Modeling analogy as probabilistic grammar[*]

Adam Albright
MIT

March 2008

1 Introduction

Formal implemented models of analogy face two opposing challenges. On the one hand, they must be powerful and flexible enough to handle gradient and probabilistic data. This requires an ability to notice statistical regularities at many different levels of generality, and in many cases, to adjudicate between multiple conflicting patterns by assessing the relative strength of each, and to generalize them to novel items based on their relative strength. At the same time, when we examine evidence from language change, child errors, psycholinguistic experiments, we find that only a small fraction of the logically possible analogical inferences are actually attested. Therefore, an adequate model of analogy must also be restrictive enough to explain why speakers generalize certain statistical properties of the data and not others. Moreover, in the ideal case, restrictions on possible analogies should follow from intrinsic properties of the architecture of the model, and not need to be stipulated post hoc.

Current computational models of analogical inference in language are still rather rudimentary, and we are certainly nowhere near possessing a model that captures not only the statistical abilities of speakers, but also their preferences and limitations.[1] Nonetheless, the past two decades have seen some key advances. Work in frameworks such as neural networks (Rumelhart and McClelland 1987; MacWhinney and Leinbach 1991; Daugherty and Seidenberg 1994, and much subsequent work) and Analogical Modeling of Language (AML; Skousen 1989) have focused primarily on the first challenge, tackling the gradience of

[*] Thanks to Jim Blevins, Bruce Hayes, Donca Steriade, participants of the Analogy in Grammar Workshop (Leipzig, 22–23 September 2006), and especially to two anonymous reviewers, for helpful comments and suggestions; all remaining errors are, of course, my own.

the data. This work has had several positive influences on the study of analogy, particularly as a synchronic phenomenon. First, it has fostered a culture of developing computationally implemented models. These allow for objective tests of the extent to which a particular pattern can be extracted from the training data, given an explicitly formalized set of assumptions. In a few cases, such work has even led to implemented models of analogical change over time (e.g., Hare and Elman 1995). More generally, it has inspired a good deal of empirical work probing the detailed statistical knowledge that native speakers have about regularities and subregularities surrounding processes in their language. The overall picture that has emerged from such work is one of speakers as powerful statistical learners, able to encode a wide variety of gradient patterns.

In this paper, I will take on the latter side of the problem, which has so far received far less attention in the literature: why do speakers generalize some regularities and not others? I discuss three general restrictions on analogical inference in morphophonology. The first is a restriction on how patterns are defined, which distinguishes between patterns that can be noticed and extended, and those that are evidently ignored. The second is a restriction on how patterns are evaluated, and concerns what it means for a pattern to be "well attested" or strong enough to generalize to novel items. The last is a restriction on which forms in a morphological paradigm are open to analogical change, and what determines the direction of influence. I argue that in all three cases, the observed restrictions correspond to limitations imposed by formulating processes as SPE-style rewrite rules (A → B / C ___ D). This observation is not a trivial one, since this rule notation is a very particular hypothesis about how linguistic knowledge is structured, and how it makes reference to positions, variables, and so on. I demonstrate ways in which statistical models that lack this type of structure suffer in their ability to model empirical data, by overestimating the goodness of various possible but unattested types of analogical inference. Based on this observation, I argue that the best formal model of analogy is one that adds a probabilistic component to a grammar of context-sensitive statements.

The outline of the paper is as follows: for each of the three proposed restrictions, I first present empirical data illustrating how it distinguishes attested from unattested analogies. Then, I compare two representative models, one with and one without the restriction imposed by rule-like structure. Finally, I discuss the broader implications of these observations for formal models of analogy.

2 What is a linguistically significant pattern?

2.1 Structured vs. unstructured inference

To illustrate the role that a formalism can play in restricting possible analogies, it is instructive to start by considering the most traditional of all formalisms: four-part analogy. In four-part notation, analogies are expressed in the form in (1):

(1) Four-part notation: A:B :: X:Y

“Whatever the relationship is between A and B, it should also hold between X any Y”

Discussions of four-part analogy frequently point out that the relation between words A and B is in many cases part of a much more general pattern, and that the examples A and B should be construed as representative members of a larger analogical set, consisting of more words (A_1:B_1 :: A_2:B_2 :: A_3:B_3 :: $\ldots$) and perhaps also more paradigmatically related forms (A_1:B_1:C_1 :: A_2:B_2:C_2 $\ldots$). The notation itself does not provide any way to indicate this fact, however, and thus has no formal means of excluding or disfavoring analogies supported by just one or a few pairs. Furthermore, the notation does not impose any restrictions on what properties particular A_i:B_i pairs can have in common with one another. In fact the pattern itself—i.e., the relation between A and B, and the equation for Y—is left entirely implicit. This means that there are many possible ways to construct analogical sets, and few concrete ways to compare competing analogical inferences.

As an example, consider mid vowel alternations in Spanish present tense indicative verb paradigms. In some verbs, when the mid vowels /e/ and /o/ are stressed, they irregularly diphthongize to [jé] and [wé], respectively. This occurs in the 1sg, 2sg, 3sg, and 3pl (as well as the entire present subjunctive). In other verbs, the alternation does not occur, and invariant mid vowels or diphthongs are found throughout the paradigm.

(2) Spanish present tense diphthongization

a. Diphthongizing verbs

Verb stem	Infin.	3sg. pres. indic.	Gloss
sent-	sent-ár	sjént-a	‘seat’
kont-	kont-ár	kwént-a	‘count’

b. Non-alternating verbs

Verb stem	Infin.	3sg. pres. indic.	Gloss
rent-	rent-ár	rént-a	‘rent’
mont-	mont-ár	mónt-a	‘ride/mount’
orjent-	orjent-ár	orjént-a	‘orient’
frekwent-	frekwent-ár	frekwént-a	‘frequent’

Since diphthongization is lexically idiosyncratic, Spanish speakers must decide whether or not to apply it to novel or unknown words. For example, if a speaker was faced with a novel verb [lerrár], they might attempt to construct analogical sets that would support a diphthongized 3sg form [ljérra]. Using the four-part notation, there are numerous ways this could be done, including:

(3) Analogical set 1:

$$\left\{\begin{array}{l} \textit{errar}:\textit{yerra} \\ \textit{enterrar}:\textit{entierra} \\ \textit{aserrar}:\textit{asierra} \\ \textit{aferrar}:\textit{afierra} \\ \textit{cerrar}:\textit{cierra} \\ \ldots \end{array}\right\} \quad :: \quad \textit{lerrar}:\textbf{\textit{lierra}}$$

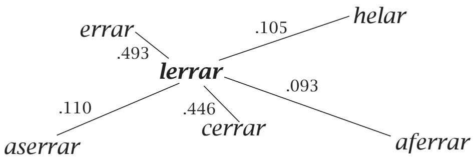
(4) Analogical set 2:

$$\left\{\begin{array}{l} \textit{serrar}:\textit{sierra} \\ \textit{alentar}:\textit{alienta} \\ \textit{helar}:\textit{hiela} \\ \textit{querer}:\textit{quiere} \\ \ldots \end{array}\right\} \quad :: \quad \textit{lerrar}:\textbf{\textit{lierra}}$$

The first set looks more convincing, since all of its members rhyme with [lerr-] and belong to the *-ar* inflectional class. Intuitively, this provides better support for the outcome [ljérra] than set 2 does; however, such a high degree of similarity is neither required nor rewarded by the formalism. In addition, nothing formally rewards a larger set (a point we will return to below). In sum, while the generality and flexibility of four-part notation have made it convenient tool for describing analogical changes, as is often noted in the literature, an explanation theory of analogy depends on being able to impose restrictions on possible proportions (Morpurgo Davies 1978).

Let us start by addressing the first shortcoming of four-part notation, namely, its inability to capture the relative similarity of different analogical pairs to the target word. A common intuition about analogical sets is that they are not chosen randomly from the lexicon at large, but rather should represent the words that are expected to have the greatest influence because they are phonologically most similar to the target word—i.e., the closest analogs. For example, the existing Spanish verbs that are most similar to the novel verb [lerrar] are shown in (5) (similarity values are in arbitrary units, higher = more similar):

(5) Existing Spanish verbs similar to [lerrar]



The restriction that we want the model to obey, then, is that generalization of a pattern to novel items must be supported by sufficiently many close analogs. One obvious way to do this is to adopt a similarity-based classification model, which decides on the treatment of novel items by considering its aggregate similarity to the set of known items. In such a model, the advantage of being similar to many existing words is anything but accidental; it is built in as core principle of the architecture of the model.

