Adam Albright (MIT)
albright@mit.edu

Defective Paradigms Workshop
10 April 2008

# Cautious generalization of inflectional morphology, and its role in defectivity

## 1 Introduction: lexically specific paradigm gaps

(1) A lexically specific gap: the past participle of *dive* in American English

- University of Chicago student prospectus, "The life of the mind"[1]

| | |
|---|---|
| Brianna: | You should learn to scuba dive. |
| Rodrigo: | I've never scuba **dived**. |
| Dickie: | I've snorkeled. |
| Rodrigo: | Do you say "scuba **dived**" or "scuba **dove**?" |
| Brianna: | It's "scuba **dove**." |
| Lakshmi: | I would just avoid it altogether and say, "**I went scuba diving**." |
| Rodrigo: | Dickie will just stay on his surfboard conjugating the Marshallese for "relax" |

- Similar examples of uncertainty about *dive* abound
  - "I believe they routinely do dumpster **dived** (**dove**? **doven**? **diven**?) meals"[2]
  - "I try to set out stuff that I need to get rid of in places that I know are well-picked or well-**dived** (or is it well-**diven**? well-**dove**? well-**doven**? help, word geeks)."[3]
- Seemingly parallel verbs cause no such difficulty

| | |
|---|---|
| *slide* ∼ *slid*: | ✓ I've backslid on my diet recently. |
| *drive* ∼ *driven*: | ✓ I've never test-driven a Toyota before. |
| *ride* ∼ *ridden*: | ✓ They have overridden the decision. |

(2) Several factors potentially at work in *scuba dive*

- Headedness: uncertainty as to whether head of phrase is not really *dive*?
  - Parallel to **forgoed*/*forwent*: is *forgoing* a type of *going*? (Pinker 1999)
  - Does not seem likely in this case ('dive' is transparently contained in 'scuba dive')
- Cross-category derivation: is *scuba dive* denominal?
  - V→N→V derivations often fail to inherit irregularity of original V

$$\textit{to}\ [\textit{drive}]_{\mathrm{V}} \rightarrow \textit{a}\ [[\textit{drive}]_{\mathrm{V}}]_{\mathrm{N}} \rightarrow \textit{a}\ [\textit{line}\ [\textit{drive}]_{\mathrm{N}}]_{\mathrm{N}} \rightarrow \textit{to}\ [[\textit{line drive}]_{\mathrm{N}}]_{\mathrm{V}} \sim \textit{line-drived}$$

  - The result of such derivations can be rather awkward, esp. if very transparent
    - ?? "The idea of my truck being **joyrided** around town doesn't sit too well with me."[4]
    - ?? "In 1980, Jimmy Carter survived a challenge by Senator Edward Kennedy for the Democratic nomination, only to be **landslided** by Ronald Reagan"[5]
  - However, no nominal compound *scuba dive*, so doesn't appear to be denominal[6]
- Failure of past participle formation: even if $[\textit{dive}]_{\mathrm{V}}$ is the head, how should it be inflected?
  - "He has **diven**, **dived**, **dove**, **doved**, **doven** or whatever…in every major and most minor bodies of water on this planet."[7]
  - "Because I have **dived** (**deeved**? **diven**? **doven**?) head first into online dating, I frequently chat with folks in whom I have no real interest."[8]

[1] `http://collegeadmissions.uchicago.edu/lifeofthemind` (Accessed 5 Apr 2008)
[2] `http://londoncommons.net/node/4945` (Accessed 5 Apr 2008)
[3] `http://people.tribe.net/alexwebster/blog?page=4` (Accessed 5 Apr 2008)
[4] `http://www.ford-trucks.com/forums/8968-am-i-being-too-uptight.html` (Accessed 5 Apr 2008)
[5] `http://deafdemocrats.wordpress.com/2008/02/03/who-are-the-superdelegates/`
[6] I leave aside the possibility that *scuba dive* is a back-formation from *scuba diver*.
[7] `http://www.scubaboard.com/forums/basic-scuba-discussions/4550-whom-would-you-spend-your-day.html` (Accessed 5 Apr 2008)
[8] `http://123valerie.blogspot.com/2007/04/reminders.html` (Accessed 5 Apr 2008)

- I will be concerned here with the last case, in which speakers are uncomfortable with forms of particular words, even in the simplest (underived) morphological contexts
  - The problem with *scuba-dive* reduces to a more basic problem with *dive*

(3) The challenge: explaining how these lexically restricted gaps arise

- As the examples show, there is no semantic reason why 'dive' should not have a past participle
- Furthermore, as is often noted for other cases, there is no phonological reason why the most likely candidate outputs would be unpronounceable (*dived*, *dove(n)*, *diven*)
  - For related observations, see also Albright (2003) for Spanish, Baerman (2008) for Russian
  - Note that many other cases do seem to involve phonological factors (Hetzron 1975; Iverson 1981; Hansson 1999; Orgun and Sprouse 1999; Raffelsiefen 2004; Rice 2007; Rebrus and Törkenczy, in press; and many others)
- Preview of claims:
  - Problem is related to change in past tense *dived* > *dove* in American English
  - Past tense form: competition between regular pattern (*dive* ∼ *dived*) and strong irregular (*dive* ∼ *dove*) led to "irregularization"
  - In principle, might expect equivalent change in participle form: *dive* ∼ *dived* → *dive* ∼ *diven* (cf. *risen*, *ridden*, *written*)
  - Claim: English past participles are formed from pasts, not presents
    - ➠ NOT: "What is the past participle for a monosyllabic verb with [aɪ] in the present?"
    - ➠ Rather: "What is the past participle for a verb whose past has [oʊ], and no suffix?"
  - Unfortunately, the comparison set is small, and the evidence is mixed
  - Gaps arise when speakers have too few relevant data to extract a rule that can be applied with any confidence
- ☞ I.e., this is a corner of the language for which "there is no (usable) grammar"

