they have a different underlying structure. However, this underlying structure is not apparent to a speaker who needs to determine their gender, and only has the surface structure to work with.

The relationship between the gender of a noun and its phonological shape is by no means straightforward. The lack of one-to-one correspondence led Harris to conclude that in spite of the many generalizations that exist, "the correlation between word marker and grammatical gender is random and arbitrary" (1985, p. 37). Since most words ending in *-o* are masculine, and most ending in *-a* and *-d* are feminine, it appears that only the final phoneme comes into play in assigning gender. However, with *z*-final words one would have to consider at least the last two phonemes; most words ending in *-az* and *-uz* are masculine, while those ending in *-ez* are generally

feminine (Teschner, 1983). Words ending in *-iz* would require an analysis of the entire final syllable in order to make a gender determination; those ending in *-briz*, *-driz*, and *-triz* are feminine, while most others ending in other consonants followed by *-iz* are masculine. Words ending in *-ón* and *-is* cannot be correctly assigned gender without considering the third-to-the last phoneme. Those ending in *-ión* are mostly feminine, while those in another consonant plus *-ón* are generally masculine. Similarly, *-sis* and *-tis* words are mainly feminine, while those ending in another consonant plus *-is* are masculine. At times, it appears that elements belonging to the penultimate syllable are also germane to gender assignment. For example, about 89% of *-e* final nouns are masculine (Teschner & Russell, 1984). However, there are a substantial number of high-frequency, bisyllabic, *e*-final nouns whose penultimate nucleus is /a/, and whose gender is feminine.

It should be evident from this brief sketch, that it is unclear which factors are relevant to gender assignment. This is problematic if one's goal is to devise a rule-based global characterization of gender assignment. In such an account, the same factors must be relevant for all items. In other words, it would not be possible to view only the final phoneme to be relevant for words ending in *-o* and *-a*, but to have to consider more of a word's phonemic material for words ending in *-z*. Nevertheless, the correspondences between a word's grammatical gender and its phonological shape is far from fortuitous as Harris suggests. Instead, definite tendencies exist, and may play a part in native speakers' intuitions about gender assignment. Therefore, finding the best way to model

gender/phoneme correlations is of theoretical interest, as is determining if the resulting model reflects how Spanish speakers process grammatical gender.

The present study is not intended as a rehash of the correspondences between gender and phonology, which have been adequately dealt with elsewhere (Bergen, 1978; Rosenblat, 1952; Teschner, 1983; Teschner & Russell, 1984). Instead, it explores this question: How much phonological information must be taken into consideration in order to make a gender determination for a noun? An analogical simulation was put to the task on a database comprised of the 2416 most frequent nouns in Spanish. As will be seen, the present study suggests that analogy is quite proficient at the task of gender assignment, and that better results are obtained when more than the final phoneme are considered. Data from a gender assignment experiment, and errors made by children are also presented as evidence in favor of the analogical account.

2.0. DATABASE AND VARIABLES. The database for the present study included all of the single gender nouns in the Juilland & Chang-Rodríguez frequency dictionary of Spanish (1964).¹ These nouns represent the most frequent nouns in the language. This is important since psycholinguistic experiments have shown that frequent words are more easily retrieved from memory and are retrieved with fewer errors (e.g. Allen et al., 1992; MacKay, 1982; Scarborough et al., 1977). Therefore, they are assumed to be more readily available to exert their influence on language usage.

2.1. Method. The singular form of all 2416 nouns was encoded in terms of eight variables,² and included a specification of gender for each item. Eight variables are sufficient to encode the phonemic make-up and syllabic structure of the penultimate rime and final syllable of each word. This choice of variables is consistent with the consensus expressed in the literature that the most relevant part of a noun, as far as gender assignment is concerned, appears at the end of the word. In those instances in which both the singular and the plural form of the same noun appeared in the database, only the singular was included.

The nouns were encoded in this manner (see Table 1): Variable groups 1-5 incorporate the final syllable. Variable groups 6-8 incorporate the nucleus and rhyme of the penultimate syllable. The algorithm for deriving the variables centers on the syllable nuclei. Starting from the nucleus and working outward, the next tautosyllabic phoneme is assigned to the next variable. If there is none, the next variable is given the syllable boundary symbol '0'. Any other missing segments are marked with the null symbol '-'. This algorithm was altered slightly so that the first variable contains the word final phoneme.

Examples:	Gender:	Variable Groups:
		87654321
tiempo	M	em0-0p-o
ojo	M	o0--0x-o
origen	M	i0--0xen

consecuencia F en000i-a

3. RULELESS MODELS. In recent years, a number of models have emerged that exploit statistical patterns in languages, without being statistical analyses per se (Bybee, 1985, 1988, 1998, 1994, 2001; Daelemans, Zavrel, van der Sloot, & van den Bosch 1999; Stemberger, 1994). The best known model of this sort is, of course, a family of related models known as connectionism (e.g. McClelland 1988). The major contention of these models is that linguistic knowledge need not be rooted in rules, but that it emerges through language use.

For example, a number of researchers have studied phonotactic constraints and co-occurrence of phonemes (Frisch, 2000; Frisch, Large, & Pisoni, 2000; Frisch, Large, Zawaydeh, & Pisoni, 2001; Frisch & Zawaydeh, 2001; Pierrehumbert, 1994; Treiman, Kessler, Knewasser, Tincoff, & Bowman, 2000). Using methods such as wordlikeness ratings of nonce words and memory recognition tasks, they have found that this sort of linguistic knowledge is gradient and probabilistic, the probabilities being closely related to the frequency of occurrence in a dictionary. Such findings are incompatible with rule-based accounts of phonotactics.

Of course, ruleless models differ in regards to how they conceive linguistic information to be stored in the mind. Connectionism suggests a more abstract representation in which knowledge is stored as patterns of association of varying strengths that exist in a network of interconnected nodes. The exemplar model of

Daelemans et al., (1999) holds that tokens of past experience are stored in the mind without any sort of abstract generalizations that summarize their similarities. Such information is subsequently used in language processing. Exemplar models are corroborated by evidence that detailed information about individual word tokens is stored in memory (Brown & McNeill, 1966; Bybee, 1994; Pisoni, 1997; Palmeri, Goldinger, & Pisoni, 1993; Goldinger, 1997) .

