A Quantitative Study of Spanish Paradigm Gaps

Adam Albright
University of California, Santa Cruz

1. Introduction

Paradigm gaps pose an interesting paradox in the generative capacity of native speakers. On the one hand, inflectional morphology tends to be automatic and prolifically productive, even for rare or made up words (Berko 1958; Bybee and Moder 1983). Once in a while, however, inflection fails for particular existing words. Pinker (1999) points out that even an inflectionally simple language like English has verbs for which many speakers cannot confidently produce a past tense form, such as *forgo* or *bespeak*.

Similar examples can be found in many languages, including Spanish. The 1sg present indicative is virtually always marked in Spanish by the suffix *-o*, accompanied in some verbs by an additional change in the root vowel, or by insertion of a velar stop before the suffix: *cant-ar/cant-o* 'sing-infin/1sg', *viv-ir/viv-o* 'live-infin/1sg', *cont-ar/cuent-o* 'count-infin/1sg', *sal-ir/sal-g-o* 'leave-infin/1sg'. In general, Spanish speakers have no trouble producing 1sg forms for rare or even made-up verbs (Albright, Andrade and Hayes 2001). For a handful of existing verbs, however, there is no 1sg present form, and all possible outcomes are deemed unacceptable. For example, for the verb *abolir* 'abolish', speakers are typically unsatisfied with any possible 1sg form (**abol-o*, **abuel-o*), and likewise for *as-ir* 'grasp' (**as-o*, **as-g-o*).

In both Spanish and English, there is no apparent semantic reason why these particular forms should not exist, and for this reason, I will refer to the phenomenon as *arbitrary lexical paradigm gaps* (though I will ultimately argue that they are neither arbitrary nor lexical).[1] As Hetzron (1975) observes,

---

1. This phenomenon has gone under many names in the literature. A traditional term for such words is *defective*, but this fails to distinguish between forms that are missing for purely semantic/syntactic reasons (such as of impersonal verbs), and those with morphophonological difficulties. Fanselow and Féry (2002) and others adopt the more general term *ineffability* to refer to all cases in which the grammar fails to produce a usable output in syntax, morphology, or phonology. I use the term *arbitrary*

such gaps are especially puzzling because speakers can generally say what the form would be *if it existed*, but then reject it as awkward or unacceptable.

Arbitrary lexical paradigm gaps raise a number of empirical and theoretical questions. Are forms like *forwent* or *abolo* categorically ungrammatical, or are they merely degraded? Are gaps a sporadic phenomenon affecting isolated words, or do the words mentioned in dictionaries represent one extreme of a gradient range of uncertainty? Do speakers generally agree on which words suffer from gaps? What are the factors that create uncertainty, and how should they be captured theoretically? Should all cases of gaps be analyzed in the same way, or can grammars fail in a variety of different ways?

For such a curious phenomenon, paradigm gaps have attracted surprisingly little attention in the literature. When they are discussed at all, it is generally with the hope that all cases can be described with single theoretical device, such as filters or inviolable constraints (Halle 1973; Orgun and Sprouse 1999; Fanselow and Féry 2002). In this paper, I make a rather different starting assumption that arbitrary lexical paradigm gaps may be caused by a variety of factors, and that different causes for uncertainty may correspond to different types of grammatical failure. Thus, detailed studies of individual cases are needed before we can decide whether or not all cases may be subsumed under a single analysis.

This paper has three goals. The first goal is an empirical one, to provide more detailed data about paradigm gaps in one particular language (Spanish). To this end, in §2, I give an overview of the relevant Spanish verbal morphology and the dictionary description of gaps, and in §3, I present quantitative data from a production experiment on potentially problematic 1sg forms. The results show that uncertainty is gradient, is relatively consistent across speakers, and is apparently not limited to a particular closed class of lexical items. Next, in §4, I consider the factors that cause uncertainty, concluding that 1sg gaps in Spanish are due to a combination of unfamiliarity with the lexical item, and uncertainty about whether to apply morphophonological alternations. Finally, in §5, I ask whether lexical paradigm gaps in Spanish are amenable to the same type of analysis that has been proposed in other cases. I argue that the Spanish data demands a different type of analysis, in which uncertainty arises within the derivation itself (Hetzron 1975).