There are many ways to be similar, however, and it is an empirical question what types of similarity matter most to humans in deciding how to treat novel words. For instance, the existing Spanish verbs *errar* ‘err’ and *cerrar* ‘close’ are similar to novel *lerrar* by ending in root-final [err]. The verb *helar* ‘freeze’ is also (at least somewhat) similar to *lerrar*, but this is due to the shared [l] (or perhaps the similarity of [l] and [r]), a similar syllabic structure, and so on. Hypothetical verbs like *lerdar*, *lenar*, and *lorrar* also share commonalities with *lerrar*, but each in its own unique way. Looking back at analogical set 1 in (3), there are intuitively two factors that make this group of analogs seem particularly compelling. First, all of these verbs share a set of common properties with each other and with the target word: they all end in [err] and all belong to the -*ar* inflectional class. In addition, those shared properties are perceptually salient (involving rhymes of stressed syllables), and are local to the change in question (being either in the same syllable as the stressed mid vowel, or in the adjacent syllable). Albright and Hayes (2003) refer to this situation, in which the comparison set can be defined by their shared properties, as STRUCTURED SIMILARITY. If we compare analogical set 2 in (4), we see that *serrar*, *alentar*, *helar*, and *querer* share no such properties.[2] Albright and Hayes refer to this as VARIEGATED SIMILARITY.

Not all similarity-based models care about the exact source or nature of similarity. In principle, the similarity of the novel word to each existing word could be calculated independently. (An example will be given in the next section.) In order to give preference to structured similarity, a model must be able to align words with one another, determine what they have in common, and ignore what is unique to individual comparisons. This requires that the model have the capacity to encode the fact that a number of words all have the same type of element in the same location—that is, the model must be able to impose structure on the data, and encode its knowledge in terms of these structures (features, prosodic positions, etc.). This sounds like a simple requirement, but in fact it represents a fundamental divide between two classes of models: those that generalize using “raw” (unstructured) similarity to known words, and those that generalize by imposing structure on novel items and parsing them for elements in common with known words.

The goal of the rest of this section is to show that structured similarity is an important component in modeling how speakers generalize morphophonological patterns. The strategy will be as follows: first, in sections 2.2–2.3, I will present two computationally implemented models, one lacking structured representations, and one that encodes its knowledge in structural terms. Then in section 2.4, the performance of the two models will be compared against experimentally obtained data concerning the relative likelihood of different novel Spanish verbs to undergo diphthongization. To preview the results, it will emerge that the ability to make use of variegated (unstructured) similarity turns out to be not only unnecessary, but even harmful in modeling human intuitions.

2.2 Analogy without structure: “pure” similarity-based classification

To asses the contribution of structured similarity to the performance of a model, we first need a baseline model that does not require structured comparisons. One commonly used model of similarity-based classification that has been widely applied in many domains is the GENERALIZED CONTEXT MODEL (GCM; Nosofsky 1986, 1990). For some applications in linguistics, see Johnson (1997), Nakisa, Plunkett, and Hahn (1997), and Albright and Hayes (2003). In this model, the treatment of a novel item is determined by calculating its similarity to classes of known items (exemplars). In deciding whether to assign a novel item *i* to a particular class *c*, the model compares item *i* to each existing member *j* of class *c*. The similarity of *i* to the entire class is a function of the summed similarities of each individual class member:

(6) Similarity of novel item *i* to class *c* (with members *j*) $= \sum e^{(-d_{i,j}/s)}$, where

- $d_{i,j} =$ the psychological distance between *i* and *j*
- $s =$ sensitivity (a free parameter of the model)

The probability of actually treating *i* as a member of class *c* is simply proportional to its similarity to the individual members:

(7) Probability of assigning item *i* to class $c = \dfrac{\text{Similarity of } i \text{ to } c}{\text{Total similarity of } i \text{ to all classes}}$

This model is based on the premise that analogical sets are more compelling when they contain more members, and when those members are more similar to the novel item. In this way, the model satisfies the restriction that analogical generalization must be sufficiently supported by known items. The model does not place any inherent restrictions on the nature of the similarity relations, however, specifying only that it reflect some generic notion of the *psychological distance* between two words. At its simplest and most neutral, this would simply be their *perceptual distance*, or some holistic measure of how similar the words sound. Intuitively, words sound similar to one another if their component segments are similar—that is, if the sounds of one word are well-matched to those of the other. In order to calculate this, we need perceptual similarity values for arbitrary pairs of sounds, and also a method of determining the optimal alignment of sounds, given their similarities.

One technique for estimating the similarity of pairs of segments is to consider how many natural classes they both belong to. Frisch, Pierrehumbert and Broe (2004), following Broe (1993) and Frisch (1996), propose the following ratio:

(8) $$\text{Similarity of sounds } s_1, s_2 = \frac{\text{Number of shared natural classes}}{\text{Number of shared + unshared natural classes}}$$

Given these similarity values, an optimal alignment of the sounds in two words is one in which they can be transformed into one another in as few steps as possible (Bailey and Hahn 2001; Hahn, Chater, and Richardson 2003). This can be calculated by finding the minimum string edit (Levenshtein) distance (Kruskal 1983); see Bailey and Hahn (2001) and Albright and Hayes (2003) for details of how this is implemented based on segmental similarity. The result is a score for each pair of words, reflecting the degree of similarity between corresponding segments and the extent of mismatches (non-corresponding material). For example, the similarity of the novel verb *lerrar* to the existing Spanish verb *errar* is calculated to be 0.493 (in arbitrary units), while the similarity of *lerrar* to *reglar* is 0.268, and to *lograr* is 0.203.

We can use this model to calculate the likelihood of diphthongizing a novel Spanish verb, by simply comparing the aggregate similarity of that verb against the set of existing diphthongizing and non-diphthongizing verbs. For example, the summed similarity of the novel verb *lerrar* to diphthongizing verbs is 4.936 (again, in arbitrary units), with the top contributors including verbs like *errar* (0.493), *cerrar* (0.446), *aserrar* (0.110), *helar* (0.105), and *aferrar* (0.093). The summed similarity of *lerrar* to non-diphthongizing verbs is 15.551, with top contributors including *reglar* (0.268), *orlar* (0.240), *ahorrar* (0.213), *forrar* (0.211), and *lograr* (0.203). We see that the higher score for the non-diphthongizing comes not from greater similarity of any individual member—in fact, *errar* and *cerrar* in the diphthongizing class are much more similar than any non-diphthongizing verb. Rather, this advantage is due to the fact that there are many more non-diphthongizing verbs, so small amounts of moderate similarity sum up to outweigh a small numer of very similar verbs. Using the equation in (7), the overall probability of applying diphthongization to *lerrar* is predicted to be 4.936 / (4.936 + 15.551), or 24.09%.

There are a couple points to note about the workings of this model. First, the model has the ability to make use of variegated similarity, since similarity is based on the optimal alignments of individual pairs of items. However, the examples in the preceding paragraph show that not all inferences make equal use of it; in fact, the closest analogs supporting diphthongization almost all contain *-errar*. This turns out to be quite typical, and analogical sets are frequently dominated by words that all happen to share the same feature(s) in common with the target word—i.e., a structured similarity. This aspect of the model will be important to keep in mind when evaluating the performance of the GCM, since we are interested not only in how well the model does, but also in the question of whether it benefits from its ability to use variegated similarity.

2.3 Analogy with structure: probabilistic context-sensitive rules

As noted above, an ability to refer to particular properties of words (having a certain type of sound in a certain location, having particular prosodic properties, etc.) is crucial in requiring that analogical sets share structural similarities. In fact, many modeling frameworks use structural properties to decide how to treat novel items. Feature-based classification models (Tversky 1977), such as TiMBL (Daelemans, Zavrel, Van der Sloot, and Van den Bosch 2000) and AML (Skousen 1989) directly incorporate the idea that in order for a group of items to be similar, they must share certain properties (feature values). Linguistic rules impose an even more specific structure. For example, context-sensitive readjustment rules (e → je / X ___ rro]$_{\text{1sg}}$) specify a change location, immediately adjacent left and right contexts, precedence relations, and so on. Although rule application is often thought of as fundamentally different from (and incompatible with) analogical inference, in fact, it is possible to think of rules as a very specific theory of how analogical sets are constructed—namely, by picking out groups of words that can be captured using the rule notation format.

The Minimal Generalization Learner (MGL; Albright and Hayes 2002) is a computationally implemented model that finds rules covering sets of words that behave consistently (belong to the same inflectional class, share the same morphophonemic change, etc.). It employs a bottom-up inductive procedure to compare pairs of words in the input data, find what they have in common, and encode these commonalities using a grammar of stochastic rules. For details of the model, the reader is referred to Albright and Hayes (2002) and Albright and Hayes (2003); in this section I provide a brief overview.

The model takes as its input pairs of forms that stand in a particular morphological relation, such as present/past, or infinitive/1sg, as in (9). In the present case, the relation between diphthongized and non-diphthongized stem variants is conditioned by stress placement, rather than any particular morphological category. Therefore, in the simulations reported here, input data are represented as pairs of stressed and stressless stem allomorphs, abstracting away from the suffixal material of the particular inflected forms that require one or the other, but retaining an indication of inflection class information (*-ar*, *-er*, *-ir*).