(4) Outline of the rest of the talk

- English past participle gaps
  - Comparison of past vs. past participle formation suggests that speakers require minimum threshold of data before they are willing to generalize a morphological pattern to new items
  - Echoes findings from rule induction in other (non-linguistic) domains
  - At the same time requires a specific hypothesis about the form of the grammar (speakers are willing to entertain some rules, but not others)
- Non-linguistic excursus: uncertainty in another cognitive domain
- Application of this type of analysis to other cases?
  - Spanish: gaps in stressed forms of present tense verbs (Albright 2003)
    - ➠ As with English, gaps occur in an irregular class with very few relevant examples for learning what the default rule should be for that class
  - Russian: gaps in 1sg of non-past verbs (Halle 1973; Baerman 2008, and others)
    - ➠ Gaps likewise target an irregular class, but data available to speakers does not appear to be so limited
    - ➠ However, although there are relatively many verbs in the class in question, they show many other (orthogonal) irregularities
    - ➠ Claim: speakers are limited in their ability to extract patterns across examples with other ("irrelevant") differences
- Broad conclusion
  - Lexically restricted gaps reflect an interaction of grammatical limitations and general learning principles, which under specific circumstances prevent speakers from learning the rules needed to generalize to unknown forms

## 2 English past participle formation

The data: gaps for past participles of certain [aɪ] ∼ [oʊ] verbs

(5) American English present, past, and past participle forms:

- *dive* ∼ *dove* ∼ ???
  - "I have **doven**, umm **diven**, ahh **dived**, whatever. Yes, I have before."[9]
- *strive* ∼ *strove* ∼ ???
  - At any rate, i have **striven** (**strove**? **stroven**?) to ensure that this journal was not a simple log of how much laundry i did in the day[10]
- *smite* ∼ *smote* ∼ ???
  - All previously established terms and conditions apply, unless of course we get **smote**. . . **smited**. . . **smitten**?. . . eh, whatever, **smacked down** by God first.[11]

(6) Some vowel correspondences in English verb forms

| | Correspondence | Count | Mean Log Freq | Present | Past | Past part. | Also |
|---|---|---|---|---|---|---|---|
| a. | X ∼ X+d ∼ X+d | ≈3000 | 1.70 | tire | tired | tired | try, guide, spy,… (dive) |
| | | | | store | stored | stored | flow, close, pose, … |
| b. | aɪ ∼ aʊ ∼ aʊ | 4–5 | 2.98 | find | found | found | wind, bind, grind (unbind) |
| c. | aɪ ∼ ɪ ∼ ɪ+(ən) | 3 | 2.98 | bite | bit | bitten | hide, light |
| d. | aɪ ∼ ɪ ∼ ɪ | 1–2 | 2.80 | slide | slid | slid | (backslide) |
| e. | aɪ ∼ eɪ ∼ eɪ+n | 1 | 3.73 | lie | lay/laid | lain/laid | |
| f. | aɪ ∼ oʊ ∼ ɪ+(ə)n | 5–8 | 3.42 | drive | drove | driven | write, ride, rise, arise (override, re-write, underwrite) |
| f′ | | 3–4 | 2.21 | strive | strove | striven | stride, smite (dive) |
| g. | aɪ ∼ oʊ ∼ oʊ | (1) | 2.79 | shine | shone | shone | (outshine) |
| h. | eɪ/ɛ ∼ oʊ ∼ oʊ+(ə)n | 6–8 | 3.23 | break | broke | broken | wake, wear, tear, bear, swear (awake, forbear) |
| i. | iː ∼ oʊ ∼ oʊ+(ə)n | 4–6 | 3.02 | speak | spoke | spoken | steal, freeze, weave (interweave, bespeak) |

- Types (a)–(g) have present tense vowel [aɪ]
- Types (f)–(i) have past tense vowel [oʊ]
  - Type (f′) has uncertainty in the past participle, but is historically like type (f)

(7) The role of frequency in encouraging change

- In general, words with gaps have overall lowish frequency
  - *dive*, *stride*, *strive*, *smite* occur about as often in CELEX as *shudder*, *exert*, *dazzle*, *envisage*, *flutter*, *envy* (log lemma frequency ≈ 2.3)[12]
  - By comparison, *write*, *ride*, *rise*, *drive* have frequencies an order of magnitude higher (log lemma freq ≈ 3.3–3.9)
- Thus, speakers are quite likely to need to infer/construct past tense forms for these words
  - Real-life "wug test" (Berko 1958)
- Plausibly the reason why *dive* was able to change in American English (*dived* > *dove*)
- Presumably also why past participle formation is an issue for some verbs and not others
  - *driven*, *risen*, *spoken*, etc. are all high enough frequency to be memorized reliably
- For related observations concerning Modern Greek, see Sims (in press)

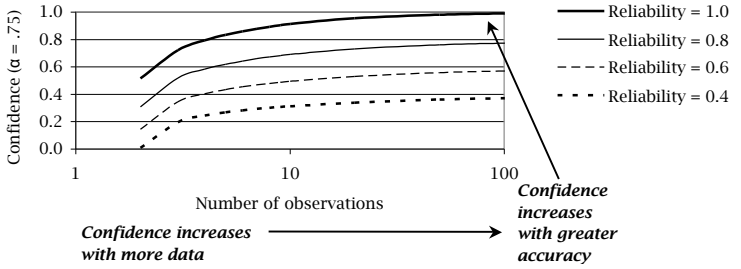
[9] `http://www.ironmagazineforums.com/open-chat/51998-scuba-diving.html` (Accessed 5 Apr 2008)
[10] `http://www.frogtoggle.com/?p=34` (Accessed 4 Apr 2008)
[11] `http://www.whereintheworldisjesus.com/` (Accessed 4 Apr 2008)
[12] Note that these counts reflect written frequencies of British English. According to my own subjective intuitions, they somewhat underestimate the frequency of *dive* and overestimate the frequency *stride*, *strive*.

