UCLA Working Papers in Linguistics—no.4, October 2000
Papers in Phonology 4—Albright and Cho (eds.)

DISTRIBUTIONAL ENCROACHMENT AND ITS CONSEQUENCES FOR MORPHOLOGICAL LEARNING

ADAM ALBRIGHT
aalbrigh@ucla.edu

BRUCE HAYES
bhayes@humnet.ucla.edu

We describe a common but neglected pattern of linguistic exceptions, which involve "distributional encroachment." This occurs when the distribution of allomorphs is determined by phonological context, but a few exceptional forms take the "wrong" allomorph. For learning algorithms, this can complicate the task of identifying distributions. We present an algorithm for learning allomorph distributions, then show how it can be modified to handle distributional encroachment.

1. INTRODUCTION

Exceptions create difficulties in rule induction. To formulate a rule, we must "factor out" the exceptions, but the exceptions are not labeled as such in the input data. Thus the questions of what are the rules and what are exceptions must be considered in tandem.

Some exceptions are easily identified: *ring ~ rang* clearly violates the English past tense rule "add -*d*." However, sometimes exceptions "hide" themselves by appearing with an allomorph that is regular in another context. We call this *distributional encroachment.*

For example, consider the forms of the past tense suffix:

(1) a. jump ~ jump[t]
kick ~ kick[t]
miss ~ miss[t]
laugh ~ laugh[t]
b. rub ~ rub[d]
sag ~ sag[d]
tease ~ tease[d]
seem ~ seem[d]
fill ~ fill[d]
c. smell ~ smell[t]
spell ~ spell[t]
dwell ~ dwell[t]
burn ~ burn[t]
learn ~ learn[t]

This suffix is pronounced as [-t] after a voiceless consonant (1a), but as [-d] after a voiced consonant (1b). However, the exceptional forms in (1c), found in various dialects, take [-t] even though they end in voiced consonants. In these forms, [-t] encroaches on the regular context for [–d]. This distributional encroachment is an obstacle to a learner trying to discover the generalization that [-t] occurs only after voiceless consonants.

We present here an algorithm for generating hypotheses about the distribution of allomorphs, then propose a method for identifying correct generalizations in the face of distributional encroachment.

2. "Minimal generality": a bottom-up approach

Following earlier work (e.g. Pinker and Prince 1988), we assume that an effective strategy for morphological learning in the face of exceptions is to explore and evaluate multiple hypotheses. In principle, hypotheses can be explored in order of increasing or decreasing generality. However, since the latter approach requires the entire data set in advance of learning, we consider bottom-up approaches to be more realistic models of human acquisition.[1]

2.1. The Minimal Generality Algorithm

Our algorithm inputs morphologically related pairs, and finds rules which map one to the other. It starts with a single pair and factors it into the change and the context, yielding a rule $A \rightarrow B$ / P__Q. Thus, for English *sip* ~ *sipped* we would obtain (2), which adds [-t] to just *sip*:

(2) $\varnothing \rightarrow$ [-t] / [sɪp] ___]$_{\text{word}}$

By generalizing over rules that have the same change, but different contexts, one obtains rules of greater generality. Thus, given $A \rightarrow B$ / P__Q and $A \rightarrow B$ / P′__Q′, the algorithm locates P_{share}, the maximal right-justified string shared by P and P′, and analogously Q_{share}. The generalized context is formed by retaining shared portions, and replacing residues with variables:

$$\begin{array}{lllllll}
\text{(3) Comparing:} & A \rightarrow B & / & P_{\text{residue}} & P_{\text{share}} & \_\_\_\_ & Q_{\text{share}} & Q_{\text{residue}} \\
\text{with:} & A \rightarrow B & / & P'_{\text{residue}} & P_{\text{share}} & \_\_\_\_ & Q_{\text{share}} & Q'_{\text{residue}} \\
\text{yields:} & A \rightarrow B & / & X & P_{\text{share}} & \_\_\_\_ & Q_{\text{share}} & Y
\end{array}$$

[1] For a top-down approach to learning morphological irregularity, see, for instance, the FOIDL approach of Mooney and Califf (1996) and Dzeroski and Erjavec (1997).