(9) Input to the minimal generalization learner: some sample *-ar* verbs

Stressless	Stressed	Gloss	Orthography (infinitive)
jeg	jég	‘arrive’	(*llegar*)
dex	déx	‘leave’	(*dejar*)
jeb	jéb	‘bring’	(*llevar*)
ked	kéd	stay	(*quedar*)
enkontr	enkwéntr	‘find’	(*encontrar*)
pens	pjéns	‘think’	(*pensar*)
kont	kwént	‘tell, count’	(*contar*)
entr	éntr	‘enter’	(*entrar*)
tom	tóm	‘take’	(*tomar*)
kre	kré	‘create’	(*crear*)
empes	empjés	‘start’	(*empezar*)
esper	espér	‘wait, hope’	(*esperar*)
rekord	rekwérd	‘remember’	(*recordar*)
tembl	tjémbl	‘tremble’	(*temblar*)

The first step in learning is to analyze individual (stressless, stressed) pairs, by factoring them into changing and unchanging portions. This allows each pair to be expressed as a rule, encoding both the change (A → B) and the non-changing portion (C ___ D). For example, the pair (tembl, tjémbl) has a vowel change surrounded by unchanging consonants: e → jé / t ___ mbl (“stressless [e] corresponds to stressed [je] when preceded by [t] and followed by [mbl]”). The pair (jeg, jég) on the other hand differs only in stress: e → é / j ___ g.

Once the input pairs have been re-cast as word-specific rules, they are compared to find what they have in common, according to the rule scheme in (10):

(10) Comparing *tembl*/*tiembl*- ‘tremble’, *desmembr*-/*desmiembr*- ‘dismember’:

Residue	Shared	Shared	Change	Shared	Shared
	feats	segs	loc.	segs	feats
	t		___	mb	l
des	m		___	mb	r
X	$\begin{bmatrix}-\text{syllabic}\\-\text{continuant}\end{bmatrix}$		___	mb	$\begin{bmatrix}-\text{syllabic}\\+\text{sonorant}\\+\text{continuant}\\+\text{voice}\\+\text{coronal}\\+\text{anterior}\end{bmatrix}$

The comparison in (10) yields a very specific rule that retains all of the properties shared by *tembl*- and *demembr*-, subject to the restriction that they can be encoded in the structural components of the rule. Shared material is expressed in terms of phonological features, while unshared material is expressed as variables. By convention, unmatched material on the left side is collapsed into a variable called ‘X’, and material on the right into a variable ‘Y’. When such comparisons are carried out iteratively across the entire data set, however, much broader rules can emerge through comparison of diverse forms, while further comparison of similar forms will yield additional narrow rules. A small sample of the many possible rules that could be learned from a set of Spanish verbs is given in (11).

(11) Representative rules for Spanish verbs[3]

i. o → wé / [+consonantal] ___ rs

ii. $\text{o} \rightarrow \text{wé} \ / \begin{bmatrix}-\text{continuant}\\-\text{voice}\end{bmatrix} \text{r} \ \_\_\_ \begin{bmatrix}-\text{continuant}\\-\text{syllabic}\end{bmatrix}$

iii. $\text{o} \rightarrow \text{wé} \ / \begin{bmatrix}-\text{syllabic}\\+\text{consonantal}\end{bmatrix} \_\_\_ \ [-\text{syllabic}]$

$$\text{iv.}\quad \text{o} \rightarrow \text{ó} \;/\; \begin{bmatrix} -\text{syllabic} \\ -\text{sonorant} \\ +\text{consonantal} \end{bmatrix} \;\_\_\_\; \begin{bmatrix} -\text{syllabic} \\ \text{+consonantal} \\ \text{-continuant} \end{bmatrix}$$

$$\text{v.}\quad \text{o} \rightarrow \text{ó} \;/\; \begin{bmatrix} -\text{syllabic} \\ +\text{voice} \end{bmatrix} \;\_\_\_\; [-\text{syllabic}]$$

$$\text{vi.}\quad \text{o} \rightarrow \text{ó} \;/\; \_\_\_\; [-\text{syllabic}]$$

These rewrite rules incorporate many types of structure that limit possible comparisons. Rules specify linear relations such as precedence and adjacency. This notation rules out many logically possible sets of words, such as those that all have a certain sound, but its location is variably either the right or the left of the change. This particular procedure also compares words by starting immediately adjacent to the change and working outwards, meaning that the descriptions of the left and right side contexts are limited to the local contexts.[4] Rule notation also embodies a form of strict feature matching: rules apply if their structural description is met, and not otherwise. Finally, although SPE-style rewrite rules are written in a way that could theoretically make use of the full power of context-sensitive grammars, the rules employed by this model obey commonly observed conventions for phonological rewrite rules by referring to a fixed number of positions and applying non-cyclically, and thus are restricted to expressing regular relations which can be captured with a finite state transducer (Johnson 1972; Kaplan and Kay 1994; Gildea and Jurafsky 1996). The system thus embodies a very strong form of structured similarity: all that matters is that words are the same in the relevant respect, and there are no penalties or rewards for additional similarities or differences.

Once all of the possible rules have been discovered, it remains to decide which dimensions of similarity the speaker should actually pay attention to. In order to do this, the rules are evaluated according to their accuracy in the training data. The RELIABILITY of a rule is defined as the number of cases that it successfully covers (its HITS), divided by the number of cases that meet its structural description (its SCOPE). Raw reliability scores are then adjusted slightly downwards using lower confidence limit statistics, to yield a score called CONFIDENCE. This has the effect of penalizing rules that are based on just a small amount of data (a small scope). The confidence scores for the rules in (11) are shown in (12):

(12) Representative rules for Spanish, evaluated (hits/scope ⇒ confidence)

i.	$\text{o} \rightarrow \text{wé} \,/\, [+\text{cons}] \,\_\_\_\, \text{rs}$	4/4	⇒	.786
ii.	$\text{o} \rightarrow \text{wé} \,/\, \begin{bmatrix} -\text{contin} \\ -\text{voice} \end{bmatrix} \text{r} \,\_\_\_\, \begin{bmatrix} -\text{contin} \\ -\text{syll} \end{bmatrix}$	6/8	⇒	.610
iii.	$\text{o} \rightarrow \text{wé} \,/\, \begin{bmatrix} -\text{syll} \\ +\text{cons} \end{bmatrix} \,\_\_\_\, [-\text{syll}]$	68/545	⇒	.116
iv.	$\text{o} \rightarrow \text{ó} \,/\, \begin{bmatrix} -\text{syll} \\ -\text{sonor} \\ +\text{cons} \end{bmatrix} \,\_\_\_\, \begin{bmatrix} -\text{syll} \\ +\text{cons} \\ \text{-contin} \end{bmatrix}$	101/106	⇒	.934
v.	$\text{o} \rightarrow \text{ó} \,/\, \begin{bmatrix} -\text{syll} \\ +\text{voice} \end{bmatrix} \,\_\_\_\, [-\text{syll}]$	19/22	⇒	.795
vi.	$\text{o} \rightarrow \text{ó} \,/\, \,\_\_\_\, [-\text{syll}]$	588/668	⇒	.871

Finally, the grammar of rules can be used to generalize patterns to novel items. The probability of generalizing a process is defined as in (13). Since this calculation is intended to mimic the probability with which a particular pattern will be employed to produce a target output, it is referred to as the PRODUCTION PROBABILITY of that pattern:

(13) Production probability

$$= \frac{\text{Confidence of the best rule applying the pattern to the input}}{\text{Summed confidence of best rules applicable to the input, for each pattern}}$$

For example, in calculating the likelihood to diphthongize the novel verb *lerrar*, the best (= most confident) applicable diphthongization and non-diphthongization rules are:

(14) Likelihood to diphthongize *lerrar*

- Best applicable diphthongization rule:
  - $\text{e} \rightarrow \text{jé} \;/\; \begin{bmatrix} +\text{consonantal} \\ +\text{coronal} \end{bmatrix} \_\_\_ \begin{bmatrix} +\text{consonantal} \\ +\text{voice} \end{bmatrix}$

    Reliability = 10/29; Confidence = .290

- Best applicable non-diphthongization rule:
  - $\text{e} \rightarrow \text{é} \;/\; \begin{bmatrix} -\text{syllabic} \\ +\text{voice} \end{bmatrix} \_\_\_ \; [+\text{sonorant}]$

    Reliability = 86/86; Confidence = .989

- Production probability(*lierro*) = $\dfrac{.290}{.290 + .989} = 23\%$

For both the Minimal Generalization Learner and the Generalized Context Model, support for generalizations comes from large numbers of words that are similar to the target word and behave consistently. In the MGL, however, similarity is defined (in boolean fashion) as presence of certain structural features. This prevents the model from using variegated similarity, since such diverse sets of relations cannot be captured in the rule notation. We can contrast this with the GCM, in which the supporting words need not be similar to one another in any particular way. This leads to the possibility that analogical inference may be based on variegated support. In the next section, we attempt to test whether this additional ability is helpful or harmful to the GCM.

