Segmental Environments of Spanish Diphthongization

Adam Albright
Argelia Andrade
Bruce Hayes

UCLA

July 21, 2000

Abstract

Spanish diphthongization is a well-known example of an exceptionful phonological alternation. Although many forms do exhibit an alternation (e.g. [s**e**ntámos] ~ [s**jé**nto] 'we/I sit', [k**o**ntámos] ~ [k**wé**nto] 'we/I count'), many others do not (e.g. [r**e**ntámos] ~ [r**é**nto] 'we/I rent', [m**o**ntámos] ~ [m**ó**nto] 'we/I mount'). Previous accounts of the alternation have largely accepted this unpredictability at face value, focusing on setting up appropriate lexical representations to distinguish alternating from non-alternating roots. Our interest is in whether Spanish speakers go beyond this, internalizing detailed knowledge of the ways in which diphthongization is conditioned by segmental environments.

We employed a machine-implemented algorithm to search a database of 1698 mid-vowel verbs. The algorithm yielded a large stochastic grammar, specifying the degree to which diphthongization is favored or disfavored in particular segmental contexts. We used this grammar to make predictions about the well-formedness of diphthongization in novel roots. The predictions were then checked in a nonce probe experiment with 96 native speaker consultants. We found that the consultants' intuitions (both in volunteered forms and in acceptability ratings) were significantly correlated with the predictions of the algorithmically learned grammar. Our conclusion is that Spanish speakers can indeed use detailed segmental environments to help predict diphthongization. We discuss this conclusion in light of various models of linguistic irregularity.

Segmental Environments of Spanish Diphthongization

1 Introduction

The diphthongization alternations of Spanish have been the subject of extended study. In inflectional and derivational paradigms, many instances of [e] and [o] occurring in stressless position correspond to [je] and [we] in stressed position, as in (1):

(1)	[s**e**ntámos]	~	[s**jé**nto]	'we/I sit'
	[t**e**ndémos]	~	[t**jé**ndo]	'we/I stretch'
	[p**o**démos]	~	[p**wé**do]	'we/I can'
	[k**o**ntámos]	~	[k**wé**nto]	'we/I count'

However, there are also large numbers of [e] and [o] that do not alternate; that is, they appear in stressed position with [e] and [o], as in (2):

(2)	[rentámos]	~	[rénto]	'we/I rent'
	[b**e**ndémos]	~	[b**é**ndo]	'we/I sell'
	[p**o**dámos]	~	[p**ó**do]	'we/I prune'
	[m**o**ntámos]	~	[m**ó**nto]	'we/I mount'

Thus, given a paradigmatic form with unstressed [e] or [o], there is no general way to predict whether its stressed correspondent will be [jé, wé] or [é, ó].[1] Note further that nonalternating [je] and [we] also occur, as in [al**je**námos] ~ [al**jé**no] 'we/I alienate', [frek**we**ntámos] ~ [frek**wé**nto] 'we/I frequent', so the alternation is unpredictable in both directions.

Given this basic unpredictability, analyses of the phenomenon have centered on mechanisms to distinguish alternating from non-alternating roots. In some accounts (Harris 1969, 1977, 1978, 1985, Schuldberg 1984, García-Bellido 1986, Carreira 1991), alternating and non-alternating mid vowels have distinct underlying representations, involving diacritics, abstract vowels, or linked vs. floating X slots. Hooper (1976) advocates representations in which alternating roots have a listed choice of vocalism: thus /k {o, we} nt/ depicts a root that alternates as /kont/ and /kwent/.

Less attention has been devoted to the question of how Spanish speakers handle words whose diphthongization properties are not known. We aim here to show that Spanish speakers know more about the diphthongization pattern than just the behavior of existing verbs. In particular, they can assess the likelihood of a novel root to undergo diphthongization, based on

[1] The bifurcation of alternating and non-alternating [e] and [o] is well understood from a historical point of view (e.g. Penny 1991): alternating [e, o] descend from Proto-Romance [ɛ, ɔ], which diphthongized in stressed position; whereas non-alternating [e, o] descend from Proto-Romance [e, o], which did not diphthongize. The merger of earlier [ɛ, ɔ] with [e, o] in stressless position created the present-day pattern.

its phonological shape. To the extent that this is true, analyses based solely on lexical representation do not capture the full linguistic knowledge of Spanish speakers.

Our general hypothesis is that in an attempt to make sense of the allomorph distributions that confront them, children comb through the data, looking for generalizations about phonological environments. When the data don’t pattern cleanly, the result is a rather messy set of conflicting learned generalizations. We further hypothesize that tacit knowledge of these generalizations persists into adulthood and can be detected experimentally.

This hypothesis already has considerable support from experimental work in the domain of *morphological* irregularity. To give just two examples, Zubin and Köpcke (1981), in their experimental study of German gender, show that while gender is not completely predictable, speakers do learn a large set of generalizations that help them to predict it. These generalizations are based on phonological shape, semantics, and other factors. Albright’s (1998) experimental work addresses the predictability of Italian verb conjugation classes. As with German gender, the conjugation class to which an Italian verb belongs is not generally predictable, but it appears that speakers learn a large set of generalizations, based on segmental environment, that help them to predict conjugation class. For further cases and literature review, see Bybee (1995).

The present case involves phonological, not morphological irregularity. If children respond to irregular phonology by conducting a search for phonological environments, then (assuming that this knowledge persists) it should be possible to show that adult speakers of Spanish are tacitly aware of these environments.

Our approach is as follows. We employ a machine-implemented algorithm to carry out inductive learning on a large data set of Spanish verb forms. The algorithm learns a detailed stochastic grammar, which projects the form of the stressed allomorph of a verb root given the unstressed allomorph. The grammar also provides well-formedness “intuitions” concerning how novel roots should be inflected. These synthetic intuitions are checked against intuitions obtained from real Spanish speakers in a nonce-probe task, or “wug test” (Berko 1958). To the degree that the intuitions match, we have evidence that humans have a capacity similar to the algorithm’s for noticing detailed environments.

The remainder of this article is organized as follows. §2 describes the learning algorithm, the data set that was fed to it, and the grammar it learned. §3 describes a wug-testing experiment designed to test the predictions of the learning algorithm. In §4, we offer some interpretation of what we found.

2 Modeling The Spanish Data With A Learning Algorithm

The machine-implemented algorithm that we used for discovering diphthongization environments is described in detail in Albright and Hayes (1998). It carries out a comprehensive search of the data, the rationale being that it is feasible to explore a large number of hypotheses simultaneously, so long as one includes a system of evaluation that permits the system to retain good hypotheses and discard bad ones.

2.1 Discovery of Contexts

The learning algorithm takes as its input pairs of forms that stand in a particular morphological relationship; in this case, the stressless allomorph of a verb root and the corresponding stressed allomorph. The method it uses for exploring segmental environments is to proceed bottom up from the lexicon. This pursues an idea of Pinker and Prince (1988, 134), which we refer to here as Minimal Generalization. The starting point of Minimal Generalization is to consider each pair of related forms as a (highly ungeneral) rule. Thus the pair in (3):

(3) [tembl] ~ [tjémbl] ‘tremble’

is construed as the rule in (4):

(4) e → jé / [t ___ mbl]

We refer to such rules as “word-specific rules.”

Further rules are built up from the word-specific rules by a process of generalization. Every newly created word-specific rule is compared with every rule already present in the system. Generalization occurs when two rules have the same structural change. The structural descriptions of the two rules are compared, and factored into material that both forms share and material that is unique to just one form. Thus, for instance, if the next data pair given the algorithm is [desmembr] ~ [desmjémbr] ‘dismember’, the comparison will proceed as follows:

(5)		***change***		***residue***	***shared segments***	***change location***	***shared segments***	***residue***
Form 1: (*tembl ~ tjémbl*)		e → jé	/	t		____	mb	l
Form 2: (*desmembr ~ desmjémbr*)	+	e → jé	/	desm		____	mb	r
	=	e → jé	/	X		____	mb	Y

In forming the factorization, the strings labeled “shared segments” are defined as the maximal identical strings that immediately precede and follow the structural change. The residues are the material not shared by the two forms.[2] The generalization process yields a rule in which shared material is retained, and residues are replaced by variables, in this case, e → jé before /mb/ clusters. The process is iterated, with each new form in the learning set compared with all rules that have been hypothesized thus far.

[2] It will be noted that the right-side residues in (5), /l/ and /r/, form a natural class, which could be characterized by the use of features. The algorithm in fact has this capacity; however, we have found that for the particular case of Spanish diphthongization, the use of features neither helps nor hurts in the task of modeling human judgments. Therefore, for convenience we have used a purely segmental approach.