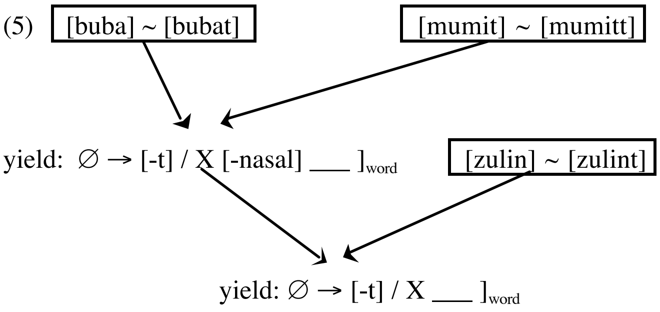
For instance, comparing the *sip ~ sipped* context in (2) with the analogous context for *grip ~ gripped*, one obtains the more general context in (4). In this example, P_{share} is further broken down into shared segments and shared features for the segment immediately preceding the shared segments. Since these forms involve a suffix, Q_{share} is omitted.

(4)

	change	$P_{residue}$	$P_{share/features}$	$P_{share/segments}$	
Comparing:	∅ → [-t] /			sɪp	____
with:	∅ → [-t] /			grɪp	____
yields:	∅ → [-t] /	X	$\begin{bmatrix} \text{+consonantal} \\ \text{+coronal} \\ \ldots \end{bmatrix}$	ɪp	____

Generalization applies iteratively across the entire data set as new forms are learned, generating an ever-larger set of hypotheses. Generalization across heterogeneous forms yields very general contexts, as shown below for a hypothetical language with [-t] suffixation:

(5)


[buba] ~ [bubat]
[mumit] ~ [mumitt]
yield: ∅ → [-t] / X [-nasal] ___]word
[zulin] ~ [zulint]
yield: ∅ → [-t] / X ___]word

2.2. *An evaluation metric for contexts*

Minimal generalization, applied across the entire data set, generates an enormous number of hypotheses. Pinker and Prince (1988, 134), in their outline of this approach, anticipated that many hypotheses could be eliminated, perhaps even reducing the grammar to a single, context-free rule. However, to do this, we need a metric for evaluating which hypotheses are worth keeping and which are not.[2]

[2] There is evidence that human learners discard less than what Pinker and Prince imagined. Psycholinguistic experiments on English (Prasada and Pinker, 1993) and Italian (Albright, 1998) have shown that speakers possess detailed knowledge about the phonological environments in which regular morphological processes apply.

What is the proper metric for evaluating rule contexts? One idea is to use the raw number of forms a context covers: "add [-t] after *ip*" explains 16 forms in our English database, while the more general "add [-t] after *p*" explains 72 forms. However, more general contexts often introduce more exceptions: English has zero irregular verbs ending in *ip*, but six ending in *p*. Therefore, we need a metric which is sensitive not only to the number of forms a context explains, but also the number of forms it includes but does not explain. One such metric is:

(6) *Raw reliability* =

$$\frac{\text{\# of forms where rule applies successfully in context C}}{\text{number of forms that contain C}}$$

Continuing the previous example, the raw reliability of adding [-t] after *ip* is 16/16 = 1, while that of adding [-t] after *p* is only 66/72 = .92.

When raw reliability is equal, we are more certain of patterns that are better attested. Therefore, raw reliabilities are adjusted using lower confidence limits ($\alpha = .75$), following a suggestion of Mikheev (1997). For example, adding [-t] after *ip* (perfect for 16 examples) has an adjusted reliability of .94, while a context that worked for 475/500 examples would have an adjusted reliability of .943. This evaluation metric involves a trade-off between generality and accuracy. The above examples show that when exceptions are present, the evaluation metric does not necessarily pick out a single, obviously best context. Rather, the result is a large, detailed grammar containing hypotheses at all levels of generality, annoted for their reliability.