Finally, it is worth noting that proportional analogy is most often used in a way that conforms to the structural restrictions imposed by the rule-based model, since the antecedent in four-part notation requires that there is a well-defined relation, and ideally also a group of words that all share the same relation. Although individual analysts may disagree about what constitutes a valid relation (see Morpurgo Davies 1978 for a review of some prominent points of view), in practice, relations are most naturally thought of as a single rewrite relation, much as in SPE-style rules. This is not to say that the formalisms are equivalent, however, since proportional analogy is certainly flexible enough to encompass relations that cannot be expressed in rule-based terms. For example, nothing formally precludes setting up proportions showing relations that involve multiple changes (prefixation of [s] and nasalization of final consonant: *tick*:*sting* ::

crab:*scram* :: *cat*::*scan*?), or changes that depend on the presence of an element somewhere in the word regardless of linear order (change of [ɪ] → [ʌ] adjacent to e [p]: *pinch*::*punch* :: *sip*::*sup* :: *pig*:*pug*?). A hypothesis of the rule-based model is that in order for a relation to be linguistically active—i.e., extended systematically to new forms—it must involve a change defined in terms of phonological features, applied to a set of words that share a common structure (again, defined over linearly arranged combinations of natural phonological classes).

2.4 An empirical test: modeling diphthongization in novel words

In order to test whether humans are restricted to inferences based on structured similarity, we can compare the performance of the two models against experimentally obtained data in which Spanish speakers were likewise tested on how they would produce stressed forms of novel verbs. Albright, Andrade, and Hayes (2001) asked 96 native speakers to inflect novel verbs containing mid vowels, to measure the relative likelihood of diphthongized responses in different contexts. Participants were given novel verbs in an unstressed form (e.g., [lerrámos] 'we *lerr*') and were asked to produce a stressed form (e.g., [lerro]/[ljérro] 'I *lerr*'), For each verb, the production probability of diphthongization was calculated by dividing the number of diphthongized responses by the total number of diphthongized + undiphthongized responses. For example, for the verb *lerrar*, 19 participants volunteered [ljérro] and 76 volunteered [lérro],[5] yielding a 20% production probability of diphthongization. (For additional details of the experimental design and results, see Albright, Andrade, and Hayes (2001)).

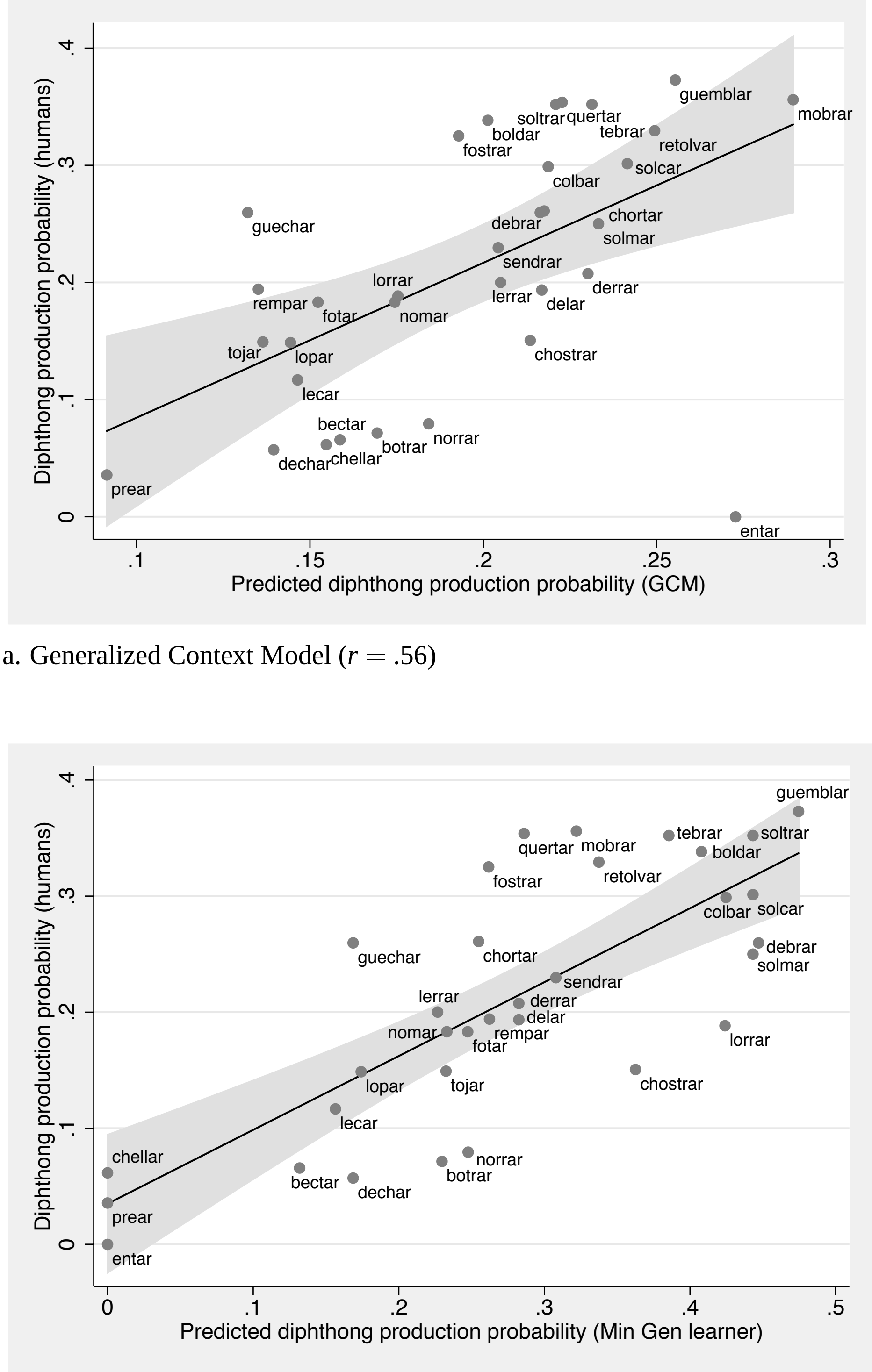
In order to test the models, predictions were obtained by training each model on a lexicon of Spanish. Two different data sets were tested: one that included all of the verbs in the LEXESP corpus containing stressable mid vowels (1,881 verbs total), and another that included just the subset of verbs that fall in the *-ar* inflectional class (1,669 of the total set). The choice of data set turns out to matter slightly for the results, with the GCM performing slightly better on the full set and the MGL performing slightly better on the smaller set. The differences were relatively small, however, and I simply report here the better result for each model (i.e., treating the choice of dataset as a parameter that can vary independently across models).

Figure 1 shows the overall ability of the two models to predict the probability of diphthongization on a verb-by-verb basis. We see that both models do reasonably well, though the MGL does somewhat better (*r*

= .77) than the GCM (*r* = .56). Most of this difference comes from the exceptionally poor performance of the GCM on a single outlier, however (*entar*); if this one item is excluded, the performance of the GCM is approximately as good as the MGL (*r* increases to .74).

So what do we conclude from this result? Clearly, neither model can be rejected outright based on raw performance. In fact, the predictions of the two models are also significantly correlated with one another (*r* = .53). This means that the models are not merely making equivalently good predictions—in fact, to a large extent they are making the very same predictons.[6] When the outputs of the two models are inspected, the reason is not hard to find: in very many cases, the two models pick out overlapping analogical sets. For example for the novel word *solmar*, the MGL found that the most confident applicable diphthongization rule was $\text{o} \rightarrow \text{wé} \,/\, \text{s} \;\_\_\_\; \text{l Y]}_{\textit{-ar}\text{ class}}$ (including such words as *solar* 'pave', *soltar* 'release', and *soldar* 'solder'). These same words figure prominently in the analogical set that the GCM employs; the five top contributors are *solar* (similarity .493), *soldar* (.417), *soltar* (.338), *cerrar* (.214), and *dormir* (.164). Similarly for the verb *lorrar*, the MGL used a rule $\text{o} \rightarrow \text{wé} \,/\, \begin{bmatrix} +\text{coronal} \\ +\text{continuant} \end{bmatrix} \;\_\_\_\; \begin{bmatrix} +\text{coronal} \\ +\text{voice} \end{bmatrix} \text{Y]}_{\textit{-ar}\text{ class}}$, supported by positive examples like *solar*, *sonar*, *soldar*, *rodar*, and *soltar*. Here too, the rule includes three of the GCM's five closest analogs: *errar* (.278), *cerrar* (.252), *solar* (.095), *rodar* (.094), and *soldar* (.085). The upshot is that although the GCM has access to variegated similarity—seen, for example, in the presence of analogs like *cerrar*—there is no guarantee that it is actually using it to a significant extent in any particular case. Thus, overall comparisons like the one in Figure 1 are unlikely to be illuminating about what mechanism speakers actually use to make analogical inferences.[7]