2. Overview of Spanish present tense forms

2.1. Spanish present tense morphology

Spanish verbs fall (roughly) into three conjugation classes, defined by the vowel that occurs in the present tense ([a], [e], or [i]), as shown in (1). Of

---

lexical paradigm gap, following Hetzron (1975), to emphasize that I am considering cases where only some words are affected.

particular interest here is the fact that the 1sg suffix is the same in all three classes (*-o*). Stress falls on the root in some parts of the paradigm (1,2,3sg and 3pl), and on the suffix otherwise.

(1)   Three conjugation classes

Class 1: [a]		Class 2: [e]		Class 3: [i] (~ [e])	
'to speak'	*hablar*	'to eat'	*comer*	'to live'	*vivir*
hábl-o	*habl-**á**mos*	*cóm-o*	*com-**é**mos*	*vív-o*	*viv-**í**mos*
*hábl-**a**s*	*habl-**á**is*	*cóm-**e**s*	*com-**é**is*	*vív-**e**s*	*viv-**í**s*
*hábl-**a***	*hábl-**a**n*	*cóm-**e***	*cóm-**e**n*	*vív-**e***	*vív-**e**n*

In addition to person and number suffixes, the present tense paradigms of many verbs exhibit unpredictable morphophonological alternations. The most common alternations involve diphthongization or raising of mid vowels in those parts of the paradigm where the root is stressed, as in (2).

(2)   a.   Diphthongization of [e]→[je], [o]→[we]

'to feel'	*sentir*	'to count'	*contar*
*s[**jé**]nt-o*	*s[**e**]nt-**í**mos*	*c[**wé**]nt-o*	*c[**o**]nt-**á**mos*
*s[**jé**]nt-es*	*s[**e**]nt-**í**s*	*c[**wé**]nt-as*	*c[**o**]nt-**á**is*
*s[**jé**]nt-e*	*s[**jé**]nt-en*	*c[**wé**]nt-a*	*c[**wé**]nt-an*

b.   Raising of [e]→[i]

'to request'	*pedir*
*p[**í**]d-o*	*p[**e**]d-**í**mos*
*p[**í**]d-es*	*p[**e**]d-**í**s*
*p[**í**]d-e*	*p[**í**]d-en*

The mid vowel alternations in (2) display an interesting asymmetry: in class 1 ([a]), diphthongization is a minority pattern, raising is unattested, and non-alternation is the default pattern for novel verbs (Albright, Andrade and Hayes 2001). In class 2 ([e]), on the other hand, diphthongization is more prevalent (though there is still no raising), while in class 3 ([i]), every single mid-vowel verb alternates, either by diphthongizing or by raising under stress.

A second alternation that commonly affects present tense paradigms is the insertion of velar stops ([k] or [g]) in the 1sg. Insertion of [k] occurs exclusively after roots ending in [s]/[θ], while insertion of [g] occurs in a wider variety of environments (*pongo* 'I put', *salgo* 'I leave', *traigo* 'I bring'). Velar insertion also shows an asymmetry: it is limited to classes 2 and 3.

(3)   Velar insertion in the 1sg

Insertion of [k]		Insertion of [g]	
'to grow'	*crecer*	'to leave'	*salir*
*cré[**sk**]-o*	*cre[**s**]-**é**mos*	*sál**g**-o*	*sal-**í**mos*
*cré[**s**]-es*	*cre[**s**]-**é**is*	*sál-es*	*sal-**í**s*
*cré[**s**]-e*	*cré[**s**]-en*	*sál-e*	*sál-en*

2.2. Two types of paradigm gaps

With this background in mind, we may now turn to the "textbook" description of Spanish paradigm gaps (de Gámez 1973; Butt 1997). Traditional sources distinguish between two types of present tense gaps: the first, which I will call ANTI-STRESS VERBS, lack all forms in which stress would fall on the root (4a). The most commonly cited verb of this type is *abolir* 'to abolish'; others claimed to exhibit this pattern include *agredir* 'assault', *aguerrir* 'harden for battle', *arrecirse* 'stiffen', *aterirse* 'be numb', *colorir* 'color' (=*color(e)ar*), *denegrir* 'blacken', *descolorir* 'de-color' (=*descolorar*), *empedernir* 'harden', *garantir* 'guarantee' (=*garantizar*), *transgredir/trasgredir* 'transgress', *trashumar* 'move pastures'.

The second type of gap consists of verbs lacking just the 1sg; I call these ANTI-EGOTISTIC VERBS (4b), borrowing from the literature on Russian, which has a similar phenomenon (Halle 1973). Some verbs that are claimed to be anti-egotistic include *asir* 'grasp', *balbucir* 'stammer' and *pacer* 'graze'.