Frisch et al. (2001) find evidence for exemplar influence in their study on phonotactics, but feel that some of the probabilistic knowledge is too fine-grained to admit an exemplar-based explanation. Therefore, they conclude that some knowledge of phonotactics must be abstracted away from individual exemplars. However, Frisch et al.'s need to resort to abstraction may be explained by their particular conception of what exemplar-based influence entails. It appears that they envision a nearest neighbor model in which only one, or a small handful of exemplars, are available to exert analogical influence on a given word. In an exemplar model, such as Analogical Modeling of Language, (which is described below), many, even hundreds of exemplars, may affect phonotactic judgements, each to a differing degree depending on how similar they are to the word in question. It is in this way that an exemplar model may account for influences which appear to be abstractions, because they are spread out across the lexicon and not uniquely associated with a few specific exemplars.

4.0. ANALOGICAL MODELING OF LANGUAGE. The remainder of this paper will

be dedicated to describing Analogical Modeling of Language (AM) and utilizing it to predict gender, and more importantly, to determine how much phonological information is needed in order to assign gender. AM is an exemplar-based 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, & Albert, 1991; Medin & Schaffer, 1978; Pierrehumbert, 2001; Riesbeck & Schank, 1989; see Daelemans, Gillis, & Durieux, 1994 for a comparison of AM and Aha et al.). AM is founded on the premise that all known words, whether they are morphologically simple or complex, are stored as wholes in the mental lexicon.³ When the need arises to determine the behavior of a word, the lexicon itself is accessed. A search is conducted for the words most similar to the unknown word.⁴ The behavior of the word(s) most similar to the word in question generally predicts the behavior of the word in question, although the behavior of less similar words has a small chance of applying as well.

The probability that a word will be chosen as an analog is dependent on three derived properties (Skousen, 1995, p. 217):

- (1) *proximity*: the more similar the example is to the word in question, the greater the chances of that example being selected as the analogical model;
- (2) *gang effect*: if the example is surrounded by several other examples having the same behavior, then the probability of selecting these similarly behaving examples is substantially increased;

(3) *heterogeneity*: an example cannot be selected as the analogical model if there are intervening examples, with different behavior, closer to the word in question.

These derived properties are important since they constrain what examples can constitute analogs, as well as deciding between competing analogs.

AM bears some similarity to connectionism in that neither model computes characterizations of the data in the form of rules. Nevertheless, there are significant differences between AM and connectionist models (Chandler, 1995; Jones, 1996; Skousen, 1989, 1995). Connectionism can only predict one outcome for an input, while AM can be used to predict the probability that one or more outcomes will be chosen. Connectionist networks also require extensive training, while AM does not. In connectionism, information is stored as patterns of activation in a network of interconnected nodes; there is no representation of individual words. In AM, the information is contained in a database of exemplars representing the contents of the mental lexicon. This database may be added to at any time. In contrast, connectionist networks cannot readily accept new data without having to be completely retrained to include the new data.

One important difference between AM and rule models is that in the latter, acquisition is a matter of making a global determination as to what variables are most crucial to the task at hand. As already discussed above, this is somewhat problematic for gender assignment, since for many nouns, (e.g. those ending in *-a*, *-o*), gender can be determined by reference to the final phoneme alone. On the other hand, correct

assignment for other nouns, such as those ending in *-n* and *-z*, require more phonological information.

In an AM analysis, in contrast, it is not necessary to determine beforehand exactly which variables are most important. It is actually desirable to include many variables which may seem irrelevant at the outset. Consider Skousen's simulation with a group of Finnish past tense forms (1989). For most of these verbs, the choice of the past tense morpheme appears to be dependent on what the final two phonemes of the stem are, or if the vowel of the verb stem is *a*. However, *sorta-* 'to oppress' appears to be an exceptional case. It does not become *sorsi* as a rule-based analysis would predict. Instead, it becomes *sorti*. Nevertheless, AM correctly predicts this outcome, but the prediction is made on the basis of the *o* in the stem which *sorta-* has in common with a group of other verbal stems, each of which has a past tense form ending in *-ti*. A stem-internal *o* may be an irrelevant variable for the vast majority of these verbs, but not for *sorta-*. This would not have become evident if only the variables which appeared most relevant were included in the analysis.

A similar phenomenon was found in a connectionist simulation designed to predict the German definite article (MacWhinney, Leinbach, Taraban, & McDonald, 1989). In two simulations, the variables encoded the presence or absence of 38 carefully chosen pieces of morphological, semantic, and phonological information about the word (e.g. whether the word contains a specific morpheme, or a phoneme in a certain position). Each of these cues is thought to govern definite article usage. In another

simulation, the only variables were the strings of phonemes that comprise the word that the article agrees with. That is, no effort was made to include only those elements thought relevant to the task and separate them from those thought to be irrelevant. Nevertheless, this simulation yielded better results than the previous ones that carefully eliminated cues that were considered unimportant to the task of definite article assignment. This sort of evidence suggests that speakers do not make a global determination of which variables are relevant in advance, as rules imply. Instead, all variables may potentially take part. In AM, the most important variables can only be determined only after the analogical set is constructed and inspected. In short, what is crucial for correct categorization of one word may not be for another.

4.1. The Psychological Validity of AM. While most rule models claim to be relevant to linguistic competence and linguistic structure, they do not claim to mirror the psychological mechanisms responsible for language comprehension and production. AM, on the other hand, is presented as a psychologically plausible model (Chandler, 1995, 2002). In one sense, a model may be considered psychologically valid to the extent that it correctly accounts for linguistic data. AM has proven successful at modeling several different language phenomena: linking elements in Dutch noun compounds (Krott, 2002), phonological alternations in Turkish stems (Rytting, 2000), Dutch stress assignment (Daelemans et al., 1994)

However, a more crucial test of the psychological import of a model is that it

makes empirical predictions about phenomena such as actual language usage, the formation of neologisms, language acquisition data, slips of the tongue, historical developments, etc. Evidence of this type is beginning to appear. For example, an AM simulation of Spanish stress assignment (Eddington, 2000b) not only accounts for the phenomenon better than an empirically testable rule approach, but make the same sort of errors as children do. In addition, it helps explain the different patterns of errors made by children of different ages.

Evidence from AM has also come to bear on the debate between those who assume that a single system can handle both regular and irregular inflections (e.g. Daugherty & Seidenberg, 1994), and those who espouse separate mechanisms for regular and irregular inflection (e.g. Pinker & Prince, 1994). For instance, in a study of English past tense formation (Prasada & Pinker, 1993), subjects were asked to inflect nonce words. The results of this experiment were originally interpreted as support of the dual-route model. However, an AM simulation of the same nonce items (Eddington, 2000a) mirrored the subject's responses closely, demonstrating that a single-route model is equally capable of accounting for the test subjects' intuitions. In a similar study, Say & Clahsen (2001) asked Italian speakers to determine what past participle, (and hence conjugation), they would assign to nonce verbs. Say and Clahsen claim that their findings favor a dual-route model for Italian verbal stem formation. Notwithstanding, the nonce words were assigned conjugational class by AM, which again closely mirrored the responses of the test subjects (Eddington, 2001).