An important aspect of Minimal Generalization is that, although it posits the most detailed possible rule at any given stage of generalization, it is nevertheless capable of learning very general structural descriptions. This happens when the same structural change occurs in a heterogeneous set of environments. As the algorithm is iterated, the differing environments cancel each other out and are replaced by variables, in a series of ever more general contexts. Thus, after exposure to a sufficient variety of diphthongizing pairs, Minimal Generalization ultimately hypothesizes a version of diphthongization constrained only by the location of stress (which we assume to be assigned by a separate mechanism):

$$\text{(6) a. e} \rightarrow \text{je} \; / \; [\text{ X } \begin{bmatrix} \underline{\qquad} \\ \text{+stress} \end{bmatrix} \text{ Y }]$$

$$\text{b. o} \rightarrow \text{we} \; / \; [\text{ X } \begin{bmatrix} \underline{\qquad} \\ \text{+stress} \end{bmatrix} \text{ Y }]$$

2.2 Evaluation of Contexts

Merely discovering a large set of possible diphthongization contexts is of little use in itself. To match native speaker intuitions we need a measure of productivity, specifying the degree of confidence with which diphthongization can be applied in any given context. We do this by computing a reliability score for each rule.

Reliability is computed in the following way. For each rule, we define the **scope** as the number of forms that meet its structural description. We define **hits** as the number of forms for which a rule can apply and derives the correct output. Both scope and hits are measured by count of types, not tokens. The **raw reliability** of a rule is defined as *hits/scope.* For example, the environment / [X ___ rr Y] (i.e., before trilled *r*) predicts diphthongization of /e/ correctly in 11 out of 11 cases in our learning corpus,[3] for a raw reliability of 1.

Raw reliability predicts productivity reasonably well when there are large numbers of forms covered by a rule. However, many rules have structural descriptions that are so specific that only a small number of forms fit them. In these cases, an adjustment must be made. The need for this can be illustrated by an example. If a particular rule R matches five roots and works for all five, that is not the same as if another rule R′ matches 1000 roots and works for all 1000. Although both have a raw reliability of 1, we intuitively give greater credence when testimony is more abundant.

In our algorithm we therefore (following Mikheev 1997) use an **adjusted reliability**, which is defined as the 75% lower confidence limit on raw reliability. Using adjusted reliability penalizes rules that are poorly instantiated in the learning set. For example, a rule that works for 11/11 forms (e.g. e → jé / [X ___ rr Y]) has an adjusted reliability of .916, whereas a rule that works for 1000/1000 forms has an adjusted reliability of .999. Albright (2000) evaluates this and other methods of calculating reliability using experimental data from Italian and English.

[3] The eleven are: *aferrar* ‘grasp’, *aserrar* ‘saw’, *aterrar* ‘terrify’, *cerrar* ‘close’, *desenterrar* ‘disinter’, *desterrar* ‘banish’, *encerrar* ‘lock’, *enterrar* ‘bury’, *errar* ‘miss’, *serrar* ‘saw’, and *soterrar* ‘bury’.

Adjusted reliability as defined here achieves the best match to the intuitions of the experimental consultants.

2.3 Using the Grammar

The grammar learned by the minimal generalization algorithm consists of a large number of rules, each annotated for its adjusted reliability. What remains is to define how this grammar is used. In principle there are two things that we want a grammar to do: derive forms, and rate their well-formedness. The existence of word-specific rules in the grammar, which constitutes a kind of memorization, guarantees that existing forms will be derived correctly. Therefore, the true test of a grammar is its ability to derive novel forms.

Machine-learned grammars can be tested with two kinds of novel forms: existing forms that were deliberately excluded from the learning set, or completely novel, made-up forms. Since our interest here is in comparing the performance of the model against humans, we think that the use of made-up forms is most appropriate: they guarantee that both humans and algorithm are generating their outputs productively, rather than making use of memorization.

People often judge more than one outcome to be possible, and have intuitions about the relative well-formedness of the various outcomes. To derive multiple outputs with the model, we rely on the fact that the grammar has multiple, often conflicting rules. Thus, when we apply the rules of the grammar to a novel input, the rules compete to produce different outputs. For example, if we feed the grammar the imaginary stressless root allomorph [lerr-], asking it to provide the corresponding stressed allomorph, then some of the applicable rules (those whose structural change is /e/ → [jé]) would derive [ljérr-], and others would derive the non-changing form [lérr-].

We then assign well-formedness scores to each output. For each form, this is defined as the adjusted confidence of the best rule that generates it. In the simulation below, we find that [ljérr-] is assigned a score of .92 (by the rule e → jé / [X ___ rr]), and [lérr-] .91 (by the rule e → é / [X l ___ Y]). These values can then be matched up with data obtained from human consultants, in a way that we will describe in §3.

2.4 Data Submitted to the Learner

We fed the minimal generalization learner pairs of verb root allomorphs. The first member of each pair was a stressless allomorph such as [kont-], as is found in the 1st plur. pres. [kontámos] 'we count'. The second member of each pair was the corresponding stressed allomorph [kwént-], as is found in the 1st sg. pres. [kwénto] 'I count'. The algorithm developed a grammar to project stressed allomorphs from stressless. For verbs with mid vowels, this requires deciding whether the stressed allomorph will have a diphthong or not.

In roots like [empes-] ~ [empjés-] 'begin', there are two locations where diphthongization could in principle apply: should the structural change /e/ → [jé] affect the first /e/ or the second? In fact, the vowel that diphthongizes is always that one that receives the stress, following the verb stress rules of Spanish. Since the minimal generalization learner does not at present incorporate a capacity for stress rules, we bypassed this problem by marking in the learning data the vowel whose stress is changed. This appears to be a legitimate idealization, for two reasons. First, the stress pattern for the relevant forms is quite predictable (see, e.g. Harris 1987). Second, the experimental work of Graham (1977) indicates that Spanish-learning children command the verbal stress pattern quite solidly at a period when they still have little control over the pattern of vocalic alternation.

The roots we chose were those of all 1,698 mid-vowel first conjugation verbs in a 5.5 million word corpus of Spanish (LEXESP: Sebastián, Cuetos and Carreiras, forthcoming). The format in which the learning data were presented to the algorithm is illustrated by the boxed columns below, with the 12 most frequent verbs.[4] Underlining represents the arbitrary mark we used to encode the location of stress alternation:

(7)	Verb	Stressless	Stressed	Gloss	LEXESP
		allomorph	allomorph		frequency
	dejar	[dex]	[déx]	‘to let’	4999
	quedar	[ked]	[kéd]	‘to stay’	4504
	encontrar	[enkontr]	[enkwéntr]	‘to find’	4044
	pensar	[pens]	[pjéns]	‘to think’	3891
	contar	[kont]	[kwént]	‘to count’	2732
	entrar	[entr]	[éntr]	‘to enter’	2661
	tomar	[tom]	[tóm]	‘to drink’	2584
	crear	[kre]	[kré]	‘to create’	2527[5]
	empezar	[empes]	[empjés]	‘to begin’	2375
	esperar	[esper]	[espér]	‘to wait’	2353
	recordar	[rekord]	[rekwérd]	‘to remember’	2085
	considerar	[konsider]	[konsidér]	‘to consider’	1741

2.5 The Machine-Learned Grammar

The grammar that resulted from feeding the learning data to the minimal generalization learner had 3,346 rules, of which 1,698 were word-specific and 1,648 were generalized. The vast majority of these would never be used in deriving novel forms, since they would be overridden by competing rules that have the same structural change but a higher adjusted reliability. In some versions of our learner these useless rules are actually discarded; since no empirical consequences follow from whether or not we do this, we will ignore this issue here.

[4] Our verb corpora, along with full data from all the experimental subjects, are posted at http://www.humnet.ucla.edu/linguistics/people/hayes/SegEnvSpanDiph/.

[5] The frequency of *crear* was inflated in the lemmatization of the LEXESP corpus because the automated lemmatization algorithm did not distinguish *creo* ‘I create’ from *creo* ‘I believe’, a second conjugation verb (Lluís Padró, personal communication). Since token frequencies were not used in our learning algorithm, the error is harmless.

The question of grammar size impinges on current controversies over what constitutes grammatical knowledge, which we address in §4.1 below.

Some of the rules that do derive output forms are listed below. After each rule, we give its adjusted reliability, along with the lexical statistics (hits and scope) that were used in calculating it.

(8)	Change	Environment	Effectiveness Statistics (hits/scope)	Adjusted Reliability
	a. e → jé	/ [X ___ rr]$_{root}$	11/11	0.92
		/ [X ___ mb Y]$_{root}$	4/4	0.79
		/ [X ___ nt]$_{root}$	24/88	0.24
	b. o → wé	/ [s ___ l Y]$_{root}$	3/3	0.72
		/ [X ___ st Y]$_{root}$	8/15	0.44
		/ [X ___ b Y]$_{root}$	11/33	0.28

These environments represent “islands of reliability” for the diphthongization alternations: they are phonological structural descriptions for which the Spanish verb lexicon is rich in verbs that diphthongize.[6]

The algorithm also learned specific rules for the “no-change” mapping, as in [r**e**ntámos] ~ [r**é**nto]. These rules reflect the preference for non-alternation in particular environments. Thus, for example, the contexts below are islands of reliability for the no-change outcome:

(9)	a. e → é	/ [X ___ tʃ]$_{root}$	15/15	0.94
		/ [X ___ k]$_{root}$	11/11	0.92
		/ [X ___ l]$_{root}$	36/38	0.91
	b. o → ó	/ [X ___ t Y]$_{root}$	38/38	0.98
		/ [X ___ m Y]$_{root}$	24/24	0.97
		/ [X ___ r Y]$_{root}$	121/134	0.88

Note that these rules have segmental environments, but specify no change, other than the automatic shift of stress.