(4) a. Anti-stress verbs: b. Anti-egotistic verbs:

'to abolish'	*abolir*		'to grasp'	*asir*
—	*abol-imos*		—	*as-imos*
—	*abol-ís*		*as-es*	*as-ís*
—	—		*as-e*	*as-en*

The gap patterns in (4) are extremely suggestive, since they mirror exactly the distribution of unpredictable morphophonemic alternations. In particular, anti-stress verbs are missing forms where diphthongization and raising occur, while anti-egotistic verbs are missing the form where velar insertion occurs.[2] Furthermore, verbs with gaps generally meet the structural description for alternations: anti-stress verbs mostly have mid vowels, while anti-egotistic verbs have stem-final [s]. In addition, virtually all of the defective verbs belong to class 3 ([i]), which is most susceptible to alternations.

It is also worth noting that most of the verbs that are claimed to have gaps are rare or archaic (including *abolir* itself). Many are being replaced by doublets in class 1 (the productive, default class): *balbucir* ⇒ *balbuc**e**ar*, *garantir* ⇒ *garan**ti**zar*, *cocer* ⇒ *coc**in**ar*, etc. Still other verbs in this list are simply falling out of use as inflected verbs. These facts beg the question of whether hesitance to produce inflected forms might merely be due to the fact that the words are unfamiliar. This argument has the potential to be circular (are speakers uncertain because the verbs are rare, or are the verbs rare

---

2.   Additional corroborating evidence comes from the fact that the entire present subjunctive paradigm exhibits both velar insertion and mid vowel alternations, and all verbs with paradigm gaps are missing the present subjunctive as well.

because speakers cannot inflect them?), but it raises an important methodological point: we must be careful to distinguish between uncertainty that is due to not knowing the word and uncertainty due to other causes.

Another possible pitfall when looking only at the dictionary description of paradigm gaps is that we have no guarantee that speakers actually dislike the forms that prescriptive sources say to avoid. A Google search for the form *abole* turns up a small, but non-negligible number of bona fide hits, and one online grammar resource notes that:

> Despite this 'rule,' however, supposedly unacceptable conjugations are used in real life. A recent news story ... stated that "*el presidente ucraniano ha promulgado una ley que abole la pena de muerte*" ("*the Ukranian president has promulgated a law that abolishes the death penalty*").[3]

The upshot is that although the patterns in (4) are intriguing, there remain some fundamental questions about the data. Are gapped forms absolutely ungrammatical, or is uncertainty gradient? Is this a comprehensive list of verbs with gaps, or are there more? And can uncertainty about inflected forms be teased apart from uncertainty about lexical items in general? In the next section, I present an experimental study designed to help answer these questions.

3. A production experiment on Spanish paradigm gaps

3.1. Experimental design and methods

In order to elicit data on potentially gapped forms, I constructed a list of 38 existing Spanish verbs (see Appendix 6). This list included 28 verbs that could potentially take mid vowel alternations, and 10 verbs that could potentially take velar insertion. The list contained some verbs that alternate (at least prescriptively), some that do not, and some that are cited as gapped. The list also contained a balanced mix of low and high frequency verbs (as found in the LEXESP corpus; Sebastián et al. 2000). For example, *cerrar* 'close' is a high frequency verb that prescriptively diphthongizes (*cierro*), *tronzar* 'slice' is a low frequency verb that does not (*tronzo*), *guarecer* 'shelter' is a low frequency verb with velar insertion (*guarezco*), and *ejercer* 'exercise' is a high frequency verb that does not alternate (*ejerzo*, not *ejierzo, *ejerzco).

This list of 38 items was augmented with 22 filler items, some containing an irrelevant ambiguity concerning stem-final glides, and some judged to be implausible in the 1sg (e.g., *italianizar* 'Italianize', *descafeinar* 'decaffeinate', *alechugar* 'curl up like a leaf'). The resulting list of verbs was designed to cause varying degrees of uncertainty, and embody a variety of morphological behaviors.

---

3. http://spanish.about.com/library/questions/aa-q-defective-verbs.htm

As noted above, we must distinguish the uncertainty that speakers feel about the task of inflecting words from the uncertainty they feel about using unfamiliar words. Therefore, before eliciting any judgments about inflected forms, a pretest was conducted to check which verbs the participants actually knew (Nusbaum, Pisoni and Davis 1984). Participants rated the familiarity of verbs (in the infinitive) on a scale from 1 to 7; if a participant gave a verb a score of 3 ("May have seen the word before") or lower, that participant's judgments about the word were excluded from the analysis.