The examples in the preceding paragraph show that although in practice the role of variegated similarity is less than what is theoretically possible, the GCM does use it at least to a certain extent. What we need, then, is a way to focus specifically on the contribution of the variegated analogs, which the MGL cannot include as support for inferences. This requires a means of separating analogs that share a structured relation from those that do not. For a set like {*solar*, *soldar*, *soltar*, *cerrar*, *dormir*}, the intuitive division is between the first three, which share *#sol* (and the *-ar* inflectional class), as opposed to *cerrar* and *dormir*, which look like odd men out. Strictly speaking, however, it is not the case that these verbs completely lack structural properties with the remaining forms. In fact, all five verbs share the set of properties in (15):



a. Generalized Context Model ($r = .56$)

b. Minimal Generalization Learner ($r = .77$)

Figure 1: Predicted vs. observed production probability of diphthongization

(15) Structural commonality: *solar*, *soldar, soltar, cerrar*, and *dormir*

$$\#[-\text{sonorant}]\begin{bmatrix}+\text{syllabic}\\-\text{high}\\-\text{low}\end{bmatrix}\begin{bmatrix}+\text{consonantal}\\+\text{sonorant}\\-\text{nasal}\end{bmatrix}\text{Y}$$

The description in (15) expresses a structured similarity, but expanding the context to include *cerrar* and *dormir* comes at a price. The description is now so general that it includes not just these five verbs, but also many others—including, importantly, some that do not diphthongize. In other words, although the description in (15) unifies all of the members of the GCM's analogical set, it does not accurately or uniquely describe what sets them apart from the rest of the verbs in the language. A rule-based model like the MGL could state a rule that applies diphthongization in this context, but it would not be a useful rule since it has too many exceptions.

This suggests a refinement to how we isolate sets of structured analogs: they must not only have in common a set of shared properties, but those properties must also be reliably associated with class membership. For example, it is not enough to be able to state what *cerrar* has in common with *soldar* and *soltar*; the properties that they share must also distinguish these verbs from non-diphthongizing verbs. In order to separate structured from unstructured analogs, then, we need a hypothesis about what those distinguishing properties are. Not coincidentally, this is precisely what the MGL model is designed to identify. For example, as noted above, the MGL determines that the properties of *solmar* that are most reliably associated with diphthongization are the preceding /s/ and the following /l/, making *solar*, *soldar, soltar* the analogs that share the set of most relevant structural properties. It should be possible, therefore, to use the structural descriptions that the MGL selects to help identify when the GCM is making use of unstructured, or variegated similarity.

In order to quantify the contribution of non-structured analogs in the predictions of the GCM, I first ran the MGL, finding for each nonce form the set of properties that were found to be most reliably associated with diphthongization (i.e., the structural description of the best applicable rule that could derive a diphthongized output). I then ran the GCM, collecting the set of diphthongizing analogs. For each nonce verb, the analogical set was then separated into two groups: the structured analogs, which contained the best context identified by the MGL, and the variegated analogs, which fell outside this context. Examples for the novel verbs *solmar* and *lorrar* are given in (16).

(16) Separating structured vs. variegated analogs

a. *solmar*: best context $=$ [sol. . .]$_{\text{-}\textit{ar}\text{ class}}$)

Structured analogs		Unstructured analogs	
solar	.493	*serrar*	.214
soldar	.417	*dormir*	.164
soltar	.338	*sonar*	.157
		serner	.139
		socar	.126
		(and 235 others)	

b. *lorrar*: best context $= \; [\begin{bmatrix} +\text{coronal} \\ -\text{continuant} \end{bmatrix} \text{o} \begin{bmatrix} +\text{coronal} \\ +\text{voice} \end{bmatrix}]_{\text{-}\textit{ar}\text{ class}}$

Structured analogs		Unstructured analogs	
solar	.095	*errar*	.278
rodar	.094	*serrar*	.252
		soldar	.085
		forsar	.084
		(and 236 others)	

The contribution of variegated analogy was then defined as the summed similarity of the unstructured analogs divided by the summed similarity of all analogs (structured and unstructured). This ratio is taken as a measure of the extent to which the GCM is relying on variegated similarity for any particular nonce word.

We are now in a position to evaluate the usefulness of variegated similarity. If speakers make analogical inferences in a way that is blind to structure, then the MGL model should suffer in cases where variegated similarity is needed, since it is unable to make use of a crucial source of support. Conversely, if structure is critical to how speakers generalize, then the GCM should do worse the more it relies more on variegated similarity. The word-by-word performance of each model was tested by fitting the predictions of each model against the experimentally obtained human responses using a linear regression. For each word,

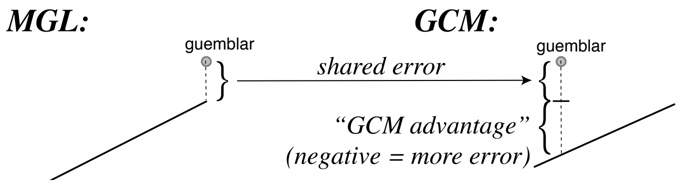


Figure 2: Calculation of "GCM advantage" score based on residuals

it was then determined how far off the model was, by subtracting the observed from the predicted values (i.e., calculating the residuals). The performance of the two models was collapsed into a single "GCM advantage" score by subtracting the GCM error from the MGL error for each word; this score is positive for a particular word if the MGL's prediction is less accurate than the GCM's, and negative if the GCM that is farther off. This comparison is illustrated in Figure 2. Finally, the GCM advantage scores were correlated against the contribution of variegated analogy, as defined in the preceding paragraph. If variegated analogs are important to speakers, we expect a positive correlation, since in cases where variegated similarity plays a larger role, the MGL should suffer more (positive GCM advantage). If speakers do not use variegated similarity, we expect a negative correlation, since the GCM's reliance on variegated analogs would encourage generalizations that humans do not make. In fact, when the correlation is calculated as described above, the result is weakly negative ($r = -.195$). Thus, we fail to find any support for the idea that variegated similarity is needed—and in fact, there is an indication that it may even be harmful.

The same result can also be seen another way, by calculating for each novel item the degree to which the GCM overestimated the goodness of each output. This amount will be positive if the GCM assigned too high a score (overpredicting the goodness of the output), and zero if the GCM is right on or under. The rational for restricting the analysis to *overpredictions* is the following: suppose that speakers do not notice variegated similarity, and that the GCM is incorrect to use it. If this is true, then access to variegated analogs should let the GCM (incorrectly) gather extra support for some outputs, leading to overestimation of their goodness. Therefore, the negative effects of variegated similarity should be seen most clearly in the GCM's overprediction errors. To test this, the GCM's overestimation scores were correlated against the relative contribution of variegated similarity, as defined above. The result here was a positive correlation between variegated similarity and overestimation ($r = .33$). This shows that the extra sources of support that the GCM has access to are not helpful in modeling speakers more accurately—in fact, they are deleterious, causing the model to overestimate the probability of diphthongization. Albright and Hayes (2003) also make a similar point about the GCM, using data from English past tense formation.

There is finally one last way in which structure can be seen to matter. If we examine the GCM predictions in Figure 1a, we see that the most blatant gaffe by far that the GCM makes is in overpredicting the probability of diphthongization in *entar*. This prediction is based on the support of diphthongizing analogs like *sentar* ‘seat’, *mentar* ‘mention’, *tentar* ‘touch’, *dentar* ‘teethe’, *ventar* ‘sniff’, and so on. All of these analogs have a preceding consonant, and in fact diiphthongization of initial vowels is overall quite rare in Spanish (particularly in the *-ar* class). The MGL is able to encode this fact by requiring that a consonant is a crucial part of the context when formulating rules. The GCM, on the other hand, has no way to encode this beyond the standard penalty for inserted or deleted a single segment in the process of calculating the optimal string alignment; therefore, it cannot categorically block analogy to similar consonant-initial words. This is yet another indication that speakers encode knowledge of patterns in terms of properties of elements that appear in particular positions—that is, in terms of linguistic structure.

2.5 Local summary

In this section, I have discussed a major restriction on what type of pattern can be generalized through analogy: it must be supported by sets of words that share a particular combination of properties in common, both with each other and with the target word. This may seem like an obvious or trivial restriction, and in fact many models simply assume it without argument. However, it is certainly not a logically necessary part of how analogy is formalized. Many examplar-based models, such as the GCM, do not obey this restriction. This allows them to capture a wider range of patterns, and thereby makes them them less constrained models. I have shown that the extra power afforded by unstructured comparisons does not help—and indeed, it seems to hurt by inflating the predicted goodness of certain generalizations. This confirms similar results shown previously for English by Albright and Hayes (2003).

Importantly, the restriction to structured comparisons is exactly what we would expect if speakers encode patterns using something like probabilistic context-sensitive rules, of the sort employed by the MGL. Of course, this is not the only model that imposes structure on its representations; similar restrictions are also found in feature-based models, such as TiMBL (Daelemans, Zavrel, Van der Sloot, and Van den

Bosch 2000) and AML (Skousen 1989).