4.2. The AM Algorithm⁵

To this point, AM has been introduced and its ability to model linguistic processes discussed. However, its actual internal workings have yet to be elucidated. Perhaps the best way to understand the AM algorithm is with a concrete illustration. Predictions are always made in terms of specific exemplars, therefore, for the purposes of the example, the following seven monosyllabic nouns and their corresponding gender will be considered: *paz F*, *plan M*, *tren M*, *cal F*, *ron M*, *par M*, and *rey M*. The task will be to predict the gender of *pan* on the basis of these seven items as the database. The three variable groups to be considered are the phoneme or phoneme cluster of the onset, nucleus, and rhyme. To begin with, all possible database items are assigned to a series of subcontexts which are defined in terms of the given context *pan*. For the task at hand, there are eight subcontexts (Table 2). A bar over a variable indicates that any value of the variable except the barred value is permitted in the subcontext.

++Insert Table 2 here++

By assigning members to subcontexts it is possible to determine disagreements. Disagreements are marked with asterisks in Table 2. A disagreement occurs when words that are equally similar to the given context, exhibit different behaviors, in this case, different genders. The number of disagreements is determined by pairing all members of a subcontext with every other member, including itself, by means of

unidirectional pointers, and counting the number of times the members of the pair have different behaviors. In this example, the only subcontext containing any disagreement is $pa\hat{u}$

Once the disagreements have been found, the subcontexts are arranged into more comprehensive groups of subcontexts called supracontexts as shown in Table 3. In Table 3, A hyphen indicates a wildcard.

++Insert Table 3 here++

The subcontextual analysis consists of adding all of the subcontextual disagreements which appear in each supracontext. The next step is to perform a supracontextual analysis which consists of analyzing all of the words that appear in a given supracontext, and determining disagreements (Table 4).

++Insert Table 4 here++

In the supracontextual analysis, it can be seen that words that have more than one feature in common with *pan* appear in more than one supracontext. This is AM's way of allowing the gender of words which are more similar to *pan* to influence the gender assignment of *pan* to a greater extent.

The purpose of AM's algorithm is to determine which members of the database are most likely to affect the gender assignment of *pan*, and also to calculate the extent of analogical influence exerted. Much of this is accomplished by calculating heterogeneity. Heterogeneity is determined by comparing the number of disagreements in the supracontextual and subcontextual analyses. If there are more

disagreements in the supracontextual analysis, the supracontext is heterogenous, and its members are eliminated from consideration as possible analogs. If the number of disagreements does not increase, the supracontext is homogenous. Words belonging to homogenous supracontexts comprise the analogical set. In the example under consideration, disagreements increase in the supracontexts - - -, and - a -. Therefore, their members are eliminated from consideration. *Cal*, and *rey* appear exclusively in these heterogenous supracontexts. As a result, they do not form part of the analogical set. *Plan* is also a member of both - - -, and - a -, however, it is also a member of the homogenous supracontexts - a n, and - - n, so it will still be available to influence *pan*.

It should not be surprising that *rey* would be eliminated through heterogeneity; it has no phonemes in common with *pan*. However, consider the words *ron* and *cal*. Both share only one feature with *pan*, yet heterogeneity eliminates only *cal*, and not *ron*. This is due to the fact that *ron* appears in the supracontext - - n, and all of the members of that supracontext are masculine, therefore, there is no disagreement. *Cal* F, on the other hand, competes with masculine words in all of the supracontexts in which it appears.

As already mentioned, the analogical set (Table 5) includes words from all non-empty homogenous supracontexts.

++Insert Table 5 here++

The analogical set contains all of the database items that can possibly influence the gender assignment of *pan*. There are two methods for calculating the influence of the

analogical set on the behavior of the given context. One is to assign the most frequently occurring behavior to the given context (selection by plurality). Of the 18 pointers in the set, 14 point to masculine. Therefore, selection by plurality would assign masculine gender to *pan*. The other method (random selection) involves randomly selecting a pointer, and assigning the behavior of the word indicated by the pointer to the given context. In this case, the probability of masculine gender assignment would be 77.78% (14/18). Random selection allows the fuzziness between behaviors to be calculated, something which is not possible in either a rule-model or a connectionist simulation.

5.0. ANALOGICAL ASSIGNMENT OF GENDER. In the current simulation, AM was put to the task of assigning gender. Four different experimental conditions were established according to how much of the noun's phonemic and syllable structure were included: 1) the word's final phoneme, 2) the rhyme of the word's final syllable, 3) the word's final syllable, 4) the word's penultimate rhyme and final syllable. In each condition, the 2416 words in the database were removed one at a time. Each word's gender was then determined on the basis of the similarities it bears to other words in the database, according to AM's algorithm. In essence, this procedure treats each word as if it were new and unknown. If the word were left in the database, AM would obviously find it when constructing the analogical set, along with its specified gender, and would predict the correct gender. This state of affairs is tantamount to accessing a known lexical item from memory along with its known characteristics, which is of little

theoretical interest for the task at hand. Table 6 contains some sample outcomes calculated by AM when the elements of the penultimate rime and final syllable are used as variables. It indicates that *planeta* and *catedral* are among those words incorrectly assigned gender by AM.

++Insert Table 6 here++

In order to determine how successful the simulation was, it was necessary to establish some sort of benchmark. Bull 's (1965) pedagogical rules, which are based on a dictionary search, seemed adequate. Accordingly, words ending in *-a*, *-d*, *-ción*, *-sión*, *-tis*, and *-sis* are feminine, while words ending in any other phoneme(s) are masculine. The gender of words ending in *-z* are thought not to lend themselves to any sort of easy categorization, and need to be learned on a case-by-case basis. However, in the database, 21 of the 28 *-z* words are feminine, so for the purposes of the study, Bull's rules will be extended so that *-z* final words are considered feminine. By applying these rules to the 2416 items in the database a benchmark success rate of 95.0% may be established.