The main portion of the experiment consisted of a fill-in-the-blank production task for potentially gapped forms. Participants were presented with verbs in the infinitive (e.g., *abolir*), and had to use them to complete a simple sentence, such as *Ahora yo* _____ ('Now, I am _____ing'). The sentences elicited 1sg forms for verbs with potential uncertainty about velar insertion, and 3pl forms for verbs with potential mid vowel alternations. After providing an inflected form, participants also gave confidence/certainty ratings for their own production, on a scale of 1 (not at all sure) to 7 (completely certain).

Twenty native Spanish speakers, all with at least some college-level education in a Spanish-speaking country, participated in the study. In addition to discarding responses for which the participant did not know the verb (see above), responses were also discarded if the wrong verb was used, or a related verb from a different class was substituted (e.g., *garantizar* for *garantir*).

3.2. Results

From the remaining responses, two values were calculated for each verb: the between-speaker agreement rate from the production task (1.0 = single consistent output, .5 = evenly split between 2 outputs, etc.), and the mean confidence rating from the ratings task (1=low, 7=high). The results, shown in Fig. 1, show two things. First speakers do not always rate their own productions with the highest confidence (7); in fact, ratings fall along a continuum that approaches a normal distribution (Shapiro-Wilk W Test, $W(36) = .94$, $p < .05$). This suggests that uncertainty about inflected forms is a gradient, not categorical effect. Second, we see that when speakers are less confident, they also tend to disagree with one another ($r(35) = .75$). In other words, when one speaker says "I think it's *abole*, but I'm not sure," another speaker is likely to volunteer *abuele*, but be equally ambivalent.

The results also show that uncertainty is not limited to verbs that dictionaries list with gaps. Among the verbs that received the lowest confidence ratings and the most disagreement, some are listed (e.g., *pacer*, *empedernir*, *aguerrir*), but others are not (e.g., *despavorir*, *discordar*, *distender*, *hender*). There is no reason to think that these particular items are an exhaustive set of the verbs that cause speakers uncertainty in their inflected forms.
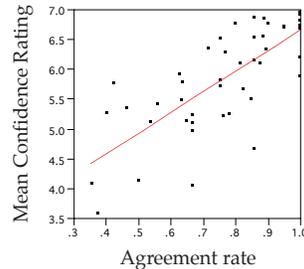
Figure 1: Correlation between confidence and between-speaker agreement.

4. Factors contributing to uncertainty

The results in §3 show two converging types of data that Spanish speakers feel gradient uncertainty about inflected forms. We are now in a position to ask what factors contribute to this uncertainty. In this section, I consider a variety of possible factors, but conclude that only two seem to play a role here: familiarity and uncertainty about irregular morphophonological processes.

4.1. Semantic plausibility

One explanation that is often given for otherwise arbitrary-seeming gaps is that the missing forms would be semantically or pragmatically odd. This factor certainly does play a role in some patterns of defectiveness, such as impersonal verbs (*I behoove*), and one might wonder whether it is also responsible for the unacceptability of forms like *pazco* 'I graze (on grass)'. Implausibility cannot explain the bulk of the Spanish data, however. A number of rather unlikely forms, such as *italianizo* 'I Italianize' or *descafeíno* 'I decaffeinate', were produced consistently and received high confidence values. I see no reason to think that forms meaning 'they abolish' or 'they distend' would be less felicitous than 'I decaffeinate'. Thus, implausibility is not a promising explanation for the Spanish data.

4.2. Homophony avoidance

In a few instances, the expected form in a gapped paradigm is homophonous with another existing word. For example, when speakers reject *abuelo* 'I abolish', they often comment that it is blocked by *abuelo* 'grandfather'. There are many reasons not to accept this explanation, however. First, not all parts of the paradigm would be affected by homophony, so even if *abuelo* happens to mean 'grandfather', there would be no reason to avoid the 3pl *abuelen*, which is not a possible noun form. Furthermore, not all

gapped verbs suffer from potential homophony—for example, *empedierno* and *agriedo* would be unique word forms—so this explanation would not account for all of the data. Finally, and most importantly, there are many cases in which homophony is tolerated: *creo* 'I create'/'I believe', *avengo* 'I avenge'/'I reconcile', *suelo* 'I am used to'/'I pave', etc. (Halle (1973) makes this same point for Russian.) Avoidance of homophony between lexical items is not usually a strong enough force to block inflectional morphology.[4]

4.3. Phonological ill-formedness

A third factor that is often implicated in paradigm gaps is phonotactic ill-formedness. Orgun and Sprouse (1999) argue that certain paradigmatically expected forms in Turkish and in Tagalog are blocked because they would violate phonological well-formedness constraints, and Fanselow and Féry (2002) argue the same for German diminutives. Hetzron (1975) likewise argues that certain Hungarian stem+suffix combinations are blocked because they would lead to illegal clusters—e.g., *csukl* 'hiccup' combined with the potential suffix *-hat* should yield **csuklhat* 'he may hiccup', but this form is impossible due to the [klh] cluster.