(8) The role of phonological form in influencing the outcome of change

- In general, low frequency leads to REGULARIZATION, not gaps
  - *melt* ∼ †*molt* > *melted*; *climb* ∼ †*clomb* > *climbed*; etc.
- The IRREGULARIZATION of the past tense form of *dive* over time in American English (†*dived* > *dove*) is presumably related to the fact that it fits within a small but coherent set of [aɪ] ∼ [oʊ] verbs
  - aɪ → oʊ / # (C) $\begin{bmatrix} +\text{voi} \\ +\text{cor} \end{bmatrix}$ —— $\begin{bmatrix} +\text{voi} \\ -\text{son} \end{bmatrix}$#
  - This context includes 5–6 existing [ɑɪ] ∼ [oʊ] verbs (*drive, rise, ride, strive, stride*; variably *thrive*)
  - It includes just 5 verbs that take other patterns, but most of them are quite rare (*slide, writhe, glide, bribe, prize*)
- An ISLAND OF RELIABILITY for [ɑɪ] → [oʊ] (Albright 2002; Albright and Hayes 2003)
  - Phonological context in which a relatively large proportion of words undergo the same morphological change
  - As this example shows, the pattern need not be exceptionless—just relatively trustable

(9) Two factors that influence trustability

- Accuracy: aɪ → oʊ / # (C) $\begin{bmatrix} +\text{voi} \\ +\text{cor} \end{bmatrix}$ —— $\begin{bmatrix} +\text{voi} \\ -\text{son} \end{bmatrix}$# works most of the time
  - Works in 5/10 cases, or better (e.g., for a child who doesn't know *writhe* or *bribe*)
  - Albright and Hayes call this RELIABILITY
- Number of examples
  - The more supporting examples, the stronger the pattern appears to be
  - E.g., past tense aɪ → oʊ / str —— $\begin{bmatrix} -\text{son} \\ +\text{voi} \end{bmatrix}$ works perfectly for the two verbs it covers (*strive, stride*), but that's not a lot of data to go on
  - Albright and Hayes call this CONFIDENCE, and model it with lower confidence limits




- Taken together, these factors require that learners formulate rules that stick close to the observed phonological contexts, and prefer lots of data
  - ☞ A cautious inductive procedure (Albright and Hayes (2003): MINIMAL GENERALIZATION)

(10) Upshot: formalization of a very simple intuition

Speakers are confident in extending a morphological pattern to the extent that…

- It appears in sufficiently many existing words
- There is a phonological difference between those words and words that take other patterns (i.e., they constitute an island of reliability)

(11) How many examples is enough to feel confident enough to extend a pattern?

- Present [ɪ] ∼ past [ʌ] before velars or nasals
  - A salient pattern, seen in 16 verbs, mostly with following [ŋ] (*swing, cling, spring, fling, shrink, sting, wring, sling, stink, slink, string* (*hamstring*))
  - A handful with just nasal (*win, spin*) or just velar (*stick, dig*)
  - Moderately productive generalization to novel words, in the right phonological context

| | | | |
|---|---|---|---|
| | stɪn, splɪŋ | → | stʌn (19%), splʌŋ (32%) |
| vs. | blɪg, plɪm, glɪp | → | blʌg (8%), plʌm (7%), glʌp (3%) |

  - Sporadic historical examples: *strung, snuck*
- Shortening: present [iːd], [iːp] ∼ past [ɛd], [ɛpt]
  - 7/12 [iːp] verbs in CELEX show this pattern, esp. after sonorants (*keep, sweep, leap, weep, creep, sleep* (*oversleep*); but not *heap, seep, reap, peep, steep*)
  - 7/13 monosyllabic [iːd] verbs in CELEX have shortening (*read, lead, bleed, breed, feed,* and optionally *speed, plead*; but not *need, heed, weed, bead, cede, seed*)
  - Rates of irregular responses to novel verbs (Albright and Hayes 2003):

| | | | |
|---|---|---|---|
| | fliːp, gliːd, kwiːd | → | flɛpt (23%), glɛd (24%), kwɛd (21%) |
| vs. | tiːp, priːk | → | tɛpt (5%), prɛk (0%) |

  - Sporadic historical change: *pled, sped* are innovative
- Present [aɪ] ∼ past [oʊ], esp. before voiced obstruents
  - As noted above, 10–14 or so existing verbs, including 5–6 that fall into coherent set before voiced obstruents
  - Occasionally generalized to novel verbs

| | | | |
|---|---|---|---|
| | baɪz, daɪz, skraɪd | → | boʊz (23%), doʊz (15%), skroʊd (19%) |
| vs. | draɪs, raɪf, staɪr | → | droʊs (5%), roʊf (9.8%), stoʊr (5%) |

  - Sporadic historical examples: *dove*
- Null affixation after lax vowels+*t*
  - 12/31 verbs in CELEX ending with [{ɛ,ɔ,ʌ}t] form their past by null affixation (*let, set, cut, shut, bet, upset, wet* (*undercut, beset, offset, inset, reset*))
  - Many exceptions, but as above, many are very low frequency (*get, forget, sweat, regret, fret, putt, gut, strut, vet, pet, butt, beget, abet, silhouette, whet, rut, bayonette, glut, jet*)
  - Also moderately generalizable

| | | | |
|---|---|---|---|
| | flɛt | → | flɛt (30%) |
| vs. | glɪt, drɪt | → | glɪt (14%), drɪt (7%) |