Given the fact that token frequency is often a factor in linguistic experiments, two databases were constructed. The first, the type frequency database, contained only one set of variables for each item. The second, the token frequency database, was adjusted to reflect the token frequency of the nouns as it appears in Juilland and Chang-Rodríguez (1964). For every five words per million,⁶ one set of variables for each item

was included in the token frequency database. Table 7 specifies the success rate in assigning gender in each of the experimental conditions using both databases.

++Insert Table 7 here++

Analogy proves to be quite proficient at correctly assigning gender, even when information besides the word's final phoneme is considered. It appears that the phonemes in the penultimate rime and final syllable are the best predictors of a noun's gender, especially since the success rates there fall slightly above the benchmark of 95.0%. However, there is no statistical difference between the number of errors across the four conditions and two databases ($\chi^2(3) = 5.59, p < 0.1$). In other words, each set of variables carries out the task of predicting gender equally well, and none departs far from the benchmark.

Nevertheless, it is important to remember that the goal of AM is not only to successfully account for the items in a given database, but more importantly, to make predictions that are consistent with empirical data. This information is often evidenced in the errors AM makes. For example, it is widely known that the great majority of nouns ending in *-a* are feminine, while those that end in *-o* are masculine. Without making use of any sort of generalization to that effect, AM correctly assigns gender to nearly all such words. However, it also overgeneralizes and incorrectly assigns gender to words that are commonly considered exceptional; *mano* 'hand' is assigned masculine, and *planeta* 'planet' and many other masculine words ending in *-a* are

assigned feminine gender. While this may seem trivial, it is important in that these are precisely the words that are misassigned gender by AM.

These findings are also corroborated by other data. For example, Brisk (1976) notes that the majority of gender errors made by Spanish speaking children occur on words ending in phonemes other than /o/ and /a/. This is reminiscent of the success rates made by the most successful AM simulation (see Table 7). 99.0% of words ending in *-a* and *-o* are correctly assigned gender, while the success rate for words ending in other phonemes is 91.9%.

In a similar vein, García (1998) found that the use of nonstandard gender in Texas Spanish occurred mainly with the word *escuela* 'school', words ending in *-e*, and masculine words of Greek origin ending in *-a*. AM assigns the standard feminine gender to *escuela*. However, as far as *-e* final words is concerned, only 84% of them are correctly assigned gender, which falls far short of the overall success rate of 96%. The database contains 18 masculine words of Greek origin ending in *-a*, and only six (33%) of these are correctly assigned masculine gender. These six masculine items are able to serve as analogs for each other in spite of the strong influence exerted by over 600 *-a* final feminine words, because all six end in *-ema*.

Another general tendency noticed by Brisk (1976) was for children to make somewhat more gender errors on feminine words, than on masculine (17.7% errors on fem., and 11.5% errors on masc.). However, children whose abilities in Spanish were least developed made more errors entailing misassigning masculine words feminine

gender than did more advanced speakers. This sort of developmental phenomenon can be modeled by assuming that speakers with more advanced abilities have larger vocabularies. Therefore, a series of AM simulations was performed with databases of varying sizes. The purpose of the simulations was to calculate the number of errors that would occur (i.e. masculine nouns misassigned feminine gender, and feminine nouns misassigned masculine).

Ten simulations were performed as follows. First, the database was ordered in descending order of word frequency, and divided roughly equally⁷ into ten data sets. The first data set contained the 241 most frequent nouns. The second data set included all of the items in the first one, plus the next 241 most frequent words, and so on until the tenth data set comprised all 2416 items. This progression of data sets not only corresponds to the fact that language acquisition entails increasing the size of one's mental lexicon, but that more frequent words are learned first, and less frequent words at a later stage. The entire 2416 items were used as the test set in each of the ten runs. Comparing the number of errors on masculine and feminine nouns in the test set is a valid procedure since there are almost identical numbers of masculine and feminine items (masc. N=1207; fem. N= 1209).

++Insert Figure 1 here++

The outcome of the ten simulations is summarized in Figure 1. Errors on

masculine nouns outnumber those committed on feminine nouns in the simulations that used smaller data sets composed of high frequency words. However, when all 2416 words comprise the data set, the number of errors on feminine nouns becomes slightly higher. This mirrors quite closely the acquisitional data presented by Brisk. According to her, the split that is observed for more advanced speakers occurs because masculine nouns outnumber feminine nouns in children's vocabularies. While this may be true in her study, it does not explain the outcome of the simulations. In the simulations in which errors on feminine nouns outnumber errors on masculine nouns (2169, and 2416 items) the percentages of masculine and feminine data set items do not vary widely (2169: 51% masc., 49% fem.; 2416: 50% masc., 50% fem.) Although the frequency data from Juilland and Chang-Rodríguez (1964), on which the present study is based, indicate that there are about equal numbers of masculine and feminine nouns, their data are derived from written sources. A word count based on children's speech (Rodríguez Bou, 1966), on the other hand, demonstrates that masculine nouns dominate the most frequent nouns, which supports Brisk's assertion.

The notion of frequency and markedness are intimately entwined because it is often the case that the unmarked entity is also the more frequent. This may be true in the vocabularies of young Spanish speakers. However, there are reasons besides dominant frequency for considering the masculine the unmarked or default gender (see Prado, 1982 for an extensive list). For instance, both children and adults favor the masculine when asked to assign gender to unknown nouns (Natalicio, 1983; Perez-

Pereira, 1991).

The structure of the nouns themselves may also be responsible for establishing the masculine as the unmarked member. Plunkett and Marchman (1991) observe that unmarked or default status does not necessarily depend on numerical superiority. Instead, items belonging to the marked category tend to cluster in groups sharing many characteristics. Unmarked items, on the other hand, have less in common, and tend to be spread out across contextual space. Bull's rules for gender assignment reflect this state of affairs; words ending in *-a*, *-d*, *-ción*, *-sión*, *-tis*, and *-sis* are feminine (marked), while words ending in any other phoneme(s) are masculine (unmarked). From an analogical perspective, what this means for gender is that a random throw of the dart onto a map of nouns organized according to phonological similarities, has a much higher probability of landing in a neighborhood of masculine nouns, even if they do not dominate feminine nouns numerically.

5.1.0. Gender Assignment Task⁸. Although the simulations just presented show that AM makes empirically valid predictions, they do not help answer the question regarding how much of a noun's phonological material must be taken in to consideration in making gender determination. That is, no statistically significant difference was found between the four sets of variables used to predict gender assignment. The purpose of the gender assignment task is to determine if the same holds true when native speakers are asked to assign gender to novel words.