Phonotactic pressures do occasionally play a role in Spanish present tense forms. For example, the expected 1sg form of the verb *roer* 'gnaw' is *roo*, but many speakers find this hiatus awkward, and prefer alternatives such as *royo* or *roigo* (or have a gap). In the remainder of the paradigm, the illicit [oo] sequence does not arise, and the expected forms are used: *roes*, *roe*, etc. Might similar pressures be responsible for the other Spanish gaps?

The viability of a phonotactic explanation rests on whether one or both of the available outcomes can be ruled out on general phonological grounds (**abole*, **abuele*). This does not seem likely, however, since both outcomes find numerous parallels in the language ((5a)-(5b)). Thus, unlike Turkish and Tagalog, the forms that are avoided in Spanish do not appear unpronounce-able in any sense. (Halle (1973) makes a similar argument for Russian.)

(5)

a. Diphthongization of o / __l		b. Preservation of o / __l	
3sg Pres	*Gloss*	*3sg Pres*	*Gloss*
suele	'be used to'	controla	'control'
huele	'smell'	viola	'violate'
duele	'be in pain'	inmola	'immolate'
vuele	'fly'	tremola	'flutter'

---

4. Homophony avoidance may indeed play a role in derivational morphology. To give an anecdotal example, I myself am uncomfortable with the noun *commission* when used to mean 'the act of committing' (*commission of a crime*), and it is apparently blocked by more specialized meanings of *commission*. For me, there is no usable noun from the verb *commit* (*?commission*/*?committal*/*?commitment*). Note that this differs from the more familiar phenomenon of lexical blocking (Aronoff 1976), in which a regularly formed word is blocked by an existing irregular synonym. Here, a regularly formed word is blocked by a homophone with a different meaning.
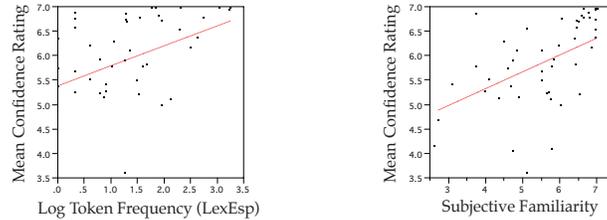
a. Effect of log token frequency
   on confidence ($r(35) = .434$)

b. Effect of subjective familiarity
   on confidence ($r(35) = .462$)

Figure 2: Effect of frequency/familiarity on confidence ratings.

4.4. Frequency or familiarity

In §2, it was noted that many of the verbs listed by dictionaries as gapped are rare or archaic, and that this fact alone might explain some of the hesitance that speakers feel when using them. It should be emphasized from the outset that a familiarity-based analysis cannot explain the more interesting aspects of the Spanish data (such as why only some parts of the paradigm are affected, and why not all rare words are affected) but there does seem to be a relation between certainty and familiarity that bears further investigation.

In order to test the effect of familiarity on certainty, the mean confidence ratings were correlated against two measures: (1) log token frequency, as found in LEXESP (Sebastián et al. 2000), and (2) the familiarity ratings gathered in the pretest. The results, shown in Fig. 2, reveal that confidence in inflected forms does depend to a certain extent on the familiarity of the verb: speakers are uncertain only when the verb is somewhat unfamiliar. This effect is unsurprising, since speakers are more likely to have encountered and stored inflected forms of high frequency words, and would thus be more confident that their production is "correct."[5] The effect is only a modest one, however, and hardly constitutes a complete explanation of the observed uncertainty.

4.5. Uncertainty about irregular morphophonology

The two types of gaps in Spanish ((4) above) suggests a close relation between gaps and morphophonological alternations such as diphthongization and velar insertion. A natural hypothesis is that gaps are due to uncertainty about whether a particular verb should undergo these alternations or not.

---

5.   The extent to which inflected forms are stored has been a topic of controversy in recent years. See Baayen, Dijkstra and Schreuder (1997) and Zuraw (2000) for formal proposals about when speakers are most likely to rely on previously stored inflections.