Overall, diphthongization tends to be disfavored in the data. We can see this by examining the adjusted reliability of the context-free rules for diphthongization and for the no-change outcome.

[6] The existence of such islands for Spanish was proposed on the basis of hand-checked data by Brame and Bordelois (1973). Our own approach uses machine checking to increase the accuracy with which the islands are located and evaluated, and experimental work to determine whether native speakers internalize them.

(10) a.	e → jé	/ [X ___ Y]$_{root}$	93/1029	0.08
b.	e → é	/ [X ___ Y]$_{root}$	918/1029	0.89
c.	o → wé	/ [X ___ Y]$_{root}$	71/669	0.10
d.	o → ó	/ [X ___ Y]$_{root}$	588/669[7]	0.87

It should be noted that the data under consideration include only first conjugation verbs; diphthongization is somewhat more widespread in the other two conjugation classes.

We will discuss further rules of the grammar below, in connection with how the grammar modeled the behavior of human speakers in our wug-testing experiment.

3 A “Wug” Test of Spanish Diphthongization

As seen in the previous section, the grammar learned by our model predicts that roots may vary widely in their likelihood to diphthongize, based on their phonological shape. For instance, a novel root like /lerr-/, since it contains the / [X ___ rr] environment of (8), should be relatively likely to diphthongize, in comparison to a root like /detʃ-/, which contains the environment / [X ___ tʃ] that favors non-alternation. To test these predictions, we developed a set of novel words that exemplified these differences.

3.1 Creating the Wug Forms

We used the grammar learned by our model as a tool for locating appropriate wug forms. We submitted to the grammar a set of approximately 4000 logically possible roots, and obtained the predicted well-formedness values for both diphthongized and non-diphthongized outcomes. We selected our wug forms from the two extremes.[8] All candidate wug verbs were looked up in a dictionary (Casares 1992), in order to avoid real verbs or verbs that included real roots. The 33 wug roots that we selected, given in the order in which they were presented to our consultants, were as in (11):

(11) 1. retolb-
2. ent-
3. pre-
4. sendr-
5. norr-
6. getʃ-
7. botr-
8. tʃostr-
9. detʃ-
10. fot-
11. derr-
12. lek-
13. del-
14. solm-
15. kert-
16. tebr-
17. gembl-
18. soltr-
19. tox-
20. tʃej-
21. lop-
22. remp-
23. nom-
24. fostr-
25. tʃort-
26. mobr-
27. bekt-
28. lerr-
29. solk-
30. debr-
31. bold-
32. lorr-
33. kolb-

[7] In the third column, the hits values do not sum to the scope values (e.g. 93 + 918 = 1011, not 1029) because there are a few irregular verbs that cannot be characterized by either structural change.

[8] We found that it was much easier to find good non-diphthongizers than good diphthongizers, since many of the good diphthongization environments are already filled by existing verb roots.

In designing the input for the minimal generalization learner, we had used stressed and stressless roots. However, verb roots are never pronounced in isolation in Spanish, so when presenting novel verbs to consultants, it is necessary to use fully inflected verbs. In the test reported here, we assigned all of the wug verbs to the first conjugation, since this is the most productive class, used for borrowings and neologisms. We presented wug verbs in the first person plural present indicative (a stressless form), and asked speakers to provide the corresponding first person singular (a stressed form). For example, consultants were given [lerrámos] 'we *lerr*', and were asked to say 'I *lerr*', which would emerge as [ljérro], [lérro], or occasionally some other form.

We chose to stick with just one morphological mapping (1 plur. pres. → 1 sg. pres.), since earlier experimental work on the productivity of diphthongization (Bybee and Pardo 1981, Eddington 1996) indicates that it is highly dependent on what morphological process is involved. By using just one mapping, we controlled for this source of variation.

3.2 Procedure

Consultants were wug-tested in individual interviews by the second author, a native speaker of Spanish. Wug verbs were presented orally, and consultants' responses were also given orally; the sessions were taped for later transcription. In the interview, the experimenter read each wug verb aloud in isolation, inflected in the 1 plur. present, for example, [lerrámos]. After hearing each verb, the consultants volunteered inflected forms that would appropriately fill in the blanks in the following dialogue:

(12) Experimental Protocol

Experimenter: [lerrámos]
Consultant (reads, filling in blanks):

Cada verano mi familia y yo _[*lerramos*]_ durante las vacaciones.
Every summer my family and I (*lerr*) while on vacation.

Hemos __[*lerrado*]__ cada verano por diez años.
We have *lerred* every summer for ten years

Me fascina ___[*lerrar*]___.
I love to *lerr*

Tengo seis meses que yo no ___________.
It's been six months since I've *lerred*

For the first blank, the expected response [lerrámos] is simply a repetition of the 1st pers. plur. form provided by the experimenter. This served as a control, to make sure that the consultant had heard the verb correctly. For the second and third blanks, the expected responses are the past participle [lerrádo] and the infinitive [lerrár]. These forms were also controls: among regular verbs, they always use the same (stressless) form of the root as the 1st plur. pres. Thus, these forms tested whether the consultant had been able to internalize the root sufficiently to perform completely predictable morphological operations on it. Finally, the last blank was the target form, the 1st sg. pres., which forced the consultant make a decision (possibly tacit) about whether to diphthongize or not: [lérro], [ljérro], or on occasion something else. Note that although the English gloss for this frame involves a past participle, in Spanish, a present tense form is required in this syntactic context.

When a consultant did not give the expected answer for one or more of the three control forms, their answer for the target form was discarded.

After the consultant had volunteered a 1st sg. pres. form, the experimenter also elicited acceptability ratings for 1st sg. pres. forms with both diphthongized and unchanged roots. Ratings were given orally on a scale from 1 (worst) to 7 (best). This was done by reading both forms in the 1st sg. pres. sentence frame from above:

(13) a. Tengo seis meses que yo no [lerro].
b. Tengo seis meses que yo no [ljerro].

Half of the consultants were asked for acceptability ratings in the order shown in (13) (no-change first); the other half were asked in the opposite order.

At the beginning of the session, consultants were trained using four real forms of Spanish. These included two diphthongizing verbs (*contar* ‘count’ and *sembrar* ‘seem’) and two no-change verbs (*cantar* ‘sing’ and *montar* ‘mount’).

The test was administered to 96 native speakers of Spanish. Of these, 48 were monolinguals in Guadalajara, Mexico, and 48 were Spanish/English bilinguals in Los Angeles. The consultants ranged in age from 9 to 70 and volunteered their time. The task took 30 to 60 minutes.

3.3 Results

3.3.1 Control Forms

The first three blanks in dialogue (12) were designed to test whether the consultant had been able to internalize and manipulate each test verb. Errors on the control forms typically involved (a) segmental misperceptions, as in *[**k**emblámos] for [**g**emblámos]; (b) misparsing the [-am-] of -*amos* as part of the root, as in *[detʃamádo] instead of [detʃádo]; (c) substitution of phonetically similar real verbs, as in *[brotámos] ‘bloom’ for [botrámos]; (d) addition of the verb-forming suffix [-e-], as in [nomeádo] for [nomádo].[9]

The number of responses that were discarded for any one consultant ranged from 0 to 33 (33 items total); however for most consultants only a few responses were discarded (median = 2, mean = 4.6). Overall, 444 out of 3168 responses, or 14%, were discarded because of an incorrect response on the control forms. We later reran the analyses discarding *all* data from any

[9] The use of this suffix by consultants has been observed in previous wug-testing studies of Spanish (Bybee and Pardo 1981:941).

subject who produced more than 3 incorrect control forms; this did not materially affect the result.

3.3.2 Production Probability

In matching the behavior of the consultants to the behavior of our formal model, it is useful to use the notion of **production probability**. Given a test form T (e.g. *lerrámos*) and a possible response R (e.g. *ljérro*), the production probability of R is defined as in (14):

(14) $\frac{\text{number of times R was volunteered}}{\text{number of valid responses for T}}$

In calculating the number of “valid responses,” we excluded any volunteered forms other than the expected no-change and diphthongized variants, since our focus was on the diphthongization alternation. However, recalculation of the correlations with these forms included yielded essentially the same results.

Here is how production probability was calculated in the case of [lerrámos]. Of the 96 consultants, 12 produced ill-formed responses for the control forms, and their data were discarded for this item. Of the remaining 84 consultants, 57 volunteered [lérro], 14 volunteered [ljérro], and 13 volunteered other forms that were not considered.[10] Therefore, the number of valid responses for this item was 70, the production probability of [lérro] was 56/70 = 0.80, and that of [ljérro] was 14/70 = 0.20.