5.1.1. Stimulus Materials. 118 nouns were extracted from *Diccionario de la lengua española* (Real Academia Española, 1995). Each of these words is considered antiquated and of infrequent use according to the compilers of the dictionary (see Table 9), and none of them appears in Juilland and Chang-Rodríguez (1964). Therefore, they were highly unlikely to be known by the subjects, which also means that their gender would be unknown. In a similar study, Natalicio (1983) had subjects assign gender to antiquated words. Their assignment reflected the phonological structure of the words, rather than the words' actual gender, which indicates that the words were truly unknown to the subjects. In the present study, words were chosen that ended in phonemes other than *o* and *a*. In this way, the more obvious gender/phoneme correspondences were eliminated, and the subjects were obliged to make gender assignments on the more ambiguous cases.

5.1.2. Subjects. 31 literate native Spanish speakers from Spain participated in the study, 18 women and 13 men. The average age of the subjects was 33.4.

5.1.3. Procedure. The 118 test items were presented in the form of a written questionnaire. The subjects were asked to circle either the feminine article *la* or the masculine article *el*, which appeared before each test item. They were instructed to choose the article that was most appropriate for the word that followed.

5.1.4. Results and Discussion. The gender of the survey word was calculated as the most common answer given by the test subjects. In the case of two words (*sorce* and *barrunte*) the subjects' intuitions were evenly divided between masculine and feminine, and these words were eliminated. The remaining 116 test items were also assigned gender by AM based on analogy with the items in the database. The four different conditions shown in Table 7 were applied and compared with the subject's preferences. Bull's rules were also applied and compared. As already mentioned, Bull does not assign a gender to words ending in -z. Nevertheless, there were 16 -z final items in the survey. The subjects preferred the feminine in eight of those cases, and masculine in the other eight. Therefore, the success rate of Bull's rules does not change regardless of the gender that is assigned to -z. Items that were calculated by AM to be equally likely to be of either gender are marked "50/50" in the database, and are only counted as half correct.

++Insert Tables 8 and 9 here++

As Table 8 demonstrates, by utilizing the final syllable or the penultimate rime and syllable gender is assigned correctly at a level that slightly exceeds that of Bull's rules. However, the success rates in the four experimental conditions do not differ significantly from each other ($\chi^2(3) = 3.052, p < 0.5$).

6. CONCLUSIONS. In the preceding pages, AM has been presented as a model

that is able to predict linguistic behavior on the basis of memory tokens. When applied to the question of Spanish gender assignment, it does a formidable job of assigning gender based on the surface properties of Spanish words, and is often more successful than Bull's rules. AM's ability to account for psycholinguistic processing is demonstrated by the fact that it makes the same sort of errors as children do. In addition, it predicted the gender of the unknown words it was presented in a way that closely mirrored Spanish speakers' intuitions regarding the gender those words. With AM as a tool of linguistic analysis, it appears that phonemic material besides the word-final phoneme may be relevant in determining gender assignment.

1. Dual-gendered nouns such as *mar* 'sea', and *estudiante* 'male or female student' were not included, however.
2. A ninth variable, the word itself, was also included. However, since each word in the database is distinct, this variable does not play a part in gender assignment by AM. What it does is to maintain the distinction between those database items that rhyme (e.g. *acusación* and *admiración*). In this way, *acusación* can serve as a possible analog for *admiración*, and vice versa. Otherwise, when predicting the gender of *admiración*, for example, all nouns ending in *-ación* would be eliminated from consideration.
3. The conception of the mental lexicon that the present study follows most closely is that of Bybee (1985, 1988).
4. In this study, the phonemic attributes of words are assumed to be the relevant variables. However, AM can also incorporate other variables such as sociolinguistic variables: age, sex, social class etc. (Skousen 1989:97-100).
5. The algorithm is available on the internet at: <http://humanities.byu.edu/am>
6. This number includes both the singular and plural form of the noun.
7. Since 2416 items do not divide equally by 10, all data sets contained 241 items except the tenth which contained 247.
8. I am most indebted to Milagros Malo Fernández and Elías Álvarez Ortigosa who generously gave of their time to administer the questionnaires.

References

- Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6, 37-66.
- Allen, P., McNeal, M., & Kvak, D. (1992). Perhaps the lexicon is coded as a function of word frequency. *Journal of Memory and Language*, 31, 826-44.
- Bergen, J. J. (1978). A simplified approach for teaching the gender of Spanish nouns. *Hispania*, 61, 865-876.
- Bradley, D. (1980). Lexical representation of derivational relation. In M. Aronoff and M. L. Keaton (Eds.), *Juncture* (pp. 37-55). Saratoga, Cal.: Anma Libri.
- Brisk, M. E. (1976). The acquisition of Spanish gender by first grade Spanish-speaking children. In G. D. Keller, R. V. Teschner, & S. Viera (Eds.), *Bilingualism in the bicentennial and beyond* (pp. 143-160). New York: Bilingual Press.
- Brown, R., & McNeill, D. (1966). The 'tip of the tongue' phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5, 325-337.
- Bull, W. E. (1965). *Spanish for teachers: Applied linguistics*. New York: Ronald Press Co.
- Burzio, L. (1996). Surface constraints versus underlying representations. In J. Durand, B. Laks (Eds.), *Current trends in phonology: Models and methods*, (pp. 97-112). Paris X and University of Salford: University of Salford Publications.
- Bybee, J. (1985). *Morphology*. Amsterdam: John Benjamins.
- _____. (1988). Morphology as lexical organization. In M. Hammond, & M. Noonan (Eds.), *Theoretical approaches to morphology*, (pp. 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. In M. Gruber, C. D. Higgins, K. S. Olson, & T.

Wysocki (Eds.), *Proceedings of the Chicago Linguistic Society*, vol. 34, (pp. 421-435).

Chicago: Chicago Linguistic Society.

_____. (2001). *Phonology and language use*. Cambridge: Cambridge UP.

Chandler, S. (1993). Are rules and modules really necessary for explaining language?

Journal of Psycholinguistic Research, 22, 593-606.

_____. (1995). Non-declarative linguistics; Some neuropsychological perspectives.

Rivista di Linguistica, 7, 233-247.

_____. (2002). Skousen's analogical approach as an exemplar-based model of

categorization. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.), *Analogical*

Modeling: An Exemplar-based Approach to Language. John Benjamins. In press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and

Row.

Daelemans, W., Gillis, S., & Durieux, G. (1994). Skousen's analogical modeling

algorithm: A comparison with lazy learning. In D. Jones (Ed.), *Proceedings of the*

International Conference on New Methods in Language Processing, (pp. 1-7).