Before we can test the effect of morphophonological uncertainty, we need an estimate of how likely a given verb is to alternate, based on the behavior of similar words in the lexicon. In order to estimate this, I used a stochastic model of rule induction, developed by Albright and Hayes (2002). This model takes as its input pairs of forms—e.g, *infinitive∼1sg*—and constructs a grammar that derives one from the other by adding or removing suffixes and applying phonological and morphophonological rules. For example, a list of pairs like [ablar]∼[ablo], [komer]∼[komo], and [bibir]∼[bibo] would lead the model to construct rules that creates 1sg forms by taking off the infinitive suffix and attaching *-o*. Verbs with alternations, on the other hand, such as [sentir]∼[sjento], [kontar]∼[kwento], and [salir]∼[salgo], would require more complex rules which also change the root vowel or insert a velar stop. In the case of competing patterns, the model assesses the reliability of each pattern in different phonological environments, at all levels of generality. For example, the model calculates the reliability of velar insertion in general, just after obstruents, just after [s], and so on. The end result is a large set of rules describing all of the processes found in existing words, along with numerical estimates of their reliability in various phonological environments. (For details, see Albright and Hayes (2002).)

The fact that rules are assessed for reliability in this model makes it a good tool for modeling certainty about irregular morphophonology. If a change occurs consistently in a particular environment, then the corresponding rule will have high reliability (approaching 100%). If, however, a change occurs in only half the words in a particular environment, then the rule for this environment will have low reliability (50%).

With this in mind, let us consider why certain environments might cause uncertainty in Spanish. It is instructive to contrast class 1 ([a]) with class 3 ([i]). Class 1 is large (comprising 84% of the verbs in LEXESP), and most verbs in this class form their 1sg by simply suffixing *-o* (3858/4050, or 95%). Thus, context-free *-o* suffixation has a very high reliability for this class, and the regular (non-diphthongized) output can always be produced with certainty, even if the word is unfamiliar. Class 3 ([i]), on the other hand, is much smaller (8% of verbs in LEXESP), and verbs in this class exhibit many more alternations. As mentioned above, every mid-vowel verb in class 3 alternates, some by diphthongizing and some by raising. In addition, for some environments, there is hardly any data at all—for example, there are very few class 3 verbs with root vowel [o] (cf. *oir* 'hear', *morir* 'die', *dormir* 'sleep'). The result is that for an unknown word in class 3, the evidence about what to do is mixed. Therefore, if the speaker cannot resort to a memorized form, the grammar is not very helpful; no high-reliability "default" pattern exists, and in many cases, the more specific local patterns are also unreliable. Multiple possible outputs may be generated, but all with low confidence.
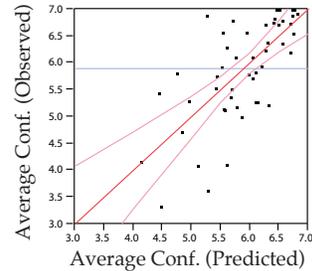
Figure 3: Combined effects of familiarity and morphological confidence.

In order to test the effect of morphophonological uncertainty in creating gaps in Spanish, I trained the model on an input file containing all of the verbs in the LEXESP corpus, deriving the 1sg from the infinitive. The model was trained separately for each conjugation class ([a] vs. [e] vs. [i]). The resulting grammar was then used to derive outputs for all of the experimental items, providing quantitative predictions based solely on likelihood that irregular morphophonological processes should apply. A multiple regression (stepwise, mixed) was then performed, considering three factors as possible predictors predictors of confidence ratings: (1) the model's predicted confidence values, (2) token frequency of the verb, and (3) subjective familiarity. The results show that confidence ratings of inflected forms are best modeled by two factors: the subjective familiarity of the word, discussed in the previous section ($F = 36.2$, $p < .0001$), and the predicted confidence values of the model ($F = 6.7$, $p < .05$). The overall model is shown in Fig. 3.

These results can be interpreted as followed: first, as discussed above, speakers are less confident about inflecting words that they have heard less often or are less familiar with. Although this effect is uninteresting from the point of view of grammatical analysis, it does predict that paradigm gaps should be most prevalent among unfamiliar verbs. More important, even if speakers know a word, they may still be uncertain about whether to apply an irregular morphophonological change. The experimental and modeling results reported here represent an initial attempt to quantify these effects, and confirm that they can play independent roles in determining the certainty with which speakers produce inflected forms.