3 Type vs. token frequency

Another possible restriction on analogical models concerns the way in which the support for competing patterns is evaluated. In principle, a pattern could be strengthened in at least two different ways: by occurring in a large number of different words (high type frequency), or by occurring in a smaller number of words that are used very commonly (high token frequency). In fact, it appears that the propensity to generalize morphophonological patterns to new forms depends primarily on type frequency, and not on token frequency. This restriction has been noted numerous times in the literature; see Baayen and Lieber (1991) for English derivational suffixes Bybee (1995) for French conjugation classes, German past participles, and others, Albright (2002) for Italian conjugation classes, Albright and Hayes (2003, p. 133) for English past tenses, Ernestus and Baayen (2003, p. 29) for stem-final voicing in Dutch, Hay, Pierrehumbert, and Beckman (2004) for medial consonant clusters in English, and additional references in Bybee (1995)). In this section, I provide further evidence for this conclusion, and suggest that it favors a model in which patterns are abstracted from individual words and encoded in some form that is separate from the lexicon (such as a grammar).

The formal definition of similarity in the Generalized Context Model ((6) above) is compatible with counting based either on type or token frequency, since “members of a class” could be taken to mean either types or individual tokens. In practice, however, the most natural interpretations of the model would lead us to expect a role for token frequency. If we assume, as is often done, that the GCM operates over exemplar representations (Johnson 1997; Pierrehumbert 2001), then every single token should contribute a measure of support to the strength of the pattern. Furthermore, even if we assume that the GCM operates over a more schematic lexicon that abstracts away from individual exemplars, there is ample evidence from on-line recognition and processing tasks that words with higher token frequency are accessed more readily than low frequency words. Therefore, even if the GCM counts over a lexicon distinct word types, it seems likely that token frequency effects would emerge simply because of the way the lexicon is accessed. Stated more generally, the premise of the GCM is that generalization is carried out by consulting the lexicon directly, and token frequency effects are characteristic—perhaps even diagnostic—of lexical access. It is important to note that the GCM is also very sensitive to type frequency, since each type contributes at least one token to the summed support for a particular class.

In principle, the Minimal Generalization Learner could also evaluate rules using types or tokens, but the rules it discovers are most naturally interpreted in terms of types. The comparisons that it carries out to abstract away from individual lexical items ((10) above) require just a single instance of each word, and nothing more can be learned from further tokens of previously seen data. In a system in which additional tokens are gratuitous, it would perhaps be a surprising design feature if token frequency played a crucial role in how rules are evaluated. In fact, calculating the confidence of rules according to their token frequency would require extra work in this model, since repeated tokens of the same lexical item could otherwise be disregarded as uninformative. This also relates to the more general hypothesis that grammars are intrinsically about kinds of words, rather than about particular instances of their use. Therefore, even if it is not strictly speaking required by the formalism, a rule-based account of analogy is most naturally limited to the influence type frequency.

Spanish diphthongization provides a direct test of the relative importance of type vs. token frequency, since although diphthongization is a minority pattern in the Spanish lexicon, affecting only a relatively small number of mid-vowel verbs (lowish type frequency), the verbs that undergo it tend to be among the most frequent verbs in the language (high token frequency). There is abundant prima facie evidence that the high token frequency of diphthongization does not make it a strong pattern: synchronically it is relatively unproductive in experimental settings (Bybee and Pardo 1981; Albright, Andrade, and Hayes 2001), and diachronically verbs tend to lose diphthongization alternations (Penny 2002; Morris 2005). Furthermore, overregularization errors among children acquiring Spanish consistently result in omitting diphthongization (Clahsen, Aveledo, and Roca 2002), even though diphthongizing tokens constitute a large portion—perhaps even the majority—of childrens' experience.

In order to test the influence of token frequency more systematically, I ran both the GCM and the MGL with and without taking token frequency into account. Specifically, a weighting term was introduced in the GCM, so the contribution of each analog was defined not only in proportion to its similarity, but also in proportion its (log) token frequency. A weighting term was also introduced into the MGL, such that the contribution of each word to the hits and/or scope of a rule was weighted according to its log token frequency. The result was that both models did slightly worse when token frequency was taken into account, as shown in (17).

(17) A negative effect of token frequency (Pearson's *r*)

	Type frequency alone	Weighted by (log) token frequency
GCM	.743	.730
MGL	.767	.742

We see that the overall effect of token frequency weighting is quite small. The reason for this is that most words in the average corpus (and presumably also the average lexicon) have very low frequency ("Zipf's Law"). As a result, weighting by token frequency influences just a small number of high frequency words. Therefore, weighting by token frequency has relatively little effect, unless the target word happens to be very similar to an existing high frequency word. It should be noted that these particular experimental items were not constructed for the purpose of dissociating type and token frequency, and ultimately the fairest test would be based on items that diverge more in their predictions. Nonetheless, the trend is clear across both models: to the extent that token frequency makes a difference, it is harmful in modeling speaker intuitions about the strength of the diphthongization pattern.

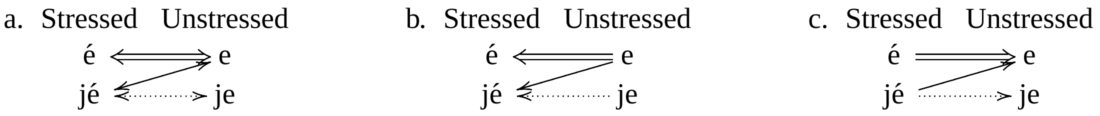
Like variegated similarity, high token frequency is a type of information that speakers could logically make use of in deciding whether or not to generalize a pattern to novel items. The fact that they apparently do not do so requires a formal model that is similarly restricted. As noted above, it is certainly possible to construct exemplar models that ignore token frequency; the amount and nature of frequency weighting is an independent parameter in the GCM that can be turned off completely, and Bybee (1995) explicitly defines schema strength in terms of type frequency. Conceptually, however, part of the appeal of exemplar models is that they rely on no special mechanisms except activating memory traces—a mechanism that intrinsically leads to token frequency effects (Bybee 2006). Insensitivity to token frequency follows quite naturally from a grammar of rules, however, since rules encode information that has been abstracted away from the particular exemplars that led to their creation. A rule-based account of analogy therefore involves no particular expectation that token frequency should play a role, and indeed is naturally restricted not to have access to information about token frequency.

4 The directionality of analogical inference

In the preceding sections, we have seen that an adequate model of analogical inference must be able to identify properties that are consistently associated with membership in a particular class, and must ensure that the association holds for sufficiently many different word types. Models that can find support for inferences in other ways, such as unstructured similarity or high token frequency, end up overestimating the goodness of many outcomes. A model without these abilities is more constrained, and has the advantage that it can more narrowly predict which analogical inferences speakers actually make. In this section I discuss one final restriction, concerning the direction of analogical inference.

Logically, statements about the relation between one form and another could be made in either direction. For example, statements about the correspondence of stressed and unstressed root allomorphs could relate either form to the other, symmetrically or asymmetrically, as in (18). This means that in principle, analogical inferences could proceed in multiple directions, both from stressless to stressed (e.g., *rentár*:*rénta* :: *sentár*:****sénta***) and stressed to stressless (e.g., *siénta*:*sentár* :: *oriénta*:****orentár***).

(18) Some logically possible directions of influence (solid and doubled lines represent progressively greater pattern strength)



What we observe, however, is a striking restriction: both in historical change (Penny 2002; Morris 2005) and child errors (Clahsen, Aveledo, and Roca 2002), there is an overwhelming (or even exclusive) tendency for analogical rebuilding of stressed forms (i.e., *rentár*:*rénta* :: *sentár*:****sénta***), consistent with (18b).[8] A typical example from the Spanish portion of CHILDES is given in (19).

(19) Overgeneralization of stressed mid vowels (Jorge, age 6;1)

y	estonces	***volo**	a	la	pastelería
and	then (=*entonces*)	fly-1sg (=***vuelo***)	to	the	pastry shop

'. . . and then I fly to the pastry shop'

Remarkably, the converse error (e.g., infinitive **vuelar* instead of *volar*) never occurs, and children also apparently never substitute mid vowels for non-alternating diphthongs (e.g., *el *frecónta* ‘he frequents’ instead of *frecuénta*). Similarly asymmetric error patterns have also been observed for Greek (Kazazis 1969), German (Clahsen, Aveledo, and Roca 2002) and Korean (Kang 2006), and appear to be the norm among children acquiring languages with morphophonological alternations. An explanatory model of analogy must be able to capture and ideally even predict such asymmetries.

Characterizing the direction of analogy has been a longstanding preoccupation in the historical linguistics literature, and numerous tendencies have been observed (Kuryłowicz 1947; Mańczak 1980; Bybee 1985, and many others). The Spanish case seems atypical in several respects. It has sometimes been claimed that more frequent paradigm members are more influential (Mańczak 1980; Bybee 1985). In Spanish, the most frequent paradigm members (3sg, 1sg, 2sg) are all stressed, which should favor a stressed → stressless direction of influence. What we observe, however, is that the more frequent stressed forms are rebuilt on the basis of the less frequent stressless forms, counter to the more usual trend. Furthermore, it is often the case that the most influential forms are also less marked (in some intuitive sense of morphosyntactic markedness). What we see in Spanish, however, is that the 3sg, which is almost universally agreed to be the least marked combination of person and number features, is rebuilt on the basis of non-singular, non-3rd person forms. Furthermore, diphthongs appear in the majority of present tense indicative forms (the 1sg, 2sg, 3sg, and 3pl = 4 out of 6), yet reanalysis is done on the basis of the minority stressless forms. In short, the direction of influence that prevails in Spanish does not appear to follow from any general principle of frequency or markedness.