Manchester: UMIST.

- Daelemans, W., Zavrel, J, van der Sloot, K., & van den Bosch, A. (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)).
- Daugherty, K., & Seidenberg, M. S. (1994). Beyond rules and exceptions: A Connectionist approach to inflectional morphology. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules*, (pp. 353-388). Amsterdam: Benjamins.
- Derwing, B. L. (1973). *Transformational grammar as a theory of language acquisition*. London: Cambridge University Press.
- Eddington, D. (1996). The psychological status of phonological analyses. *Linguistica*, 36, 17-37.
- _____. (2000a). Analogy and the dual-route model of morphology. *Lingua*, 110, 281-298.
- _____. (2000b). Spanish stress assignment within the Analogical Modeling of Language. *Language*, 76, 92-109.
- _____. (2001). Dissociation in Italian conjugations: A single-route account. *Brain and Language*. Forthcoming.
- Frisch, S. (2000). Temporally organized lexical representations as phonological units. In M. B. Broe, J. P. Pierrehumbert (Eds.), *Papers in laboratory phonology V*:

- Acquisition and the lexicon*, (pp. 283-298). Cambridge: Cambridge University Press.
- Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language*, 42, 481-496.
- Frisch, S. A., Large, N. R., Zawaydeh, B., & Pisoni, D. B. (2001). In J. L. Bybee, and P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.
- Frisch, S. A., & Zawaydeh, B. (2001). The psychological reality of OCP-place in Arabic. *Language*, 77, 91-106.
- García, M. E. (1998). Gender marking in a dialect of Southwest Spanish. *Southwest Journal of Linguistics*, 17, 49-58.
- Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in speech processing*, (pp. 33-65). San Diego: Academic.
- Harris, J. W. (1985). Spanish word markers. In F. H. Nuessel Jr. (Ed.), *Current issues in Hispanic phonology and morphology*, (pp. 34-54). Bloomington, IN: Indiana Linguistics Club.
- _____. (1991). The exponence of gender in Spanish. *Linguistic Inquiry*, 22, 27-62.
- Jaeger, J. J. (1980). Testing the psychological reality of phonemes. *Language and Speech*, 23, 233-253.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The Basis of Consistency Effects in Word Naming. *Journal of Memory and Language*, 29, 687-715.

Jones, D. (1996). *Analogical natural language processing*. London: UCL press.

Juilland, A., & Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton.

Kenstowicz, M. (1996). Base-identity constraints and uniform exponence: Alternatives to cyclicity. In J. Durand and B. Laks (Eds.), *Current trends in phonology: Models and methods* (pp. 363-393). Paris X and University of Salford: University of Salford Publications.

_____. (1998). Uniform Exponence: Exemplification and Extension. Manuscript, MIT. (ROA 218, <http://rucss.rutgers.edu/roa.html>).

Kiparsky, P. (1975). What are phonological theories about? In D. Cohen, and J. R. Wirth (Eds.), *Testing linguistic hypotheses* (pp. 187-209). Washington D. C.: Hemisphere.

Klein P. W. (1983). Spanish gender morphology. *Papers in Romance*, 5, 57-64.

_____. (1989). Spanish 'gender' vowels and lexical representation. *Hispanic Linguistics*, 3, 147-162.

Krott, A. (2002). Analogical hierarchy: Exemplar-based modeling of linkers in Dutch noun-noun compounds. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.), *Analogical Modeling: An Exemplar-based Approach to Language*. John Benjamins. In press.

Lamb, S. (2000). Bidirectional processing in language and related cognitive systems. In

- M. Barlow and S. Kemmer (Eds.), *Usage-based models of language* (pp. 87-119).
Stanford, CA: CSLI Publications.
- MacKay, D. G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review*, 89, 60-94.
- MacWhinney, B., Leinbach, J., Taraban, R. & McDonald, J. (1989). Language learning: Rules or cues? *Journal of Memory and Language*, 28, 255-277.
- McCarthy, J. (1995.) Extensions of faithfulness: Rotuman revisited. Manuscript, University of Massachusetts, Amherst. (ROA 110, <http://ruccs.rutgers.edu/roa.html>).
- McCarthy, J. & Prince, A. (1994a). The emergence of the unmarked. Manuscript, University of Massachusetts, Amherst. (ROA 13, <http://ruccs.rutgers.edu/roa.html>).
- _____. (1994b). An overview of prosodic morphology. Manuscript, University of Massachusetts, Amherst. (ROA 59, <http://ruccs.rutgers.edu/roa.html>).
- McClelland, J. L. (1988). Connectionist models and psychological evidence. *Journal of Memory and Language*, 27,107-123.
- Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238.
- Natalicio, D. (1983). Native speaker intuitions as a basis for determining noun gender rules in Spanish. *Southwest Journal of Linguistics*, 6, 49-55.
- Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice

- attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 309-28.
- Pérez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. *Journal of Child Language*, 18, 571-590.
- Pierrehumbert, J. (1994). Syllable structure and word structure: A study of triconsonantal clusters in English. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III*, (pp. 168-188). Cambridge: Cambridge University Press.
- _____. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*, (pp. 137-157). Amsterdam: John Benjamins.
- Pinker, S., and Prince, A. (1994). Regular and irregular morphology and the psychological status of rules of grammar. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules*, (pp. 321-351). Amsterdam: Benjamins.
- Pisoni, D. (1997). Some thoughts on 'normalization' in speech perception. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing*, (pp. 9-32). San Diego: Academic.
- Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, 38, 3-102.

- Prado, M. (1982). El género en español y la teoría de la marcadez. *Hispania*, 65, 258-266.
- Prasada, S., Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8, 1-56.
- Real Academia Española. (1995). *Diccionario de la lengua española*. CD ROM version of the 21st edition. Madrid: Espasa Calpe.
- Riesbeck, C. K., & Schank, R. S. (1989). *Inside case-based reasoning*. Hillsdale, N.J.: Erlbaum.
- Rodríguez Bou, I. (1966). *Recuento de vocabulario de preescolares*. Río Piedras, puerto Rico: Consejo superior de enseñanza.
- Rosenblat, A. (1952). Género de los sustantivos en -e y en consonante. In *Estudios dedicados a Menéndez-Pidal*, vol. 3 (pp. 159-202). Madrid: Consejo Superior de Investigaciones científicas.
- Rytting, C. A. (2000) An empirical test of analogical modeling: The /k/ ~ i alternation. In A. K. Melby, & A. R. Lommel (Eds.), *Lacus forum XVII: The lexicon*, (pp. 73-84). Fullerton, CA: Linguistic Association of Canada and the United States.
- Say, T., & Clahsen, H. (2001). Words, rules, and stems in the Italian mental lexicon. In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.), *Storage and computation in the language faculty*. (pp. 75-108). Dordrecht: Kluwer..
- Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology*, 3, 1-17.