5. Theoretical implications

These data suggest that paradigm gaps represent one extreme in a spectrum of uncertainty, and that uncertainty about inflected forms is a pervasive phenomenon. In addition, it appears that arbitrary lexical paradigm gaps are neither arbitrary nor lexical. They are a systematic effect, affecting a coher-

ent class of words, and a consistent part of the paradigm. Moreover, they are a grammatical effect, in the sense that they emerge when speakers must synthesize a form, but are uncertain of the outcome. The question remains, however, as to what the most appropriate theoretical analysis is of these facts.

The central question raised by paradigm gaps, identified also by Hetzron (1975) and Fanselow and Féry (2002), is where in the derivation the gapped form fails. In most models of generative grammar, surface forms are derived by taking inputs (underlying forms), applying a grammar of rules/constraints, and pronouncing the output. This leaves (at least) three logically possible loci of failure: (1) the underlying forms of certain words may be defective, (2) the grammar itself may be indeterminate or uncertain, or (3) some external mechanism blocks the output from being pronounced at the surface.

Most formal proposals have focused on surface filters. One of the earliest generative treaments of paradigm gaps is by Halle (1973), who argues that in Russian, as in Spanish, gaps are not blocked by semantic or phonotactic factors, and must thus be due to some other mechanism, such as diacritic marking ([−Lexical Insertion]). Under this proposal, the grammar generates the expected form, but speakers have an additional piece of knowledge telling them not to pronounce it. In a similar spirit, Orgun and Sprouse (1999) and Fanselow and Féry (2002) propose to model gaps in OT with a filtrative CONTROL component blocking certain illicit structures from being pronounced. In cases like Russian or Spanish, where there is no obvious phonotactic or semantic violation, Fanselow and Féry argue for parochial, morpheme-particular constraints (*pazco*, *abuelo*, etc.; p. 278).

These proposals are unsatisfying in many respects. First, they vastly overpredict possible gap patterns. In principle, any form of any word could be marked as [−Lexical Insertion] or eliminated by a parochial constraint, but in fact, only certain forms, such as 1sg, are affected. Second, how are gaps learned, if they require the grammar to contain additional, morpheme-specific statements? What evidence would lead a learner to conclude that certain words *cannot* be used in the 1sg, especially if the words involved are rare to begin with? Finally, blocking or filtrative mechanisms cannot easily account for the gradient nature of the uncertainty associated with gaps.

A different approach, which seems reasonable but has not been pursued, is to attribute paradigm gaps to incomplete or defective underlying forms. The intuition here would be that the lexical entry of *abolir* does not contain enough information to know whether diphthongization should apply to it.[6] Such an approach would also face some explanatory challenges, however, such as why the undiphthongized output is blocked (*abolo*), why the effect is gradient, and why only class 3 verbs are affected.

---

6. Or, perhaps, that diphthongization requires a listed diphthongized alternant, which these verbs lack (*lexical conservatism*; Steriade 1997; Eddington 1996).

The final possibility, advocated here, is that gaps of this type are due to uncertainty within the grammar itself. They are not so systematic that they can be explained by general filters, but they are more systematic than would be predicted by a morpheme-specific analysis. For this reason, they are best explained by a model that incorporates certainty about inflected forms directly into the grammar, such as the Albright and Hayes Minimal Generalization learner. This echoes proposals by Fillmore (1972) and Hetzron (1975), who argue for grammars with indeterminate or irreconcilably conflicting rules.

6. Conclusion

This paper represents an initial step in providing a more systematic account of both the data and the causes of paradigm gaps in Spanish. The overall picture that emerges is that the gaps that are listed in grammars lie at just one extreme of a gradient range of uncertainty that speakers feel when deciding whether or not to apply morphophonological alternations. This uncertainty is strongest when two factors collide: first, the word must be relatively infrequent or unfamiliar, so that the speaker is forced to synthesize a form. In addition, the lexicon must contain conflicting evidence about whether or not the alternation should apply. This scenario is most compatible with a view of morphology and phonology in which speakers have detailed probabilistic knowledge of the reliability of different processes in the lexicon, such as in the Albright and Hayes model of rule induction.

A useful follow-up to this study would include using a more sensitive task to tease apart true gaps from cases where speakers are willing to accept more than one form (free variation), as well as expanding the study to test a wider variety of verbs. In addition, careful cross-linguistic comparison is needed to determine which other cases are due to uncertainty about morphophonological alternations, and which cases are caused by semantics, phonotactics, and so on. It appears that a good deal more research is needed in order to determine whether all cases of gaps or ineffability should really all be treated with the same formal mechanism, or whether different cases demand different types of analyses.