If a form was volunteered by every consultant, its production probability would be 1; and if the form was never volunteered, its production probability would be 0. The sum of the production probabilities for the competing outputs is always one.

Next, we must consider how production probability should be modeled. The problem is that the minimal generalization learner does not output production probabilities per se, but rather well-formedness scores. There is no guarantee that the well-formedness scores for all of the candidates for a given root will sum to 1, as the production probabilities do. It may be that several outcomes receive a high score (sum of scores > 1) or that no outcome receives a high score (sum of scores < 1). Therefore, in order to model the production probabilities of the consultants, we need to adjust the well-formedness scores that were assigned by the model.

To do this, we made what we take to be a plausible assumption: that the production probabilities of rival forms are proportional to their well-formedness scores (that is, people are more likely to say things that sound well-formed to them). This assumption led us to adjust the scores in the following way. Given, for example, a predicted well-formedness score for [lerrámos] ~ [lérro] of 0.91 and a predicted well-formedness score for [lerrámos] ~ [ljérro] of 0.92, we summed these values, obtaining 1.83. We then divided the well-formedness scores by this sum to obtain predicted production probabilities: for [lérro], 0.91/1.83 = .497, and for

[10] These were: [lerréo] (4), [lerrámo] (2), [ljéro] (1), [léro] (1), [lwérro] (1), [légdro] (1), [béltro] (1), and two forms probably not meant as 1st sg. pres., namely [lerrámos] (1) and [e lerrádo] (1).

[ljérro], 0.92/1.83 = .503. The sum of the predicted production probabilities for competing outputs calculated in this way is always 1.

3.3.3 Results for Volunteered Forms

Table 1 lists the production probabilities for all of the wug verbs. Column 2 lists the experimentally observed production probabilities for the no-change outcomes, and column 5 for the diphthongized outcomes. Columns 3 and 6 list the corresponding values predicted by the computer model. Finally, columns 4 and 7 give the well-formedness scores from which the predicted production probabilities were calculated. The forms are sorted in decreasing order of the consultants' production probability for diphthongized forms.

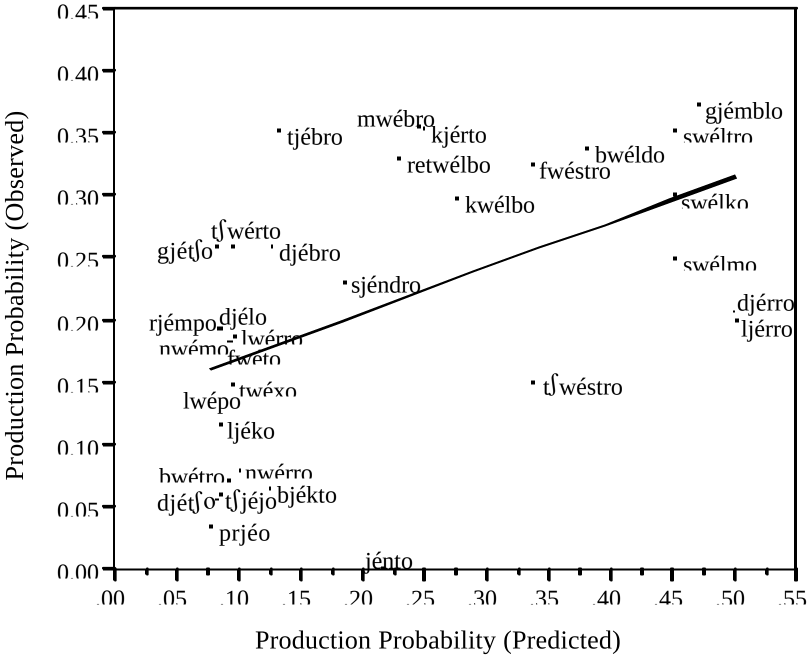
Table 1: Results from the Production Task

	No Change (e.g. *retólbo*)			**Diphthongization (e.g. *retwélbo*)**		
[1] Wug form	[2] Prod. Probability: Humans	[3] Prod. Probability: Model	[4] Well-formedness score: Model	[5] Prod. Probability: Humans	[6] Prod. Probability: Model	[7] Well-formedness score: Model
gembl-	**0.63**	**0.53**	0.89	**0.37**	**0.47**	0.79
mobr-	**0.64**	**0.76**	0.87	**0.36**	**0.24**	0.28
kert-	**0.65**	**0.75**	0.90	**0.35**	**0.25**	0.30
soltr-	**0.65**	**0.55**	0.87	**0.35**	**0.45**	0.72
tebr-	**0.65**	**0.87**	0.89	**0.35**	**0.13**	0.14
bold-	**0.66**	**0.62**	0.92	**0.34**	**0.38**	0.57
fostr-	**0.68**	**0.66**	0.87	**0.33**	**0.34**	0.44
retolb-	**0.67**	**0.77**	0.94	**0.33**	**0.23**	0.28
kolb-	**0.70**	**0.72**	0.87	**0.30**	**0.28**	0.33
solk-	**0.70**	**0.55**	0.87	**0.30**	**0.45**	0.72
debr-	**0.74**	**0.87**	0.89	**0.26**	**0.13**	0.13
getʃ-	**0.74**	**0.92**	0.94	**0.26**	**0.08**	0.09
tʃort-	**0.74**	**0.91**	0.95	**0.26**	**0.09**	0.10
solm-	**0.75**	**0.55**	0.87	**0.25**	**0.45**	0.72
sendr-	**0.77**	**0.81**	0.89	**0.23**	**0.19**	0.20
derr-	**0.79**	**0.50**	0.92	**0.21**	**0.50**	0.92
lerr-	**0.80**	**0.50**	0.91	**0.20**	**0.50**	0.92
del-	**0.81**	**0.91**	0.91	**0.19**	**0.09**	0.09
lorr-	**0.81**	**0.90**	0.93	**0.19**	**0.10**	0.10
remp-	**0.81**	**0.92**	0.93	**0.19**	**0.08**	0.09
fot-	**0.82**	**0.91**	0.98	**0.18**	**0.09**	0.10
nom-	**0.82**	**0.91**	0.96	**0.18**	**0.09**	0.10
lop-	**0.85**	**0.91**	0.95	**0.15**	**0.09**	0.10
tox-	**0.85**	**0.90**	0.94	**0.15**	**0.10**	0.10
tʃostr-	**0.85**	**0.66**	0.87	**0.15**	**0.34**	0.44
lek-	**0.88**	**0.92**	0.92	**0.12**	**0.08**	0.09
norr-	**0.92**	**0.90**	0.88	**0.08**	**0.10**	0.10
bekt-	**0.93**	**0.87**	0.93	**0.07**	**0.13**	0.13

botr-	**0.93**	**0.91**	0.98	**0.07**	**0.09**	0.10
detʃ-	**0.94**	**0.92**	0.94	**0.06**	**0.08**	0.09
tʃej-	**0.94**	**0.91**	0.91	**0.06**	**0.09**	0.09
pre-	**0.96**	**0.92**	1.00	**0.04**	**0.08**	0.09
ent-	**1.00**	**0.78**	0.89	**0**	**0.22**	0.24
mean	**0.79**	**0.79**	0.91	**0.21**	**0.21**	0.29

The same data are shown graphically in Figure 1, in which predicted production probabilities are plotted against observed probabilities for the diphthongized forms.

Fig 1. Production Probability of Diphthongized Forms
(Predicted vs. Observed)



The predictions of the model are fairly well correlated with the observed production probabilities (r(31) = .510, p = .0025).

This correlation is due to the existence of islands of reliability. For example, the island for diphthongization described by the change o → wé / [s ___ l Y], given above in (8), favors diphthongized outcomes like [swéltro], [swélko], and [swélmo], visible in the upper right portion of Figure 1. There are likewise islands of reliability for no-change outcomes, such as (9), o → ó / [X ___ t Y]. The latter island favors outcomes like [fóto] and [bótro]. Accordingly, the predicted production probability for the rival forms [fwéto] and [bwétro] is low; these forms may be seen in the lower left portion of Figure 1. An important question is whether *both* types of island have an independent effect. We return to this issue in §3.4.4.

3.3.4 Results for Well-Formedness Ratings

Recall that the consultants in the experiment were also asked to rate both no-change and diphthongized forms along a well-formedness scale ranging from 1 (worst) to 7 (best). Half the consultants rated the no-change form first, and the other half rated the diphthongized form first. We first describe overall characteristics of the data, then consider to what degree they can be accurately modeled with the minimal generalization learner.

We carried out an analysis of variance (ANOVA) on the data, with the following factors: diphthongization (whether the form being rated was diphthongized or not), presentation order (no-change first/diphthongized first), and verb identity (which of the 33 wug verbs was being rated).