Albright (2002) proposes that speakers generalize in some directions and not others because of a restriction on how paradigm structure is encoded. In particular, it is proposed that paradigms have an intrinsically asymmetrical organization in which certain forms are designated as “basic” and the remaining forms are derived from them by grammatical rules. For example, the error data suggests that in Spanish, a stressless form of the root (as found in the infinitive, 1pl, or 2pl) is taken as basic, and stressed forms are predicted—sometimes incorrectly—on the basis of a stressless form. The challenge is to understand why Spanish speakers choose this particular direction, and why paradigm organization may differ from language to language.

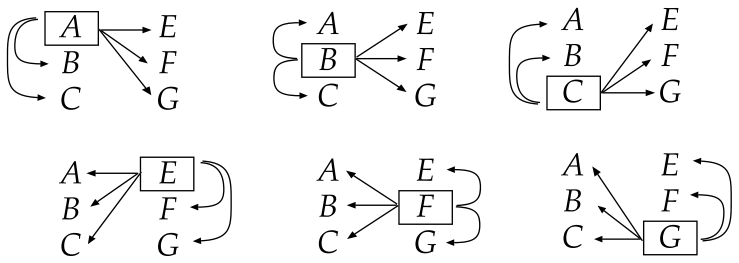


Figure 3: Candidate grammars, using asymmetrical mappings from a single base

One principle of paradigm organization, explored also by Stump and Finkel (this volume), is PREDICTABILITY: a form is basic (≈ a PRINCIPAL PART) if it contains enough information to predict other forms in the paradigm. As Stump and Finkel point out, there are many ways in which paradigms could be organized around predictive forms, depending on whether how many basic forms we are allowed to refer to, whether paradigm structure may differ from class to class, and so on. Many paradigm-based theories of morphology designate specific forms as "reference forms" in one way or another, and use these forms as the basis of computation for the remaining forms in the paradigm (Wurzel 1989; Stump 2001; Blevins 2006). Albright (2002) adopts a particularly restrictive hypothesis: paradigm structure is the same (static) across all lexical items, and each form in the paradigm is based on just one other base form. The task of the learner is to find the base forms that permit the most accurate mappings, while still obeying this restriction.

The base identification algorithm, in brief, works as follows: the learner starts with a small batch of initial input data, consisting of paradigmatically related forms (1sg, 2sg, 3sg, etc.). Each one of these forms is considered as a potential base form, and the minimal generalization learner is used to find sets of rules that derive the remaining forms in the grammar. The result is a set of competing organizations, shown in Figure 3. In the usual case, at least some parts of the paradigm suffer from phonological or morphological neutralizations, with the result that not every form is equally successful at predicting the remainder of the paradigm. In these cases, some of the competing grammars will be less certain or accurate than others. The learner compares the candidate organizations to determine which form is associated with the most accurate rules, and this is chosen as the base for the remainder of the paradigm. This process may also be run recursively among the derived forms, to establish additional intermediate bases. (See Albright 2002 for details.)

When this procedure is run on an input of Spanish present tense verb paradigms, the organization in

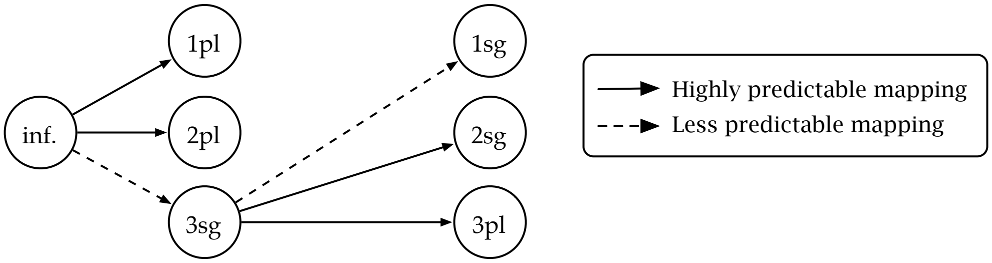


Figure 4: Predicted organization of Spanish present tense paradigms

Figure 4 results. Crucially, due to the restriction that each form be based on exactly one other base form, the model allows only five possible directions of inference (out of $5 \times 6 = 30$ logical possible pairwise relations). Some of these relations, such as infinitive→1pl. or 3sg.→2sg., are virtually 100% predictable, and leave no room for error. The greatest opportunities for analogical errors involve the mapping from stressless to stressed forms (here, infinitive → 3sg.), and to the 1sg. in particular. In fact, both of these mappings correspond to attested child errors:

(20) Stem errors among children acquiring Spanish (Clahsen, Aveledo, and Roca 2002)

a. Stressed stem replaced by stressless stem:

**volo* for *vuelo* 'fly-1sg', **juga* for *juega* 'play-3sg', **tene* for *tiene* 'have-3sg', **teno* for *tengo* 'have-1sg'

b. Irregular 1sg replaced by stem from 3sg:

**tieno* for *tengo* 'have-1sg'; **sabo* for *sé* 'I know'; **conozo* for *conozco* 'I know'; **parezo* for *parezco* 'I appear'; **salo* for *salgo* 'I leave'; **oyo* for *oigo* 'I hear'

Although this analysis is somewhat skeletal and leaves many broader questions about paradigm structure unanswered,[9] it highlights some of the virtues of a rule-based model of analogy. In particular, grammatical formalisms place strong restrictions on possible analogical inferences by dictating which forms may be effected, which patterns can be extended, and so on. Naturally, the strength and nature of these restrictions may vary considerably depending on the formalism; I have argued here in favor of a grammar of probabilistic context-sensitive rules that asymmetrically relate forms in the paradigm, but other formalisms are possible. The advantage of such a restrictive model is that it makes very specific and testable predictions about possible errors, and presumably also eventual historical changes. In the cases examined, these predictions appear to be substantially correct.

5 Conclusion

The results in the preceding sections have a common theme: in each case, the data of Spanish contains patterns that might logically lead to analogical inferences, yet speakers appear not to generalize them to novel or unknown items. I have argued that this reveals fundamental restrictions on how speakers learn to encode linguistic knowledge, which make these patterns either inaccessible or unimpressive. Furthermore, I have shown that a model based on probabilistic context-sensitive rules is well suited to capturing these restrictions. First, it limits the type of similarity relations that are relevant in supporting analogy: they must be "structured" in the sense that supporting analogs must all share a set of properties that are reliably correlated with class membership. As shown in section 2, speakers, too, appear to obey this restriction, and models that lack such structure overpredict the goodness of many logically possible inferences. In addition, attributing analogy to a grammar of rules leads us to expect that generalizations should be based on high type frequency of similar words, and that token frequency should be irrelevant; in section 3, we saw that this, too, appears to be correct. Finally, rewrite rules are an intrinsically directional formalism ($A \rightarrow B$), corresponding to the idea that inference proceeds in some directions but not others. In section 4, I argued that a model of paradigm structure based on predictability relations between related forms can predict which directions speakers actually choose, in a way that appears to line up well with data from child errors and historical change. In each case, the payoff of the more restrictive formalism is clear: it provides an account for why some errors occur and some do not, providing a more explanatory model of how speakers carry out analogy in morphophonological systems.

The examples discussed here are also intended to highlight some virtues of computationally implemented models of analogy. At the most basic level, the models facilitate a quantitative assessment of the relative contribution of different types of analogical reasoning, by allowing us to compare directly the predictions of models with and without a particular capacity. Such comparisons are potentially quite important in an area where it is easy to posit many potentially relevant factors (high token frequency, semantic effects, phonetic factors, etc.), but difficult to establish their explanatory value. Equally important, though, is the role that modeling may play in shaping and refining theoretical distinctions. An example of this was seen in section 2.4, in which comparison of the two models required a more careful definition of the concept of structured similarity, and testing the distinction was only possible by interpreting one model with respect to the other. We are only beginning to develop the analytical tools needed to construct theoretical arguments from such modeling results. I hope to have shown, however, that computational modeling can play a role not only in testing, but also in developing theories of what constitutes a possible analogy.

Notes

[1] Recent decades have seen a wealth of frameworks for modeling analogical inference and decision making more generally; see especially Gentner, Holyoak, and Kokinov (2001) and Chater, Tenenbaum, and Yuille (2006).

[2] Or, more precisely, they share only very general properties which do not distinguish them from other verbs in the language, such as having a liquid, a stressable mid vowel, and so on.