☞ Assessment: patterns with at least a half dozen examples, forming a phonologically coherent set that admits up to about the same number of exceptions, can be generalized

(12) By contrast, patterns with just 2–3 members are typically not generalized

- Present [eɪ] ∼ past [ʊ]
  - Seen in just three verbs (two common, one rare): *take, shake, forsake*
  - No generalization: novel verb [tʃeɪk] never volunteered as [tʃʊk] in past
- Likewise, [ɛ] ∼ [o]+d
  - Seen in just two verbs, both very frequent: *tell, sell*
  - Never volunteered for novel verbs [grɛl], [snɛl]
- Similarly, [ɛ] ∼ [a]
  - Found in one common and one rare verb: *get, tread*
  - Never extended to novel verbs [gɛz], [tɛʃ]

(13) Local conclusion

- Modest lower bound on amount of data needed to support morphological generalization
- Evidence from English suggests roughly half dozen items, falling into a coherent phonological description in which they constitute the majority
  - Precise threshold may vary depending on the setting/task
  - Seems likely that speakers are more willing to hazard forms for nonce words in experimental wug test than for real words in more natural setting
  - Other factors no doubt play a role as well (loanwords, deliberate archaism[13], etc.)
- Patterns with too few examples are not extended, regardless of phonological consistency

(14) So how does this account for uncertainty about past participles?

- In principle, if [aɪ] ∼ [oʊ] is sufficiently robust to be extended to past tense forms, one should expect a parallel pressure for irregularization in the past participle
  - Past: aɪ → oʊ / # (C) $\left[\begin{array}{c}+\text{voi}\\+\text{cor}\end{array}\right]$ — $\left[\begin{array}{c}+\text{voi}\\-\text{son}\end{array}\right]$# (*drove, rose, rode…*)
  - Past participle: aɪ → ɪ +ən / # (C) $\left[\begin{array}{c}+\text{voi}\\+\text{cor}\end{array}\right]$ — $\left[\begin{array}{c}+\text{voi}\\-\text{son}\end{array}\right]$# (*driven, risen, ridden…*)
  - Both cover 5–6 out of the 8–10 verbs meeting this phonological description
- Surprisingly, speakers are actually more likely to say [doʊvən] (reluctantly)
- There are no existing verbs with present tense [aɪ] → past participle [oʊ]+(ə)n !!
- However, there are several words with past tense [oʊ] → past participle [oʊ]+(ə)n, including some that are phonologically similar to *dove*
  - *wove* ∼ *woven*; *froze* ∼ *frozen*; *stole* ∼ *stolen*; *woke* ∼ *woken*; *tore* ∼ *torn*

(15) Forming past participles from past tense forms

- If we assume that past participle formation is based on the past, rather than the infinitive/present stem, a different picture emerges
- Looking back at the table in (6), we see that there are several competing past participle forms for past tenses in [oʊ]
  - Change to ɪ+(ə)n
    - Three or four in phonologically coherent set: *drove*, *strove*, *rose* (*arose*) → *driven*, *striven*, *risen* (*arisen*)
    - Among these, *striven* is so infrequent that speakers may hardly have encountered it
    - This leaves just 2–3 solid examples
  - Change to oʊ+(ə)n
    - Two with voiced fricatives (*froze*, *wove* → *frozen*, *woven*)
    - Two with [k] (*spoke*, *broke* → *spoken*, *broken*)
    - Four with [r] (*wore*, *tore*, *bore*, *swore* → *worn*, *torn*, *born*, *sworn*)
  - Identity (just one verb: *shone* ∼ *shone*)
- A wug-test based on too little data
  - "Should the participle of *dove* be like *drove*, *rose* (→ *diven*), or like *wove*, *froze* (→ *doven*), or like *shone* (→ *dove*)?"
  - "Should the participle of *smote* be like *wrote* (→ *smitten*) or like *spoke*, *broke* (→ *smoten*) or like *shone* (→ *smote*)?"
  - Although these questions involve multiple choices, the crucial factor in creating the gap is the fact that none of the available choices provides sufficient data to be feel confident (Albright 2003, in press a)

[13] Homer Simpson, season 4 episode 3: "Oh Spiteful One, show me who to smite and he shall be smoten!"

(16) Establishing the past → past participle directionality

- This analysis crucially assumes that past participles are "built from" past tense forms
- This claim is at odds with the claim that past and past participle formation are independent of one another, because they do not always share the same phonological readjustments (Embick and Halle 2005)
- It is needed, however, in order to explain the difference between the two forms
  - Pasts: no gaps, regular present → past mapping ensures productive generalization of regular pattern (or, in specific instances, strong irregular patterns)
  - Past participles: regular present → participle mapping is not sufficient to ensure productive generalization; gaps are associated with irregular past tense forms
- The directionality here is not arbitrary
  - Perhaps supported on syntactic/semantic grounds
  - Even if not: past tense form is highly predictive of past participle form in English, and is substantially more frequent than it
  - For discussion of how predictability and frequency interact to influence directionality of morphological mappings, see Albright (2008)

(17) Why aren't gaps more widespread in English?

- English past tense forms involve quite a few isolated patterns, limited to just a small handful of verbs each
- If past participle formation is based on past tense forms, why doesn't paucity of data lead to gaps for many verb types?
- Boring, but plausible answer: the more isolated and idiosyncratic the pattern, the higher frequency the verbs involved have to be in order to maintain their irregularity
  - Speakers would be hard-pressed to create new participles from *got* or *took*, but *gotten* and *taken* are so frequent that the issue simply never arises

(18) An aside: could these gaps be learned?