- Sebastián-Gallés, N. (1996). Speech perception in Catalan and Spanish. In M. Carreiras, J. E. García Albea, & N. Sebastián-Gallés (Eds.), *Language Processing in Spanish* (pp.1-17). Mahwah, N.J. : Erlbaum.
- Skousen, R. (1975). *Substantive evidence in phonology*. The Hague: Mouton.
- _____. (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 modeling. *Journal of Quantitative Linguistics*, 5, 246-255.
- Stemberger, J. P. (1994). Rule-less morphology at the phonology-lexicon interface. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules*, (pp. 147-169). Amsterdam: Benjamins.
- Stemberger, J. P., & MacWhinney, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, 14, 17-26.
- Steriade, D. (1997). Lexical conservatism and its analysis. Manuscript, UCLA.
- _____. (1999). Lexical conservatism. In Linguistic Society of Korea (Ed.), *Linguistics in the morning calm*, vol. 4, (pp.157-180). Hanshin, Korea: Linguistic Society of Korea.
- Teschner, R. V. (1983). Spanish gender revisited: -Z words as illustrating the need for expanded phonological and morphological analysis. *Hispania*, 66, 252-256.

Teschner, R. V., & Russell, W. M. (1984). The gender patterns of Spanish nouns: An inverse dictionary-based analysis. *Hispanic Linguistics*, 1, 115-132.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In M. B. Broe, J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*, (pp. 269-282). Cambridge: Cambridge University

- (1) The word final phoneme.
- (2) The nucleus of the final syllable if the syllable is closed or the non-specification marker '-' if the syllable is open.
- (3) The tautosyllabic phoneme preceding the nucleus of the final syllable or the syllable boundary symbol '0' if there is none.
- (4) The tautosyllabic phoneme preceding (3), or the syllable boundary symbol '0' if there is only one phoneme in the onset, else the non-specification marker '-'.
- (5) The tautosyllabic phoneme preceding (4), or the syllable boundary symbol '0' if the onset contains two phonemes, else the non-specification marker '-'.
- (6) The second tautosyllabic phoneme following the syllable nucleus of the penultimate syllable (9), or the syllable boundary marker '0' if only one phoneme follows (9), else '-'.
- (7) The first tautosyllabic phoneme following the syllable nucleus of the penultimate syllable (9), or the syllable boundary marker '0' if the syllable is open, else '-'.
- (8) The syllable nucleus of the penultimate syllable, or '0' if the noun is monosyllabic.

Table 1. Variable Assignment.

Subcontexts	Members of the Subcontext	Pointers	# of Disagreements
p a n	none	none	0
p a n	<i>plan M</i>	<i>plan M > plan M</i>	0
p ~n	none	none	0
p a û	<i>paz F, par M</i>	<i>paz F > paz F</i> <i>par M > par M</i> <i>*paz F > par M</i> <i>*par M > paz F</i>	2
p a û	<i>cal F</i>	<i>cal F > cal F</i>	0
p ~n	<i>tren M, ron M</i>	<i>tren M > tren M</i> <i>ron M > ron M</i> <i>tren M > ron M</i> <i>ron M > tren M</i>	0
p ~ û	none	none	0
p ~ û	<i>rey M</i>	<i>rey M > rey M</i>	0

Table 2. Subcontexts and their Disagreements.

Supracontext	Subcontexts in Supracontext	# Subcontextual Disagreements
p a n	pan	0
p a -	pan, *pa ^u	2
p - n	pan, p~n	0
- a n	pan, p ⁿ	0
p - -	pan, p~n, *pa ^u p~ ^u	2
- a -	pan, p ⁿ , *pa ^u p ^u	2
- - n	pan, p ⁿ , p~n, p ⁿ	0
- - -	pan, p ⁿ , p~n, *pa ^u p ^u p ⁿ , p~ ^u p ^u	2

Table 3. Subcontextual Analysis.

Supracontext	Words in Supracontext	Pointers	# of Subcontextual Disagreements
p a n	none	none	0
p a -	<i>par M, paz F</i>	<i>paz F > paz F</i> <i>par M > par M</i> <i>*paz F > par M</i> <i>*par M > paz F</i>	2
p - n	none	none	0
- a n	<i>plan M</i>	<i>plan M > plan M</i>	0
p - -	<i>par M, paz F</i>	<i>paz F > paz F</i> <i>par M > par M</i> <i>*paz F > par M</i> <i>*par M > paz F</i>	2
- a -	<i>plan M, cal F, par M, paz F</i>	<i>plan M > plan M</i> <i>*plan M > cal F</i> <i>plan M > par M</i> <i>*plan M > paz F</i> <i>cal F > cal F</i> <i>*cal F > plan M</i> <i>*cal F > par M</i> <i>cal F > paz F</i> <i>*paz F > plan M</i> <i>*paz F > par M</i> (not all shown; all the rest involve agreement)	6
- - n	<i>plan M, tren M, ron M</i>	(not shown)	0
- - -	<i>plan M, tren M, ron M, cal F, paz F, par M, rey M</i>	(not shown)	12

Table 4. Supracontextual Analysis.

Homogenous Supracontext	Words in Supracontext	Pointers	# of Pointers to Masculine	# of Pointers to Feminine
p a -	<i>par M, paz F</i>	<i>par M > par M</i> <i>par M > paz F</i> <i>paz F > par M</i> <i>paz F > paz F</i>	2	2
- a n	<i>plan M</i>	<i>plan M > plan M</i>	1	0
p - -	<i>par M, paz F</i>	<i>par M > par M</i> <i>par M > paz F</i> <i>paz F > par M</i> <i>paz F > paz F</i>	2	2
- - n	<i>plan M, tren M,</i> <i>ron M</i>	(not shown)	9	0

Table 5. Non-empty Homogenous Supracontexts.

Word	Actual Gender	Probability of Masculine	Probability of Feminine
<i>tabaco</i>	masc.	.992	.008
<i>inglés</i>	masc.	.814	.186
<i>edición</i>	fem.	0	1.000
<i>hierba</i>	fem.	0	1.000
<i>ideal</i>	masc.	.608	.392
<i>problema</i>	masc.	1.000	0
* <i>planeta</i>	masc.	.100	.900
* <i>catedral</i>	fem.	.950	.050

Table 6. Sample Outcomes of Gender Probability Based on AM.