The test indicated a main effect of diphthongization ($F(1) = 158.958$, $p < .0001$), with no-change items preferred over diphthongized. This means that, as expected, diphthongization is globally dispreferred.

There was no main effect of presentation order, indicating that speakers did not consistently prefer the forms presented to them first or second. However, there was a significant interaction between presentation order and diphthongization ($F(1) = 4.616$, $p < .01$). Specifically, no-change forms, but not diphthongized forms, received higher ratings when presented first. We have no explanation for this observation. However, there was no three-way interaction with verb identity, which means that the extra boost for hearing no-change forms first was not greater for some test items than for others. We therefore combined the ratings from the two presentation orders, but carried out separate analyses for the ratings of diphthongized and no-change forms.

More important, the test showed a significant two-way interaction of diphthongization and verb identity ($F(32) = 1.906$, $p < .01$). This means that diphthongization sounds better or worse, depending on the phonological shape of the root. This agrees with the result seen above for production probabilities.

Since the two-way interaction of diphthongization and verb identity shows that root shape does have an effect on the judgments, we want to know why there should be such verb-by-verb differences. We explored this by comparing the consultants' rating with those of the computer model.

Table 2 compares the averaged ratings of the consultants with the predicted values from the computer model. The results are listed in decreasing order of the consultants' ratings for diphthongized forms. The predictions of the computer model (already given in their raw values in Table 1) are rescaled here to match the 1-7 scale used in the experiment.

Table 2: Results from the Judgment Task

	No Change (e.g. *retólbo*)		**Diphthongization (e.g. *retwélbo*)**	
[1] Wug form	[2] Well-formedness score: Humans	[3] Well-formedness score: Model	[4] Well-formedness score: Humans	[5] Well-formedness score: Model
soltr-	4.2	6.2	4.4	5.3
tebr-	4.5	6.3	4.4	1.8
getʃ-	4.4	6.6	4.3	1.5
kert-	4.3	6.4	4.2	2.8
kolb-	4.3	6.2	4.1	3.0
bold-	4.5	6.5	4.0	4.4
fostr-	4.5	6.2	4.0	3.7
gembl-	4.7	6.3	4.0	5.7
lerr-	4.5	6.4	3.9	6.5
mobr-	4.4	6.2	3.9	2.7
sendr-	4.9	6.3	3.9	2.2
solk-	4.3	6.2	3.9	5.3
solm-	4.8	6.2	3.8	5.3
derr-	4.5	6.5	3.7	6.5
lorr-	4.4	6.6	3.7	1.6
retolb-	4.5	6.7	3.7	2.7
debr-	4.7	6.3	3.6	1.8
lop-	4.7	6.7	3.6	1.6
nom-	4.8	6.8	3.6	1.6
tʃort-	4.7	6.7	3.6	1.6
tʃostr-	4.8	6.2	3.6	3.7
remp-	5.1	6.6	3.5	1.5
del-	4.7	6.5	3.4	1.5
tox-	4.6	6.7	3.4	1.6
fot-	4.5	6.9	3.3	1.6
lek-	4.6	6.5	3.3	1.5
tʃej-	4.8	6.4	3.3	1.5
botr-	4.9	6.9	3.2	1.6
detʃ-	5.1	6.6	3.2	1.5
bekt-	5.3	6.6	3.0	1.8
norr-	4.7	6.3	3.0	1.6
pre-	5.3	7.0	3.0	1.5
ent-	4.9	6.3	2.8	2.5
Mean	*4.6*	*6.4*	*3.7*	*2.8*

Several things can be noticed in this table.

First, the well-formedness ratings from consultants occupy a narrower range than those of the computer model. This is a typical effect that is seen when acceptability ratings are averaged across consultants.

Second, the predicted ratings of the computer model for all forms are highly correlated with the averaged consultant ratings (r(64) = .838, p < .0001 for the no-change-first presentation order; r(64) = .678, p < .0001 for the diphthongized-first presentation order). These correlations, though high, are of little interest for present purposes. The computer model correctly apprehended that the no-change outcome is almost always a better guess than the diphthongized outcome. Since the consultants tacitly arrived at the same conclusion (see ANOVA results above), a high correlation results. Our focus is on effects of segmental environment; to test for these, we must examine the no-change forms and the diphthongized forms separately.

The predicted ratings of the computer model for diphthongized forms are correlated with the averaged consultant ratings (r(31) = .454, p < .01). This indicates, as above, that the islands of reliability for diphthongization that are used by the computer model in making its predictions are, at least to some degree, apprehended by Spanish speakers. The computer model's ratings for no-change forms are also somewhat correlated with the averaged consultant ratings of these forms (r(31) = .385, p < .05).

We note finally that that the mean ratings for particular outputs are highly correlated with the production probabilities described in the previous section: r(64) = .930 overall; r(31) = .902 for diphthong outputs; r(31) = .639 for no-change outputs. This provides some assurance that when consultants make a conscious, intuitive judgment, they are accessing the same knowledge that they use in spontaneous productions.

3.4 Discussion

The results above indicate that Spanish diphthongization is indeed influenced by segmental environments. Below, we discuss the data pattern in further detail.

3.4.1 Independent Effects on Well-Formedness

For some of the wug forms, there are alternative possibilities for why the consultants might have favored diphthongization or no-change. We list below what seem the most plausible ones.

Phonotactic problems. For some of the roots, the diphthongized form includes segment sequences that are unusual or unattested (though still pronounceable) in Spanish. These are listed in (15); the phonologically-awkward sequence is given in parentheses.

(15) a. [gjétʃo], [djétʃo] (?[jetʃ])
b. [tʃjéjo] (?[jej], ?[tʃje])
c. [prjéo] (?[jeV])

Learned status. The wug root /bekt-/ contains a /kt/ cluster, which generally occurs only in vocabulary borrowed from Latin and other languages. If such forms synchronically occupy a learned stratum (Harris 1969, McCawley 1968, Ito and Mester 1995a, 1995b), and if diphthongization occurs only in non-learned forms, we would expect speakers to avoid diphthongization in /bekt-/.

Extreme resemblance to an existing verb. The following wug roots closely resemble existing verbs of Spanish; their diphthongization behavior might be attributed to direct analogy with the existing verb.

(16)	[mwébro]	(like [mwébo] ‘I move’)
	[kjérto]	(like [kjéro] ‘I like’)
	[sjéndro]	(like [sjémbro] ‘I plant’)
	[swéltro]	(like [swélto] ‘I let go’)

Of course, every novel verb resembles numerous existing verbs to varying degrees. Although the model employed here does not rely directly on similarity to existing words, it does pay attention to shared phonological structure. This is precisely what leads to islands of reliability, and thus to differences between items based on their phonological form. We are concerned here only with the possible confound of *extreme* resemblance, where individual lexical items are most likely to assert a direct influence.

Homophony. If the wug root /fot-/ is *not* diphthongized, the resulting 1st sg. pres. form [fóto] is homophonous with the noun [fóto] ‘photo’. Conceivably, consultants may have been more likely to diphthongize /fot-/ for this reason.

To provide a stricter test of our general hypothesis (that speakers learn phonological environments for diphthongization/no-change), we recalculated the correlations given above, excluding the 10 potentially confounded items just mentioned from the data set. The results were as follows:

Table 3: Results with 10 Wug Roots Removed

(a) Production probabilities

All roots (from §3.3.3)	10 roots removed
r (31) = .510	r(21) = .427

(b) Well-formedness ratings

	All roots (from §3.3.4)	10 roots removed
Diphthongized	r(31) = .454	r(21) = .450
No-change	r(31) = .385	r(21) = .265

The values for production probability, and for the ratings for diphthongized forms, remain significant at the .05 level; the correlation for well-formedness ratings remains positive but is not significant.

3.4.2 Consultant Backgrounds

The data pattern differs for consultants of different personal backgrounds. In general, higher correlations, for both production probabilities and acceptability ratings, were found for the following groups:

- Bilingual consultants interviewed in Los Angeles (vs. the mostly monolingual consultants interviewed in Guadalajara)
- Consultants between the ages of 18 and 40
- Consultants with more education (some college > {some high school, some junior high school} > elementary school)

Some causal factors that may have been involved are as follows. Educated consultants probably know more verbs, and have more knowledge of the normative diphthongization pattern of standard literary Spanish, which was what the computer model learned. Consultants under 18 and over 40 tended to be less educated. Further, the oldest consultants tended to have more trouble understanding the experiment (they missed more of the control questions), and were most likely to give undifferentiated ratings consisting of only 1's and 7's. Finally, it can be noted that those consultants who had only an elementary school education were for the most part the oldest ones, so in this area the effects of age and education are confounded.

3.4.3 Size of the Learning Set

A computer model intended to mimic human speakers should do better if the learning data fed to it is similar to the data encountered by people in the course of acquisition. The model described here learned from a data set of 1,698 mid-vowel first conjugation verbs. In these data, the diphthongization pattern is that of standard normative varieties of Spanish, and the data include a number of relatively rare roots that might not have been known to all of the consultants.