Appendix A. Verbs used in the production experiment

(6)   Verbs with (potential) mid vowel alternations

abnegar 'abnegate'; *abolir* 'abolish'; *adherir* 'adhere'; *aferrar* 'grapple'; *agredir* 'assault'; *aguerrir(se)* 'harden for battle'; *almorzar* 'lunch'; *arbolar* 'hoist a mast'; *cerrar* 'close'; *controvertir* 'dispute'; *desovar* 'spawn'; *despavorir* 'fear'; *discordar* 'disagree'; *distender* 'distend'; *empedernir(se)* 'harden'; *encolar* 'glue'; *erguir* 'straighten'; *forzar* 'force'; *heder* 'stink'; *hender* 'split'; *hervir* 'boil'; *moldar* 'mold'; *podrir* 'rot'; *serrar* 'saw'; *soterrar* 'bury'; *tostar* 'toast'; *tronzar* 'slice'; *trovar* 'sing verses'

(7)   Verbs with (potential) velar insertion

   amortecer 'dull, dim'; *aparecer* 'appear'; *asir* 'grasp'; *balbucir* 'stammer'; *ejercer* 'practice'; *embaír* 'deceive'; *guarecer* 'shelter'; *mecer* 'swing, rock'; *pacer* 'graze'; *yacer* 'lie'

References

Albright, Adam, Argelia E. Andrade and Bruce Hayes. 2001. Segmental Environments of Spanish Diphthongization. *UCLA Working Papers in Linguistics, Number 7: Papers in Phonology 5*, pages 117–151.

Albright, Adam and Bruce 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*.

Aronoff, Mark. 1976. *Word formation in generative grammar*. MIT Press.

Baayen, R. Harald, Ton Dijkstra and Robert Schreuder. 1997. Singulars and Plurals in Dutch: Evidence for a Parallel Dual-Route Model. *Journal of Memory and Language*, 37:94–117.

Berko, Jean. 1958. The child's acquisition of English morphology. *Word*, 14:150–177.

Butt, John. 1997. *Spanish Verbs*. Oxford University Press.

Bybee, Joan and Carol Moder. 1983. Morphological Classes as Natural Categories. *Language*, 59(2):251–270.

de Gámez, Tana, editor. 1973. *Simon and Schuster's International Dictionary: English/Spanish Spanish/English*. Simon and Schuster.

Eddington, David. 1996. Diphthongization in Spanish Derivational Morphology: An Empirical Investigation. *Hispanic Linguistics*, 8:1–35.

Fanselow, Gisbert and Caroline Féry. 2002. Ineffability in Grammar. In Gisbert Fanselow and Caroline Féry, editors, *Resolving Conflicts in Grammars: Optimality Theory in Syntax, Morphology, and Phonology*. Helmut Buske Verlag, Hamburg, pages 265–307.

Fillmore, Charles. 1972. On Generativity. In Stanley Peters, editor, *Goals of Linguistic Theory*. Prentice-Hall, Englewood Cliffs, N.J.

Halle, Morris. 1973. Prolegomena to a theory of word formation. *Linguistic Inquiry*, 4:3–16.

Hetzron, Robert. 1975. Where the Grammar Fails. *Language*, 51:859–872.

Nusbaum, Howard C., David B. Pisoni and Christopher K. Davis. 1984. *Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words*. Research on Spoken Language Processing PR-10. Indiana University, Bloomington, IN.

Orgun, Cemil Orhan and Ronald Sprouse. 1999. From MPARSE to CONTROL: deriving ungrammaticality. *Phonology*, 16:191–224.

Pinker, Steven. 1999. *Words and Rules: The Ingredients of Language*. New York: Basic Books.

Sebastián, Núria, Fernando Cuetos, M. Antònia Martí and Manuel F. Carreiras. 2000. *LEXESP: Léxico informatizado del español. Edición en CD-ROM*. Barcelona: Edicions de la Universitat de Barcelona (Colleccions Vàries, 14).

Steriade, Donca. 1997. Lexical Conservatism. In *Linguistics in the Morning Calm, Selected Papers from SICOL 1997*. Linguistic Society of Korea, Hanshin Publishing House, pages 157–179.

Zuraw, Kie. 2000. *Patterned Exceptions in Phonology*. Ph.D. thesis, UCLA.