- Daland, Sims, and Pierrehumbert (2007): argue that in some cases, gaps may be learned from frequency distributions over inflected forms
  - In this case, learner would need to notice that *dive*, *strive*, etc. are used less often than expected in the past participle, and reason that therefore the forms must not exist
  - This would not explain why these verbs are used so infrequently in these forms, but could at least account for how gaps propagate and are encoded
- A very rough test of this idea
  - Took all verb forms from CELEX for which infinitive, past, and past participle are orthographically distinct[14]
    - ➠ Note that CELEX does not distinguish different uses of the past participle, so counts are inflated by adjectival uses
  - Compared frequencies of infinitive, past, and past participle, testing whether particular verbs deviated significantly from average (using $\chi^2$, to be generous)

| | Infin. | Past | Part. | Diff. sig. ($\chi^2$) |
|---|---|---|---|---|
| Avg. | 431 | 277 | 526 | — |
| *stride* | 9 | 32 | 0 | $p < .001$ |
| *strive* | 18 | 11 | 12 | n.s. ($p = .232$) |
| *swim* | 69 | 32 | 21 | $p < .001$ |
| (hypothetical) | 5 | 3 | 0 | n.s. ($p = .052$) |

[14]This is necessary because CELEX combines frequency counts for orthographically identical forms of a single lemma, so it is not possible to estimate past and past participle frequencies independently when they are homophonous.

  - Some gapped verbs (like *stride*) do indeed appear less often than expected in the past participle (though *stride* is also rather infrequent in the present?)
  - However, other verbs do actually occur in the "gapped" form approximately as often as expected, yet speakers evidently resist them[15]
  - Furthermore, many non-gapped verbs also happen to occur less often than expected in the past participle (e.g., *swim*), so low frequency alone evidently does not create uncertainty
  - Most important: for hypothetical rare verbs that are not used in past participle, but are also infrequent in other forms, small numbers make it difficult to be confident that the lack of participle forms is significant
- Thus, an account based solely on frequency distributions overpredicts the possibility of gaps in frequent words, and predicts that gaps should tend not to arise in rare words
  - Yet at least in English (and Spanish), forms that develop gaps tend to be overall rare (English *smite*, Spanish *abolir*)
  - A purely frequency-based account fails to capture the observe relation between gaps and low frequency
  - As we'll see immediately below, I'm not arguing that there could be no role for learning about distributions; merely that the challenge of knowing how to apply morphophonological patterns plays a major role in explaining where in the paradigm gaps occur, and which words suffer from them

# 3 A non-linguistic parallel: number concept games (Tenenbaum 2000)

(19) Tenenbaum (2000): investigated how people generalize based on sets of examples of numbers

- Premise: give subjects a small set of numbers, and see what they think the pattern might be by asking them to rate probability that other numbers would be in the set
- Example:
  - Present set {6,10,4,2}, then ask for ratings of other numbers, like 7, 8, or 57
  - Subjects rate 8 as very probable (hypothesis is presumably "even numbers", "even number $\leq$10", or something similar)
  - Subjects rate 7 as relatively improbable (compatible with hypothesis "integers $\leq$10", but doesn't seem so likely given that particular set of examples)
  - Subjects rate 57 as very improbable
- Different distributions of numbers support different types of hypotheses

| Set | A likely hypothesis | Type |
|---|---|---|
| {16, 12, 13, 17} | Numbers from 10–20? | (ranges) |
| {16,6,2,14} | Even numbers, perhaps $<$20? | (even/odd) |
| {16,2,8,4} | $2^n$ | (powers) |
| {16,8,40,24} | $8\times n$ | (multiples) |
| {16,7,51,23} | arbitrary list? | (rote) |

(20) Two ways that amount of data affects generalization

- Pathologically little data $\rightarrow$ uncertainty
  - Present: {16}. Ask: is 30 in the set?
  - Subjects in this condition are relatively agnostic about any number beyond what they've seen (though slightly more inclined to guess that even numbers, powers of 2, etc. might be included)
  - Even just four numbers is generally enough to make subjects confident enough to assign high probability to novel numbers (see examples above)

[15] Caveats concerning possible differences between CELEX counts on British English vs. my American intuitions are in force.

- Suspicious coincidence of multiple occurrences
  - {16, 23, 19, 20}: set seems to be numbers between ≈15–25
  - {16, 23, 19, 20, 20, 19, 16, 23, 23, 23, 16, 20}: just those four numbers?
  - Probability distribution provides implicit negative evidence that some numbers aren't occurring (Daland, Sims, and Pierrehumbert 2007)

(21) Lexically restricted paradigm gaps may well involve both factors

- Analysis in preceding section has focused on the role of very few types
  - Given just a couple examples, it's tough to evaluate distinguish between a trend and a coincidence
  - E.g., is the fact that *drove*, *rose*, and *rode* end in voiced obstruents significant?
- Daland, Sims, and Pierrehumbert (2007) focus on what can be inferred from distributions when many tokens are available
  - The fact that *dive* is moderately frequent may indeed accentuate the **diven*/**doven*/**dove* gap, making it more acute than **striven*/**stroven*/**strove*
  - However, I'm not really able to evaluate this (given that etymologically expected tokens of *striven* do actually occur, so this alone might make the form seem better)

(22) Local summary

- Factors that influence willingness to generalize morphological patterns appear to play a role in inductive inference more generally
  - See also recent work by LouAnn Gerken involving generalization from small numbers of examples in infant language tasks
  - These tasks have narrowed in on 3–4 examples as sufficient in experimental settings; data above suggests maybe around twice that in the real world, at least for morphological patterns involving phonologically similar groups of words
- However, also a vital role for grammar in the account
  - It is certainly not the case that English speakers have just 6–8 verbs at their disposal
  - The restriction to consider only evidence from other strong verbs with [oʊ] in the past tense when trying to create a past participle requires a very specific grammatical architecture

# 4 Application to Spanish and Russian

(23) Goal of this section:

- Consider to what extent these factors seem relevant in explaining the distribution of gaps in two other well-known cases (Spanish, Russian)

## 4.1 Spanish 1sg present gaps

(24) Morphophonologically irregular mid-vowel alternations in Spanish present tense forms

a. Diphthongization of /e/, /o/

i. *sentar* 'seat'

| | |
|---|---|
| s[**jé**]nt-o | s[**e**]nt-ámos |
| s[**jé**]nt-as | s[**e**]nt-áis |
| s[**jé**]nt-a | s[**jé**]nt-an |

ii. *contar* 'count'

| | |
|---|---|
| c[**wé**]nt-o | c[**o**]nt-ámos |
| c[**wé**]nt-as | c[**o**]nt-áis |
| c[**wé**]nt-a | c[**wé**]nt-an |

b. Raising of /e/

| i. *pedir* 'request' | |
|---|---|
| p[**í**]d-o | p[**e**]d-ímos |
| p[**í**]d-es | p[**e**]d-ís |
| p[**í**]d-e | p[**í**]d-en |

c. Neither

| i. *rentar* 'rent' | |
|---|---|
| r[**é**]nt-o | r[**e**]nt-ámos |
| r[**é**]nt-as | r[**e**]nt-áis |
| r[**é**]nt-a | r[**é**]nt-an |

| ii. *montar* 'mount' | |
|---|---|
| m[**ó**]nt-o | m[**o**]nt-ámos |
| m[**ó**]nt-as | m[**o**]nt-áis |
| m[**ó**]nt-a | m[**ó**]nt-an |

- Non-alternation is the default for *-ar* and *-er* verbs
- Raising is confined to *-ir* verbs
- Nearly all *-ir* verbs alternate, either by raising or by diphthongization[16]

(25) Gaps in stressed forms:*abolir* 'to abolish' (Butt 1997, p. 185)

| | |
|---|---|
| — | ab[**o**]l-ímos |
| — | ab[**o**]l-ís |
| — | — |

- All stressed forms are avoided (including also present subjunctive)
- These are the forms that would potentially involve diphthongization: *ab[**ó**]le, *ab[**wé**]le

(26) Some observations about gapped verbs

- Gaps in stressed forms are confined to the *-ir* class, and affect only verbs with mid vowels
  - Suggestively, this is precisely the set of verbs that always alternates somehow
- Among verbs that are reported to involve 1sg uncertainty, most are low frequency, particularly in inflected forms
  - *abolir* is typical in this respect: not uncommon in the infinitive and past participle (*abolido*), but generally not used in finite forms
- Some affected verbs have switched classes to avoid the problem (*colorir* → *colorear* 'color'), while others have simply died out
- Seems parallel to cases like English *stride* or *smite* (low frequency words, a real-life wug test)

(27) A paucity of data for /o/ verbs in the *-ir* class

- For the back mid-vowel /o/, there are just three *-ir* verbs
  - *dormir* 'sleep' (3sg *duerme*), *morir* 'die' (3sg *muere*), *oír* 'hear' (3sg *oye*)
  - Two show diphthongization, one has different diphthong due to stem-final /i/
- One might expect diphthongization, given that practically all verbs in this class alternate in some way (*abolir* → *abuele*?)
- Or, perhaps non-alternation should be an option, since the majority of verbs in the language are non-alternating (*abolir* → *abole*?)
- As with English, the challenge is to understand why speakers don't employ such general rules, limiting themselves instead (unsucessfully) to *-ir* verbs with stem vowel /o/

[16] An exception is *sumergir* 'submerge', which falls under a separate generalization that roots ending in [erx] never alternate, regardless of their class (Brame and Bordelois 1973).

(28) A paucity of data for /e/ verbs in the *-ir* class?

- There are relatively more /e/ verbs in the -ir class—e.g., counts from LEXESP corpus (Sebastián, Cuetos, Martí, and Carreiras 2000)
  - 32 diphthongizing (*sentir* ∼ *siento* 'feel')
  - 42 raising (*pedir* ∼ *pido* 'ask')
- However, these include many prefixed forms, based on a relatively small number of verb roots
  - Diphthongizing: about 14 roots, nearly all with /en(t)/ or /er(t)/[17]
  - Raising: about 24 roots, clustered around a few phonological forms (/eɲ-/, /e-/, /ed-/), but not many of any particular neighborhood [18]
- Frequency disparities:
  - Some of these verbs are extremely common verbs ('come', 'say', 'ask', etc.)
  - However, many of them are not commonly encountered in inflected forms
  - This is much like the English situation (*rise*, *ride* vs. *strive*, *stride*), and suggests that actual data may be less than counts of verb lemmas imply

(29) What this suggests

- For /o/ verbs in class three, there is certainly not enough data to construct any reliable rules
- For /e/ verbs, there may be almost enough forms to feel confident about alternations in particular phonological contexts (diphthongization /___ rC/, raising /___ ɲ/), but not for more general rules
- The fact that the verbs contain sub-groups of distinct phonological contexts, and are also very frequent, conspire to make learners unwilling to generalize beyond those contexts
- Number concept equivalent: {26, 22, 22, 26, 24, 26, 26, 22, 22, 74, 70, 74, 74, 74, 74, 74, 78, 74}
  - Numbers in 20's and in 70's both support the hypothesis that the set contains even numbers, but the clustering also makes it very likely that numbers in between are not included
- Upshot: speakers are left with very specific rules for this class, but nothing broader that would generalize to verbs like *abolir*

(30) Why this limitation?