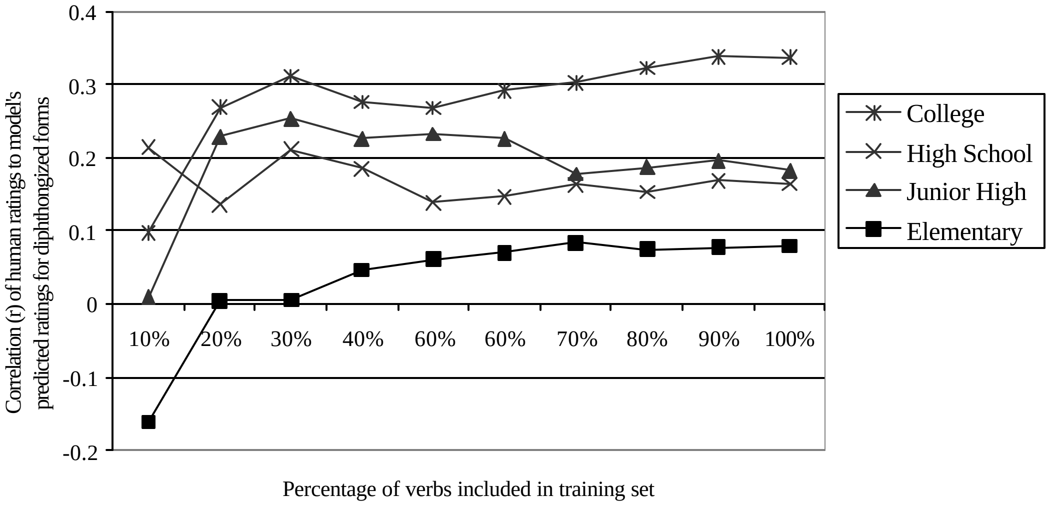
This raises the possibility that the data set that we fed to the minimal generalization learner did not accurately reflect the input data of our consultants. If consultants do not know all of the verbs contained in the full database, then there is no way that these verbs could influence their morphological intuitions. Furthermore, we might hypothesize that even for those consultants who know all of the verbs in the complete set, the rarer verbs may have been learned so late in life that morphological learning was essentially complete, and therefore they do not contribute to morphological intuitions. In both of these cases, excluding the rarer words from the training set should provide a more accurate model of consultant intuitions.

We tested this by dividing the entire database of 4,795 verbs found in the LEXESP corpus into deciles, and constructing ten progressively larger learning sets. The first learning set contained all of the (mid-vowel, first conjugation) verbs in the top decile (137 verbs), the second learning set contained all of the verbs in the top two deciles (297 verbs), and so on. The last learning set contained all 1,698 mid-vowel first conjugation verbs in the database. These ten learning sets were then fed to the minimal generalization learner, yielding predicted well-formedness ratings for each "stage."

In order to test the hypothesis that the optimal training set may differ depending on consultant's background, the consultants were divided up according to level of education in Spanish: (a) no education past elementary, (b) no education past junior high school, (c) no education past high school, (d) at least college education. Since it was not clear how one would assess the effects of education in English, the comparison was carried out only for the monolingual Guadalajara group.

Finally, the predicted ratings of diphthongized forms at each of the 10 stages were correlated with the ratings given by the four subgroups of Guadalajara consultants. The results are shown in Figure 2.

Figure 2: Size of training set vs. performance of the model



Not too much can be learned from these data. Plainly, increased education correlates with a closer match to the predictions of the minimal generalization learner, irrespective of the size of the learning data set. We conjecture that this may involve greater experience among the more educated consultants with formal testing situations. The college-educated consultants were the only ones whose data curves rises through the last quarter of the training set, which could conceivably be an effect of their larger vocabularies.

3.4.4 The Trade-Off Hypothesis

We have been assuming that diphthongization and no-change are handled by separate rules, each with their own segmental environments. An alternative possibility is that consultants' intuitions about the no-change forms are simply the complement of their intuitions about diphthongized forms: a no-change form sounds good to the extent that the corresponding diphthongized form sounds bad. We will refer to this as the "trade-off hypothesis".

The trade-off hypothesis is compatible with the view that speakers cannot learn rules with vacuous structural changes. Under this view, Spanish speakers learn only environments specific to diphthongization, and rely on these environments when producing or rating no-change forms.

The data pattern of the Spanish lexicon, as determined by the computer model, clearly involves special environments specific to both diphthongization and no-change. Some environments of each type are given in (9) above. Because both environments exist, the computer model's predictions for no-change and for diphthongization show only a moderate inverse correlation (r(31) = -.486). Indeed, there are roots for which the model's ratings are high in both the diphthongized and no-change forms. An example is [lerrámos], with [lérro] rated at .91 by the model and [ljérro] at .92.

Do Spanish speakers nonetheless behave as the trade-off hypothesis would predict? Our data suggest a mixed view. On one hand, there is evidence that speakers do find and use specific no-change environments, contrary to the predictions of the trade-off hypothesis. However, speakers, unlike the model, exhibit trade-off behavior.

The evidence that speakers apprehend independent islands of reliability for both diphthongization and no-change is based on a partial correlation. We first factored out the effects of diphthongization, to see whether there were any residual effects that could be explained by the no-change environments. Recall that the simple correlation between the computer model's no-change predictions and the consultants' no-change ratings was .385. When the predicted ratings for diphthongized forms are factored out first, the remaining correlation between the computer model's predicted ratings for the no-change forms and the consultants' ratings is .249. Thus, some but not all of the variance in the consultants' ratings of no-change forms can be attributed to the influence of the competing diphthongized output. We conclude that consultants must be using information about both diphthongization and no-change environments in forming their judgments.

On the other hand, the consultants did rate diphthongized and no-change forms in a more complementary fashion than the computer model did. The inverse correlation of the consultants' ratings for diphthongized and no-change forms was -.721. The analogous correlation for the computer model's predicted ratings was only -.486. Thus, the consultants were more likely than the computer model to treat the no-change and diphthongized outcomes as competitors. This may be a result of the experimental setup—recall that the consultants rated the two forms one after the other, and they may have adopted a trade-off strategy in rating the forms. Alternatively, the difference may reflect a real linguistic principle, morphological blocking (Aronoff 1976), although in a gradient way. If so, the computer model should be modified to incorporate such effects.

3.4.5 Modeling with All Three Conjugation Classes

In the experimental task, consultants were asked to rate the likelihood of a novel first conjugation form to diphthongize. In modeling the results, we fed the minimal generalization learner a set of verbs from the first conjugation only. This was based on the tacit assumption that real speakers learn the diphthongization environments separately for each conjugation class. Another possibility, however, is that speakers learn these environments in more general terms, going across the full vocabulary of verbs, or perhaps including other parts of speech as well.

To test this hypothesis, we ran the learner with a larger data set, consisting of 4,782 verbs from all three conjugation classes. The predicted production probabilities and well-formedness ratings were correlated against the consultant data, with the following results:

Table 4: Results with Learning Data from All Three Conjugations

(a) Production probabilities

1st conjugation only	All conjugations
r(31) = .510	r(31) = .595

(b) Well-formedness ratings

	1st conjugation only	All conjugations
Diphthongized	r(31) = .454	r(31) = .519
No-change	r(31) = .385	r(31) = .291

Table 4 shows that the correlation between consultants' production probabilities and the model's predictions is higher when only first conjugation verbs are included in the learning data. This difference is not significant, however, under Steiger's (1980) test for the significance of different correlations based on different predictor variables: t(31) = 1.585, p = .122. For the well-formedness ratings, a similar pattern is observed, also at non-significant levels (diphthongs t(31) = .924, p = .362; no-change t(31) = –.85, p = .403).

Although the results here were inconclusive, the issue at stake is important and worth considering further: is diphthongization to be considered a general phonological rule of Spanish, or is it somehow fragmented into separate cases for every affix in the language?

Some evidence supports the latter view:

- Although our consultants volunteered a wide variety of forms, no forms followed the pattern of vowel raising (e.g. [p**e**dír] ~ [p**í**do] 'to ask for'), which is fairly well attested but occurs only in the third conjugation.
- Eddington's (1996, 1998) wug-testing experiments have shown that diphthongization shows major differences of productivity across different morphological constructions.
- In the historical changes documented by Reyes (1978) analogical shifts have worked differently in different conjugations, and diphthongization has shown differing productivity for novel verbs in different conjugations (Malkiel 1984).

Thus, we have reason to think that further research will show that the different conjugation classes of Spanish do involve different diphthongization environments.

3.4.6 Is Diphthongization All One Change?

Most analyses of diphthongization in Spanish have collapsed the changes /e/ → [jé] and /o/ → [wé], expressing them as a single rule (see for example Harris 1969, 1985, Brame and Bordelois 1973, Carreira 1991, García-Bellido 1986). In this paper, we have assumed that there may be independent segmental environments that favor each of the two changes. However, if speakers really do learn diphthongization as a single operation, then they should not be sensitive to different environments for the two changes.