[3] Since the implemented model uses linear (flat) phonological representations, stress is encoded here as a feature of the stressed vowel, rather than as a property of the syllabic context.

[4] Ultimately, this is too strong an assumption, since contexts are sometimes non-local. For an attempt to extend this system to find non-local contexts, and discussion of some of the issues involved, see Albright and Hayes (2006).

[5] One additional subject volunteered an unexpected and idiosyncratic change for this verb; this response was excluded.

[6] This was confirmed by a stepwise multiple regression analysis, in which the MGL predictions were entered first with a high degree of significance ($p < .0001$), and the GCM predictions were unable to make any additional significant contribution.

[7] The reason that the GCM tends to stick to such structurally interpretable analogical sets appears to be due to the fact that diphthongizing verbs in Spanish themselves happen to fall into such clusters. The explanation for this may be partly phonological, since phonotactic restrictions on stem-final consonant combinations would restrict the set of possibilities in this position, and make it easier for commonalities to emerge. There may also be a historical component: suppose the structured model of analogy is the correct one, and structure-guided inferences have been shaping Spanish over the centuries. In this case, we would expect verbs to retain diphthongization most readily if they fall into structurally definable gangs, creating structure in the lexicon of Spanish. If this were true, then the GCM could do good job of capturing the modern language, but would be unable to explain how the language came to be this way. If, on the other hand, the GCM model were correct, we would expect diphthongizing verbs to be retained on the strength of

variegated similarity, and the set of existing diphthongizing verbs could consist of variegated analogical sets which the structured model would be unable to locate. A full diachronic analysis of verb-by-verb changes in diphthongization is left as a matter for future research.

[8] Rebuilding stressless forms to include diphthongs has been reported in some dialects of Spanish (Judeo-Spanish, New Mexico Spanish). This data should be treated with care, however, since the morphology of these dialects also differs in more radical ways from literary Spanish. A similar effect is also reported in the experimental results of Bybee and Pardo (1981), but my preliminary attempts to replicate this finding have so far been unsuccessful.

[9] In particular, it is natural to wonder how such a restrictive model could cope with systems that involve significantly more ambiguity—i.e., systems that motivate multiple principal parts in Stump and Finkel's terms. It is important to keep in mind that nothing in the current model precludes the possibility that at a given point in time, languages may exhibit patterns that may be characterized as symmetrical predictability relations. A prediction of the asymmetrical model, however, is that learners will learn implications in just one direction, and that analogical generalizations should therefore go primarily in one direction. One type of data that is often telling in this regard is the relative size and frequency of the inflectional classes involved. Frequently classes that can be distinguished only in derived (non-basic) forms are small and consist of words with high token frequency, which may be correlated with their status as memorized exceptions rather than as grammatically principled forms.

References

Albright, A. (2002). Islands of reliability for regular morphology: Evidence from Italian. *Language 78*(4), 684–709.

Albright, A., A. E. Andrade, and B. Hayes (2001). Segmental environments of spanish diphthongization. In A. Albright and T. Cho (Eds.), *UCLA Working Papers in Linguistics, Number 7: Papers in Phonology 5*, pp. 117–151. http://www.linguistics.ucla.edu/people/hayes/Segenvspandiph/SegEnvSpanDiph.pdf.

Albright, A. and B. Hayes (2002). Modeling English past tense intuitions with minimal generalization. *SIGPHON 6: Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, 58–69.

Albright, A. and B. Hayes (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition 90*, 119–161.

Albright, A. and B. Hayes (2006). Modeling productivity with the Gradual Learning Algorithm: The problem of accidentally exceptionless generalizations. In G. Fanselow, C. Féry, R. Vogel, and M. Schlesewsky (Eds.), *Gradience in Grammar: Generative Perspectives*, pp. 185–204. Oxford University Press.

Baayen, R. H. and R. Lieber (1991). Productivity and English derivation: A corpus-based study. *Linguistics 29*, 801–843.

Bailey, T. and U. Hahn (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language 44*, 568–591.

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics 42*, 531–573.

Broe, M. (1993). *Specification Theory: The Treatment of Redundancy in Generative Phonology*. Ph. D. thesis, University of Edinburgh.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins Publishing Company.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes 10*(5), 425–255.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language 82*, 711–733.

Bybee, J. L. and E. Pardo (1981). On Lexical and morphological conditioning of alternations: a nonce-probe experiment with Spanish verbs. *Linguistics 19*, 937–968.

Chater, N., J. B. Tenenbaum, and A. Yuille (2006). Special issue: Probabilistic models of cognition. *Trends in Cognitive Sciences 10*(7).

Clahsen, H., F. Aveledo, and I. Roca (2002). The development of regular and irregular verb inflection in Spanish child language. *Journal of Child Language 29*, 591–622.

Daelemans, W., J. Zavrel, K. Van der Sloot, and A. Van den Bosch (2000). TiMBL: Tilburg memory based learner reference guide 3.0. Report 00-01, Computational Linguistics Tilburg University.

Daugherty, K. and M. Seidenberg (1994). Beyond Rules and Exceptions: A Connectionist Approach to Inflectional Morphology. In S. D. Lima, R. L. Corrigan, and G. K. Iverson (Eds.), *The Reality of Linguistic Rules*. Amsterdam: J. Benjamins.

Ernestus, M. and R. H. Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language 79*, 5–38.

Frisch, S. (1996). *Similarity and Frequency in Phonology*. Ph. D. thesis, Northwestern University.

Frisch, S. A., J. B. Pierrehumbert, and M. B. Broe (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory 22*(1), 179–228.

Gentner, D., K. J. Holyoak, and B. N. Kokinov (Eds.) (2001). *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.

Gildea, D. and D. Jurafsky (1996). Learning bias and phonological-rule induction. *Computational Linguistics 22*(4), 497–530.

Hahn, U., N. Chater, and L. Richardson (2003). Similarity as transformation. *Cognition 87*, 1–32.

Hare, M. and J. L. Elman (1995). Learning and morphological change. *Cognition 56*, 61–98.

Hay, J., J. Pierrehumbert, and M. Beckman (2004). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden, and R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press.

Johnson, C. D. (1972). *Formal Aspects of Phonological Description*. The Hague: Mouton.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson and J. W. Mullennix (Eds.), *Talker Variability in Speech Processing*, pp. 145–165. San Diego: Academic Press.

Kang, Y. (2006). Neutralization and variations in Korean verbal paradigms. In *Harvard Studies in Korean Linguistics XI*, pp. 183–196. Hanshin Publishing Company.

Kaplan, R. M. and M. Kay (1994). Regular models of phonological rule systems. *Computational Linguististics 20*(3), 331–378.

Kazazis, K. (1969). Possible evidence for (near-)underlying forms in the speech of a child. *Chicago Linguistics Society 5*, 382–386.

Kruskal, J. B. (1983). An overview of sequence comparison. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pp. 1–44. Reading, MA: Addison-Wesley.

Kuryłowicz, J. (1947). The nature of the so-called analogical processes. *Diachronica 12*(1), 113–145. (trans. Margaret Winters, 1995, Diachronica 12:1.113-145).

MacWhinney, B. and J. Leinbach (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition 40*, 121–157.

Mańczak, W. (1980). Laws of analogy. In J. Fisiak (Ed.), *Historical morphology*, pp. 283–288. The Hague: Mouton.

Morpurgo-Davies, A. (1978). Analogy, segmentation and the early Neogrammarians. *Transactions of the Philological Society*, 36–60.

Morris, R. E. (2005). Attraction to the unmarked in Old Spanish leveling. In D. Eddington (Ed.), *Selected Proceedings of the 7th Hispanic Linguistics Symposium*, pp. 180–191. Somerville, MA: Cascadilla Proceedings Project. http://www.lingref.com/cpp/hls/7/paper1097.pdf.

Nakisa, R. C., K. Plunkett, and U. Hahn (1997). A Cross-Linguistic Comparison of Single and Dual-Route Models of Inflectional Morphology. In P. Broeder and J. Murre (Eds.), *Cognitive Models of Language Acquisition*. Cambridge, MA: MIT Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal* *of Experimental Psychology: General 115*, 39–57. http://www.cogs.indiana.edu/nosofsky/pubs/1986_rmn_jep-g_attention.pdf.

Nosofsky, R. M. (1990). Relations between exemplar similarity and likelihood models of classification. *Journal of Mathematical Psychology 34*, 393–418.

Penny, R. (2002). *A History of the Spanish Language* (2nd ed.). Cambridge University Press.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*, pp. 137–158. John Benjamins.

Rumelhart, D. and J. McClelland (1987). Learning the past tenses of English verbs. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Skousen, R. (1989). *Analogical Modeling of Language*. Dordrecht: Kluwer Academic Publishers.

Stump, G. T. (2001). *Inflectional morphology: a theory of paradigm structure*, Volume 93. Cambridge: Cambridge University Press.

Tversky, A. (1977). Features of similarity. *Psychological Review 84*, 327–352.

Wurzel, W. U. (1989). *Natural Morphology and Naturalness*. Kluwer Academic Publishers.