- As with English, one wonders why speakers, in desparation, are not able to make use of the broader generalization that as a whole, Spanish verbs tend not to diphthongize
- Albright (2003) attributes this to a cautious inductive mechanism
  - Don't generalize across different phonological contexts or morphological classes unless you have evidence that they truly behave the same
- Conversely, one might imagine that speakers are perfectly willing to generalize across classes, but that Spanish speakers have learned that conjugation class is relevant to vowel alternations, so behave cautiously based on this knowledge (Albright, in press b)
- In either case, the fact that rules must be sensitive to conjugation class information is a feature of the grammar, not some general cognitive learning procedure

[17] *ven-* 'come', *sent-* 'feel', *ment-* 'lie', *(arre)pent-* '(re)pent', *(dis)cern-* '(dis)cern', *diger-* 'digest', *erg-* 'prick up ears', *(re)fer-* '(re)fer',, *her-* 'injure', *adher-* 'adhere', *herv-* 'boil', *(in)jer-* 'insert', *(re)quer-* 'require', *vert-* 'turn'

[18] *dec-* 'say', *eleg-* 'choose', *correg-* 'correct', *exped-* 'dispatch', *fre-* 'fry', *re-* 'laugh', *(des)le-* 'dissolve', *(en)gre-* 'make vain', *ceb-* (-ceive), *gem-* 'moan', *med-* 'measure', *ped-* 'ask', *comed-* 'restrain', *(re)pet-* '(re)peat', *rend-* 'yield', *hench-* 'stuff', *segu-* 'follow', *reñ-* 'scold', *ceñ-* 'cling', *-streñ-* '-strain', *teñ-* 'dye', *vest-* 'wear', *serv-* 'serve'

## 4.2 Russian 1sg non-past gaps

(31) Russian 1sg non-past gaps (Halle 1973; Shvedova 1970; Zalizniak 1977; Daland, Sims, and Pierrehumbert 2007; Baerman 2008, and many others)

- *pobedit'* → 1sg *?*pobežu*, *?*pobeždu*, *?*pobedju*
- Gaps affect 1sg non-past of second conjugation (*-it'*) verbs, primarily of dental-final stems

(32) A suggestive parallel with Spanish

- Gaps target a smaller inflection class, plagued with various morphophonological changes
- Gaps target specifically that part of the paradigm that displays alternations

(33) However, as often noted in the literature, two features make Russian look different from Spanish

- More words
  - The second conjugation contains quite a few verbs, including many dental-final verbs
  - Therefore, it appears that learners have more than just a handful of forms to go on
- Less competition
  - In the current standard language, all dental-final stems alternate (*t*→*č* / *šč*, *d*→*ž*, *s*,*z*→*š*,*ž*)
  - Thus, the entire set of existing words seems to point in the same direction
  - This is unlike, say, Spanish /e/ → [je] (diphthongization) vs. [i] (raising)

☞ I will attempt to argue here that Russian may not actually be so different after all, and that perhaps an analysis rather parallel to Spanish is indeed possible

(34) First step (Baerman 2008): perhaps more competition than current grammars would suggest

- As a historical stage
  - In 18th–19th century sources, competition between several patterns may be seen: 1sg *-žd-* vs. *-d-* vs. *-ž-*
  - The sources that Baerman surveys from times before gaps were widely reported show clear variation between roots and within a single root—e.g.,
    - *xodit'* generally with *-ž -*
    - *-bedit'* typically with *-žd-*, occasionally with *-ž-*
    - *dudet'* generally with *-d-*
- Also synchronically?
  - Many modern speakers also seem willing to accept non-alternating (*-d-*) forms, especially—but not only—for nonce or defective verbs (Alley, Brookes and Sims 2006)
  - Furthermore, 1sg *-žd-* is often offered as a (reluctant) possibility for gapped verbs, presumably due either to exposure to older written sources, or to the fact that *-žd-* occurs in paradigmatically related forms (such as past passive participles)
- Thus, one possibility (endorsed by Baerman 2008) is that gaps may have arisen at a time when 1sg forms in Russian looked a lot more like stressed mid vowels in Spanish *-ir* verbs
- Given the relative timing of attested change in alternations and the rise of gaps, this historical hypothesis seems correct
  - Fracturing of language into a variety of little subpatterns created uncertainty, leading people to avoid 1sg forms of many verbs, and decreasing amount of data available about 1sg formation for subsequent generations
- However, this still leaves us with the fact that in the modern language, quite a few verbs consistently show alternation
  - E.g., there are 66 /d/-final second conjugation verbs in Zalizniak (1977) that occur at least once per million words in a modern corpus (Sharoff 2002)
  - Why do speakers fail to extend this pattern?

(35) Sixty-six Russian verbs are less data than you think (?)

- Among these 66 verbs, there are only 26 unique verb roots represented
  - The rest are prefixed form based on those roots
- Among those 26 roots, not all are common in the 1sg
  - Some words have low-frequency overall (*beredit'* 'irritate', *(na)čadit'*, etc.)
  - Some words are used primarily in 3rd person
- Furthermore, it's not obvious that the "second conjugation" forms a unitary class
  - Most important: different verbs show different stress patterns!

(36) Division of second-conjugation mutating verbs by stress pattern

- Counts from distinct roots in second conjugation verbs listed by Zalizniak, with very rare verbs (<1 per million; Sharoff 2002) removed (Albright, in press b)

| | Fixed | Varying | Total | Percent with stress alternation |
|---|---|---|---|---|
| p | 12 | 2 | 14 | 14% |
| b | 10 | 0 | 10 | 0% |
| m | 9 | 2 | 11 | 18% |
| v | 12 | 2 | 14 | 14% |
| t | 19 | 10 | 29 | 34% |
| d | 19 | 7 | 26 | 27% |
| s | 7 | 3 | 10 | 30% |
| z | 9 | 1 | 10 | 10% |

- For dental-final verbs, rather few roots instantiating any particular combination of final consonant + stress
- Note that even these counts are inflated, since they may contain rare words, or words that aren't used much in the 1sg (viz. above)
- If Russian speakers were looking to "nearby roots" (= same final consonant and same stress pattern) to decide about how to form an unknown 1sg form, they may in fact have little data to go on

(37) Why do stress, voicing, etc. make a difference?