We tested this by rerunning the learning simulation with an altered learning set, in which all instances of /o/ and /we/ were replaced by /e/ and /je/; this had the effect of letting the algorithm seek environments that generalized across front and back vowels. Table 5 gives the correlations with the judgments of our consultants, compared with the correlations obtained earlier with front and back vowels separated.

Table 5: Results with /e/ → [jé] and /o/ → [wé] collapsed

(a) Production probabilities

/e/ vs. /o/ separated	/e/ vs. /o/ merged
r (31) = .510	r(31) = .424

(b) Well-formedness ratings

	/e/ vs. /o/ separated	/e/ vs. /o/ merged
Diphthongized	r(31) = .454	r(31) = .388
No-change	r(31) = .385	r(31) = .406

As can be seen, when the front and back vowels were merged in the learning set, the correlations for diphthongization dropped, although there was a small increase in the correlation for no-change. We infer, tentatively, that the consultants' knowledge of diphthongization was sensitive to environments specific to front and back vowels.

4 Conclusions

The previous sections indicate that segmental environment does influence the propensity of roots in Spanish to undergo diphthongization, and that speakers are tacitly aware of this effect. We suggested that this knowledge may extend to differences in conjugation class, that it involves particular environments for both diphthongization and no-change outcomes, and that the crucial contexts may be different for the /e/ → [jé] and /o/ → [wé] changes. The effects we observed were most noticeable for more highly educated consultants, perhaps because their own vocabularies more closely match what was given to the computational learner.

4.1 Is This Grammar?

We must now ask whether gradient segmental effects should be considered part of the *grammars* internalized by Spanish speakers. The approach we have taken answers this question affirmatively: the minimal generalization learner constructs a grammar that is quite traditional with regard to the content of the rules, but is perhaps unorthodox in the number of rules it contains.

There are two possible responses to this position, which we consider in turn.

4.1.1 Analogical Approaches

One response is to abandon the view that the consultants' behavior is guided by a grammar. Instead, the consultants might employ analogy, projecting novel forms by carrying out an online statistical comparison of the wug root with phonetically similar roots in their lexicons. Formal models that do this include connectionism (see, e.g. Rumelhart and McClelland 1986, Daugherty and Seidenberg 1994) and the "Analogical Modeling of Language" approach of Skousen (1989) and his colleagues. These models would further claim that *none* of the knowledge speakers have of their language takes the form of a grammar; the entire system is claimed to work analogically.

At present, we have no empirical data that could distinguish our own approach from purely-analogical models. However, our model makes distinct predictions. While it attends in close detail to the patterns in the lexicon, it does so during the learning phase. For the minimal generalization learner, detailed rules are inevitable, as they must be discovered as way stations along the path to the most general rules. In contrast, all other systems that detect detailed lexical patterns do so "on line," at the moment of wug testing, by accessing similar items in the mental lexicon. Therefore, our system makes the unique claim that speakers could judge novel forms without having to carry out any kind of mental access to phonetically similar forms. Such access *could* happen, but it would not be necessary, since it is the rule system that leads to the well-formedness judgment.

4.1.2 The Dual Mechanism Approach

Another alternative to our claim of large grammars would be to adopt the "dual mechanism" theory advocated by Pinker and Prince (1988, 1994).[11] In this theory, it is claimed that grammatical rules do exist, and govern much of our linguistic behavior. However, the grammatical rule component of the system is considered to be relatively small and free of excessive contextual detail. Minor generalizations are attributed to analogy with existing lexical items, and are claimed to be best modeled with a connectionist network.

Although the advocates of the dual-mechanism model have not taken on the question of irregular phonological alternations, one might suppose that a dual-mechanism account of our data would designate one particular pattern—either diphthongization or no-change—as the default. The opposite pattern would be accounted for by the connectionist network, on the basis of analogy with existing items.

It seems clear, under this view, that diphthongization cannot be the default pattern. Applying diagnostics from Pinker and Prince (1988), we note that diphthongization tends to occur primarily in high-frequency forms, it is diachronically unproductive (at least in the first conjugation; Malkiel 1984), it is vulnerable to leveling (Reyes 1978), and it is highly sensitive to details of phonological shape. Therefore, if there is a default pattern for Spanish diphthongization, it must be the no-change pattern.

On the other hand, we have seen above (§3.4.4) that no-change may have one property that is inconsistent with default status. That is, there is at least weak evidence suggesting that our consultants were sensitive to phonological islands of reliability for the no-change pattern. The existence of these islands is problematic for a strict interpretation of the dual-mechanism view, in which default forms are always derived by the default rule. An earlier case of this type involving Italian verbal morphology, with considerably higher correlations, is reported in Albright (1998).

[11] For a recent review of the controversy surrounding this approach, see Clahsen (1999) and the replies included in the same volume.

A modified version of the dual-mechanism theory would posit that default outcomes can be generated by both the connectionist network and the default rule. Under this view, detailed phonological environments could have an influence even among the no-change forms. This would be a weakening of the dual-mechanism hypothesis, and would require reassessment of the evidence that has been presented to justify the special status of the default.

The approach taken here, like the dual mechanism model, does posit highly general rules that derive no-change outcomes—these are listed in (10) above. However, these rules have no special status, as they must compete with all the other rules in the grammar for deriving an output. Where a more specific rule is applicable to the input and is known to be more effective, it takes precedence over the default. Thus, our own model derives all patterns with a single mechanism.

4.2 The Productivity of Diphthongization

Earlier wug-testing research on the Spanish diphthongization alternation (Kernan and Blount 1966, Bybee and Pardo 1981) found that consultants were quite reluctant to extend the diphthongization alternation to novel forms. These studies therefore concluded that the alternation was not a productive one. Based on our own work, we believe that the methods used in these pioneering studies may have underestimated the productivity of the alternation. One reason for this is that, as with other moderately-productive patterns, Spanish diphthongization is well attested only in a limited set of segmental environments. In our own study, the learning model helped us to find specific environments that closely matched the diphthongization pattern of the lexicon. By including wug forms that fit these environments, we were able to test the cases where productivity is maximized. The picture that emerges is that the diphthongization alternation is in fact moderately productive, but for the most part only among forms that occupy phonological islands of reliability for diphthongization.

Our results complement those of Eddington (1996, 1998), who likewise found fairly high productivity for diphthongization in particular morphological environments.

4.3 Exceptionless Patterns and Undominated Constraints

Recent work in Optimality Theory (Prince and Smolensky 1993) addresses a number of issues that have arisen here: free variation (Anttila 1997a,b), gradient well-formedness (Hayes, in press; Boersma and Hayes, in press), matching of frequencies (Boersma 1997, Boersma and Hayes), and exceptionful phonological patterns (Zuraw 2000). Thus it is natural to wonder whether our data could be modeled using some version of Optimality Theory.

The rules found by the minimal generalization learner can be interpreted as language-specific Optimality-theoretic constraints; for example, "if the stressless allomorph contains [e] in the context / [X ___ rr], the stressed allomorph must contain [jé]." If these constraints (as we will now call them) are ranked using the continuous scale developed in Boersma (1997), the resulting grammar can produce diphthongization at different frequencies in different environments.

As a check on whether our consultants' behavior could *in principle* be modeled by an OT grammar incorporating our constraints, we used the Gradual Learning Algorithm (Boersma 1997; Boersma and Hayes, in press) to rank a set of 125 constraints, which were selected from the full set found by the minimal generalization learner.[12] The learning data fed to the Gradual Learning Algorithm were taken not from the vocabulary of Spanish, but instead were the actual production probabilities from Table 1. The resulting grammar achieved high accuracy in matching the observed production probabilities, with an error rate of only 1.98%.

This result is of little theoretical interest, however, because a learning algorithm should not operate off of the intuitions of native speakers (as revealed in their reactions to made-up words), but rather from language data similar to what real language learners encounter.

To our knowledge, no currently available constraint ranking algorithm is up to this task. The reason is that all existing algorithms assume that if a constraint is never violated in the learning data, it should be ranked at the top of the grammar. However, contrary to this prediction, we find that small but exceptionless generalizations are characteristically unable to overcome generalizations that are exceptionful but more broadly based.

Our data include an interesting case of this sort. In Spanish, every verb root containing /e/ in the environment / [X ___ rr] is a diphthongizing root. Therefore, if absence of violations implies top ranking, the constraint that requires diphthongization for such roots will emerge as undominated. Moreover, all applicable constraints that specify no-change (such as the default /e/ → [é] / [X ___ Y]) have counterexamples in the lexicon, and therefore cannot be undominated. The implication is that our consultants should have consistently produced forms like [ljérro] and [djérro] instead of [lérro] and [dérro]. But in fact, [ljérro] and [djérro] had production probabilities of only .20 and .21, respectively.