Exp. Condition	% Correctly Assigned Gender with Type Frequency	% Correctly Assigned Gender with Token Frequency
1) Final Phoneme	93.8	93.6
2) Final Rhyme	94.5	94.1
3) Final Syllable	95.5	93.8
4) Penultimate Rhyme and Final Syllable	96.4	95.0

Table 7. Outcome of the Four Experimental Conditions.

Condition	% Agreement
1) Final Phoneme	62.9
2) Final Rhyme	71.1
3) Final Syllable	81.0
4) Penultimate Rhyme and Final Syllable	77.6
5) Bull's Rules	75.0

Table 8. Percent of Agreement Between AM and Questionnaire Outcomes.

Word	Survey	Bull's	F. Phoneme	F. Rime	F. Syl.	F. Syl. & P. Rime
abarraz	M	x	x	x		
acates	M					
acemite	M					
acordación	F					
acumen	M		x			
afer	M					
afice	M					
alancel	M				x	
alcaduz	M	x	x	x	x	
alcalifaje	M					
alcamiz	M	x	x	x		
alinde	M					
alioj	M		x	50/50	x	x
alizace	M					
amarillor	M					
anascote	M					
arrafiz	M	x	x	x		
asperez	F					
atarfe	M					
avarientez	F					x
azcón	M		x	x		
acoche	M					
balizaje	M					
beudez	F					

bitumen	M		x			
bocacín	M		x			
botor	M					
broznedad	F					
cabción	F					
cabrial	M					
cafiz	M	x	x	x		
calicud	M	x	x	x	x	x
calonge	M					
cambil	M					
candelor	M					
canez	F					x
carauz	M	x	x	x		
ceción	F					
celtre	F	x	x	x	x	x
cifaque	M					
cipión	M		x	x	x	x
cobil	M					
coce	F	x	x	x	x	x
cocadriz	F					
compage	F	x	x	x	x	x
consuetud	F					
consulaje	M					
copanete	M					
cotrofe	M					

crenche	F	x	x	x	x	x
criazón	F	x			50/50	50/50
crochel	M					
chivital	M					
delate	M					
desdón	M		x	x		
deslate	M					
destín	M		x			
disfrez	M	x	x	x		
egeción	F	x				
elébor	M					
emiente	F	x	x	x	x	x
entalle	M					
entenzón	M		x	x	50/50	x
epiglosis	F		x			
escambrón	M		x	x		x
escorche	M					
escrocón	M		x	x		
esgambete	M					
esguarde	M					x
esledor	M					
estipe	F	x	x	x	x	x
estruz	F					
evagación	F					
fabledad	F					

fenestraje	M					
fluxión	F		x			
folguín	M		x			
fosal	F	x	x	x	50/50	x
gafez	M	x	x	x	x	x
gagate	M					
garifalte	M					
grasor	M					
gubilete	M					
guiaje	M					
ingre	F	x	x	x		x
jusente	M					
lailán	F	x		x	x	x
lande	M					
lavajal	M					
lerdez	F					
linamen	M		x			
mandrial	M					
mansuetud	F					
másticis	F	x	x			
menge	F	x	x	x	x	x
meridión	M		x	x	x	x
merode	F	x	x	x	x	x
nacre	F	x	x	x	x	
orebce	M					

palude	F	x	x	x	x	x
panol	M					
peraile	M					
pernicie	F	x	x	x		
pólex	M					
primaz	M	x	x	x		50/50
pujés	M					
realme	M					
rebalaj	F	x	x		x	x
riste	M					
senojil	M					
sozprior	M					x
tabelión	M		x	x	x	x
trascol	M					
velambre	F	x	x	x	50/50	
venadriz	F					
venderache	M					

F. = Final, Syl. = Syllable, P. = Penultimate, x = error

Table 9. Comparison of Errors.

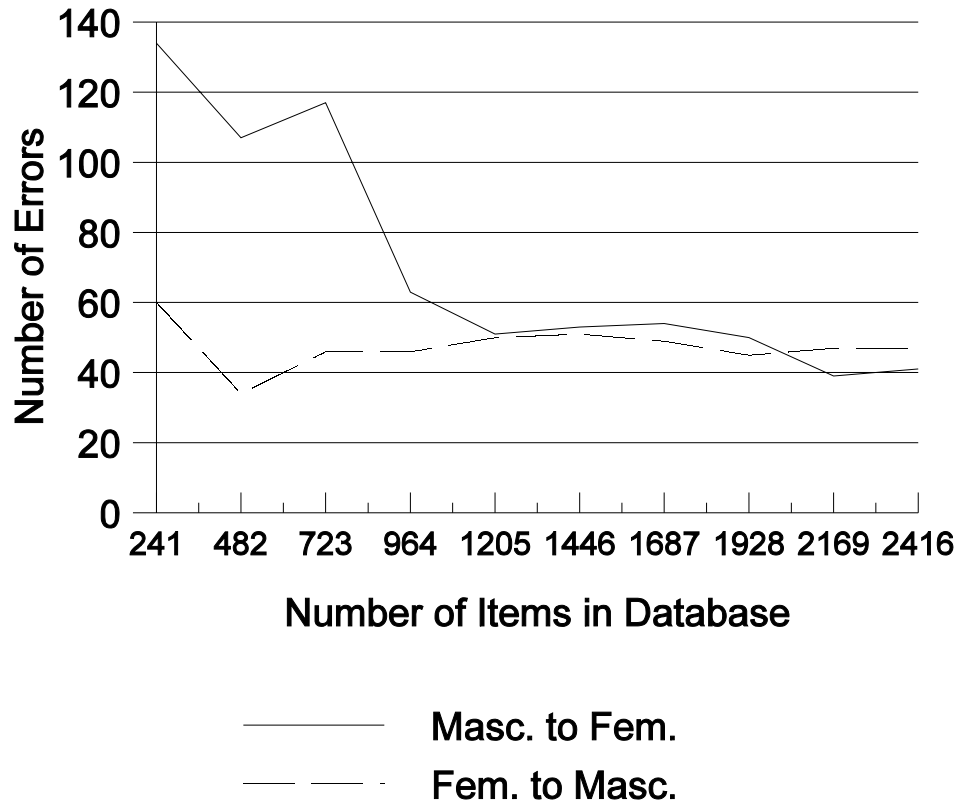


Figure 1. Errors by Size of Database.