- Same question as above: why would speakers be so cautious about generalizing across different root shapes?
- Suggestion: harder to relate data that differs along multiple dimensions
- Number concept equivalent: {60, 10, 40} vs. {60, 5710, -40}
  - How likely is 70?
  - Both sets support the hypothesis "multiples of 10"
  - In fact, the second should more strongly, since it seems less likely to be a range ("numbers from 0 to 100" vs. "numbers from -100 to 6000")
  - However, other differences (sign, magnitude) are draw attention away from the similarity
- Russian verbs differ from each other in more orthogonal respects than English or Spanish verbs do
- Tentative conclusion: the amount of data from any little pocket of the language may not be a lot greater than in other cases

# 5 Conclusion

(38) Cases of lexically restricted gaps appear to share a number of traits

- Restricted to smaller, more irregular inflection classes where alternations abound
- Typically associated with lower frequency items within those classes
- Affect those parts of the paradigm where alternations are expected

(39) Proposal

- For parts of the language with few relevant examples, or have only high-frequency examples that cluster into a few phonological neighborhoods, speakers are reluctant to infer a more general rule
- "Cautious induction": fit the rules as tightly to the data set as possible
- This may leave speakers with no rules covering words of certain shapes in certain classes
- For lower frequency words, this could be a problem, since speakers are also less likely to have recourse to memorized versions
- Result: gaps or uncertainty

(40) What would help in testing this account more thoroughly?

- Better estimate of the data available to learners
  - Which forms of which lexical items are available to learners in learning how to form a particular part of the paradigm?
- More wug testing and gaps data (Albright 2003; Alley, Brooks & Sims 2006; Sims, in press CLS)
- Artificial grammar experiments testing generalization based on linguistic examples with different distributions, parallel to number concept examples?

# References

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

Albright, A. (2003). A quantitative study of Spanish paradigm gaps. In G. Garding and M. Tsujimura (Eds.), *WCCFL 22 Proceedings*, Somerville, MA, pp. 1–14. Cascadilla Press.

Albright, A. (2008). Explaining universal tendencies and language particulars in analogical change. In J. Good (Ed.), *Language Universals and Language Change*, pp. 144–181. Oxford University Press.

Albright, A. (in press a). How many grammars am I holding up? Discovering phonological differences between word classes. In *WCCFL 26*. Cascadilla Press.

Albright, A. (in press b). Lexical and morphological conditioning of paradigm gaps. In C. Rice (Ed.), *When nothing wins: Modeling ungrammaticality in OT*. Equinox Publishing. `http://www.mit.edu/~albright/papers/Albright-Lexically%20conditioned%20gaps.pdf`.

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

Baerman, M. (2008). Historical observations on defectiveness: the first singular non-past. *Russian Linguistics 32*, 81–97.

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

Brame, M. K. and I. Bordelois (1973). Vocalic alternations in Spanish. *Linguistic Inquiry 4*, 111–168.

Butt, J. (1997). *Spanish Verbs*. Oxford University Press.

Daland, R., A. Sims, and J. Pierrehumbert (2007). Much ado about nothing: a social network model of Russian paradigmatic gaps. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics in Prague, Czech Republic, June 24th-29th, 2007*.

Embick, D. and M. Halle (2005). On the status of stems in morphological theory. In T. Geerts and H. Jacobs (Eds.), *Proceedings of Going Romance 2003*. John Benjamins.

Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic Inquiry 4*, 3–16.

Hansson, G. O. (1999). 'When in doubt…': Intraparadigmatic dependencies and gaps in Icelandic. In P. Tamanji, M. Hirotani, and N. Hall (Eds.), *NELS 29: Proceedings of the 29th meeting of the North Eastern Linguistic Society*, pp. 105–119. Amherst, MA: GLSA Publications. `http://www.linguistics.ubc.ca/People/Gunnar/GH_NELS29_Gaps.pdf`.

Hetzron, R. (1975). Where the grammar fails. *Language 51*, 859–872.

Iverson, G. (1981). Rules, constraints, and paradigm lacunae. *Glossa 15*, 136–144.

Orgun, C. O. and R. Sprouse (1999). From MPARSE to CONTROL: Deriving ungrammaticality. *Phonology 16*, 191–224.

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

Raffelsiefen, R. (2004). Absolute ill-formedness and other morphophonological effects. *Phonology 21*, 91–142.

Rebrus, P. and M. Törkenczy (in press). Covert and overt defectiveness in paradigms. In C. Rice (Ed.), *When nothing wins: Modeling ungrammaticality in OT*. Equinox Publishing.

Rice, C. (2007). Gaps and repairs at the phonology–morphology interface. *Journal of Linguistics 43*, 197–221.

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

Sharoff, S. (2002). Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. In *Proc. of Language Resources and Evaluation Conference (LREC02). May, 2002, Las Palmas, Spain*.

Shvedova, N. J. (Ed.) (1970). *Grammatika Sovremennogo Russkovo Literaturnogo Jazyka*. Moscow: Nauka.

Sims, A. D. (in press). Avoidance strategies, periphrasis and paradigmatic competition in Modern Greek. In J. Blevins and F. Ackerman (Eds.), *Periphrasis and paradigms*. Stanford, CA: CSLI.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. Solla, T. Leen, and K. Muller (Eds.), *Advances in Neural Information Processing Systems 12*, pp. 59–65. Cambridge: MIT Press.

Zalizniak, A. A. (1977). *Grammaticheskij Slovarj Russkogo Jazyka*. Izdatel'stvo Russkij Jazyk.