The problem, we believe, is that exceptionlessness is not the only kind of testimony that should weigh in favor of a constraint: the robustness with which a generalization is attested also matters. In the case of [ljérro] and [djérro], a constraint that works 11 out of 11 times competes against (among others), the default constraint for no-change ((10)b), which works 918 out of 1029 times. Our conjecture is that Spanish learners tacitly notice this robustness difference, and use it in making their judgments. The minimal generalization learner attempts to model this behavior by using adjusted instead of raw reliability; in the present case, the adjustment is not enough, but it does go in the right direction.

We conclude that it would be profitable to pursue ranking algorithms for Optimality Theory that take into account the robustness of the generalizations embodied in the constraints.

4.4 The Role of Algorithmically-Discovered Grammars

The experiment we have reported showed that the intuitions of native speakers concerning the phonological patterns in their language may be both gradient and sensitive to phonological

[12] The constraints were selected to include both the most reliable and the most general constraints which pertained to our 33 wug forms. The results of the simulation may be downloaded from http://www.humnet.ucla.edu/linguistics/people/hayes/SegEnvSpanDiph/.

detail. In order to produce a grammar that was capable of matching these intuitions, we needed two things: the capacity to consider a large number of possible phonological environments, and the capacity to provide quantitative measures of the degree to which phonological generalizations can be trusted. Both requirements go beyond what can feasibly be obtained with handcrafted grammars, and motivate the move to algorithmic analysis.[13]

References

Albright, Adam (1998). Phonological subregularities in productive and unproductive inflectional classes: Evidence from Italian. M.A. thesis, UCLA. [http://www.humnet.ucla.edu/people/aalbrigh/papers.html]

Albright, Adam (2000). The lexical bases of morphological well-formedness. Poster presented at the 9th International Congress of Morphology, Vienna, Feb 25-27. [http://www.humnet.ucla.edu/linguistics/people/grads/aalbrigh/papers.html]

Albright, Adam, and Bruce Hayes (1998). *An automated learner for phonology and morphology*. Ms., UCLA. [http://www.humnet.ucla.edu/linguistics/people/hayes/learning/learning.htm]

Anttila, Arto (1997a). *Variation in Finnish phonology and morphology*. Doctoral dissertation, Stanford University, Stanford, Calif.

Anttila, Arto (1997b). Deriving variation from grammar: A study of Finnish genitives. In *Variation, change and phonological theory*, ed. by Frans Hinskens, Roeland van Hout, and Leo Wetzels. Amsterdam: John Benjamins. [http://ruccs.rutgers.edu/ROA/search.html, #63]

Aronoff, Mark (1976). *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press.

Berko, Jean (1958). The child's acquisition of English morphology. *Word.* **14**. 150-177.

Boersma, Paul (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **21**. 43 –58. [http://fonsg3.hum.uva.nl/paul/papers/learningVariation.pdf]

Boersma, Paul and Bruce Hayes (in press). Empirical Tests of the Gradual Learning Algorithm. to appear in *Linguistic Inquiry*. [http://www.humnet.ucla.edu/linguistics/people/hayes/GLA/].

Brame, Michael K. and Ivonne Bordelois (1973). Vocalic alternations in Spanish. *Linguistic Inquiry*. 4:111-168.

Bybee, Joan and Elly Pardo (1981) On lexical and morphological conditioning of alternations: a none-probe experiment with Spanish verbs. *Linguistics.* **19**. 937-968.

Bybee, Joan (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*. 10(5):425-255.

Carreira, María (1991). The alternating diphthongs of Spanish: A paradox resolved. In Héctor Campos and Fernando Martínez-Gil (eds.) *Current Studies in Spanish Linguistics*. Washington, D.C.: Georgetown University Press.

Casares, Julio (1992). *Diccionario ideológico de la lengua española*. Barcelona: Editorial Gustavo Gili, S.A.

Clahsen, Harald (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* **22**. 991-1060.

[13] The learning software used for this project may be downloaded from http://www.humnet.ucla.edu/linguistics/people/hayes/learning/.

Daugherty, Kim G. and Mark Seidenberg (1994). Beyond rules and exceptions: a connectionist approach to English morphology. In Susan D. Lima, Roberta L. Corrigan, and Gregory K. Iverson (eds.) *The Reality of Linguistic Rules*. Amsterdam: John Benjamins.

Eddington, David (1996). Diphthongization in Spanish derivational morphology: an empirical investigation. *Hispanic Linguistics.* **8:1**. 1-13.

Eddington, David (1998). Spanish diphthongization as a non-derivational phenomenon. *Rivista di Linguistica.* 10(2):335-354.

García-Bellido, Paloma (1986). Lexical diphthongization and high-mid alternations in Spanish: An autosegmental account. *Linguistic Analysis.* **16**. 61-92.

Graham, Charles R. (1977). *The Development of Linguistic Stress in Children Learning Spanish as a First Language*, Ph.D. dissertation, University of Texas at Austin.

Halle, Morris and Jean-Roger Vergnaud (1987). *An Essay on Stress*. MIT Press, Cambridge, MA.

Harris, James (1969). *Spanish Phonology*, MIT Press, Cambridge, MA.

Harris, James (1977). Remarks on diphthongization in Spanish. *Lingua* **41**. 261-305.

Harris, James (1978). Two theories of non-automatic morphophonological alternations. *Language* 54: 41-60.

Harris, James (1985). Spanish Diphthongisation and stress: a paradox resolved. *Phonology Yearbook* **2**. 31-45.

Harris, James (1987). The accentual patterns of verb paradigms in Spanish. *Natural Language and Linguistic Theory* **5**. 61-90.

Hayes, Bruce (in press) Gradient well-formedness in Optimality Theory. To appear in Frank van der Leeuw and Jeroen van de Weijer, eds., *Conceptual Studies in Optimality Theory*. Oxford: Oxford University Press. [http://www.humnet.ucla.edu/linguistics/people/hayes/gradient.htm]

Hochberg, Judith (1987). Learning Spanish stress: developmental and theoretical perspectives. *Language* **64**. 683-706.

Hooper, Joan (1976). *An Introduction to Natural Generative Phonology*. New York: Academic Press.

Ito, Junko, and Armin Mester (1995a). Japanese phonology. In *Handbook of Phonological Theory*, ed. by John Goldsmith. Cambridge, MA: Blackwell. 817-838.

Ito, Junko, and Armin Mester (1995b). The core-periphery structure of the lexicon and constraints on reranking, in Jill Beckman, Suzanne Urbanczyk, and Laura Walsh Dickey (eds.) *University of Massachusetts Occasional Papers in Linguistics* Vol. 18: *Papers in Optimality Theory*. GLSA, Amherst, pp. 181-209.

Kernan, Keith T. and B. G. Blount (1966). The acquisition of Spanish grammar by Mexican children. *Anthropological Linguistics* **8:9**. 1-14.

Malkiel, Yakov (1984). Rising diphthongs in the paradigms of Spanish learned *-ir* verbs. *Hispanic Review* **52:3**. 303-333.

McCawley, James (1968). *The phonological component of a grammar of Japanese*. The Hague: Mouton.

Mikheev, Andrei (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics* **23**. 405-423.

Penny, Ralph J. (1991). *A history of the Spanish language*. Cambridge: Cambridge University Press.

Pinker, Steven and Alan S. Prince (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**. 73-193.

Pinker, Steven and Alan S. Prince (1994). Regular and irregular morphology and the psychological status of rules of grammar. In Susan D. Lima, Roberta L. Corrigan, and Gregory K. Iverson (eds.) *The Reality of Linguistic Rules*. Amsterdam: John Benjamins.

Prasada, Sandeep and Steven Pinker (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* **8**. 1-56.

Prince, Alan, and Paul Smolensky (1993). *Optimality Theory: Constraint interaction in generative grammar*. Rutgers University Center for Cognitive Science Technical Report 2.

Reyes, Rogelio (1978). *Studies in Chicano Spanish*. Bloomington: Indiana University Linguistics Club.

Rumelhart, David and James McClelland (1986). On learning the past tenses of English verbs. In David Rumelhart, James McClelland and the PDP Research Group (eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.

Schuldberg, Howard Kelly (1984). Diphthongization in Spanish verbs. *Hispanic Linguistics* **1:2**. 215-228.

Sebastián, N., F. Cuetos, and M. Carreiras (Forthcoming). LEXESP: Creación de una base de datos informatizada del español. DGICYT: APC93-0122

Skousen, Royal (1989). *Analogical modeling of language*. Dordrecht: Kluwer Academic Publishers.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, **87:2**. 245-251.

Zubin, D.A. and K.-M. Köpcke (1981). Gender: A less than arbitrary category. *Chicago Linguistic Society* **17**. 439-49.

Zuraw, Kie (2000). *Exceptions and Regularities in Phonology*. Doctoral dissertation, UCLA, Los Angeles, Calif.

Department of Linguistics
UCLA
Los Angeles, CA 90095-1543

Albright: aalbrigh@ucla.edu
Andrade: argelia@ucla.edu
Hayes: bhayes@humnet.ucla.edu