The predictions of these large grammars can be tested by comparing them with human judgments. In fact, people often feel comfortable with more than one form, and can rate the relative goodness of competing forms. Grammars obtained by minimal generalization can model these intuitions, since they also make multiple guesses, providing a well-formedness score for each. We define this score as the adjusted reliability for the best context that derives the output in question.

2.3. Using phonology to improve confidence

The minimal generalization algorithm uses positive evidence, relying on the contexts in which an allomorph *does* occur. However, this ignores potentially useful information: is there a pattern for the cases where the allomorph does *not* occur?

For English past tense [-d], it is helpful to recognize not only that [-d] is added to voiced segments, but that in fact it could never occur after a voiceless segment: sequences like final [pd] are unpronounceable in English. We would therefore like to learn that [-t] and [-d] are in some sense "the same" suffix — that is, there is a phonological rule turning [-d] to [-t] after voiceless sounds. We will see later that this is not just helpful but necessary; here we simply sketch a method for discovering phonological rules.

We saw above that minimal generalization compares roots that take allomorph X, and asks what structural property conditions X. In order to discover phonological processes like the devoicing of [-d] to [-t] after voiceless sounds, we must extend this question: what structural property conditions X and not some other allomorph Y? This can be answered by exploring scenarios in which Y is attached to stems that actually take X.

We propose the following: given structural changes {A→B, D→E, ...} and a context for one of them (A→B / P__Q), we must also consider / P__Q for competing changes: D→E / P__Q, etc. For example, when we compare *ripped* and *lacked* to hypothesize that ∅ → [-t] after (noncoronal) voiceless stops, we should consider the hypothesis that ∅ → [-d], another known change, can also apply after voiceless stops.

Taken alone, this modification would not help. The reliability of a rule adding [-d] after voiceless sounds would be zero, since it would generate phonologically illegal sequences like *[sɪpd]. However, if the learner also knows that *[pd] is unpronounceable in English, it can use this information to posit a phonological rule that fixes *[pd] by changing it into [pt]: [3,4]

(7) $d \rightarrow t \, / \, p\_\_ \,]_{\text{word}}$

With this rule in place, adding [-d] to verbs ending in voiceless segments is successful, with derivations like /sɪp+d/ → [sɪpt]. Verbs like *sip* thus enter the set of forms that [-d] covers. Minimal generalization can then go on to compare verbs like *sip* and *sag*,

[3] The phonological process is stated here as a rule, but it would be straightforward to translate this into a constraint-based framework.

[4] Notice that in order to posit a phonological rule, we had to make use of the information that [pd] does not occur in English, which seems suspiciously like a type of negative evidence! However, recent research has shown that infants as young as 10 months old seem to know what sequences are illegal in their language (see Hayes 1999 for a review), so it is plausible that human acquirers would already have a list of illegal sequences before embarking on morphological learning.

ultimately yielding the hypothesis that [-d] can be suffixed after any consonant regardless of voicing.

2.4. Example: English past tense allomorphy

Let us examine how minimal generalization works for English past tenses. As noted above, most verbs form the past tense by adding the regular suffix [-t]/[-d]/[-əd], with the allomorphs distributed as follows:

(8)
$$\varnothing \rightarrow \left\{ \begin{array}{ll} [\text{-əd}] & \text{after -t,-d} \\ [\text{-t}] & \text{after voiceless consonants other than } t \\ [\text{-d}] & \text{elsewhere} \end{array} \right\}$$

The 180-200 irregular verbs form their past tenses by changing vowels (*ring ~ rang*, *fling ~ flung*), by a combination of suffixation and vowel changes (*keep ~ kept*), or by more radical changes (*think ~ thought*, *is ~ was*).

We implemented the minimal generalization algorithm and tested it with a set of 2181 English verbs in phonetic transcription. This list was taken from Brian MacWhinney's web site[5] and augmented to include all irregular verbs of English. It contained 184 irregular and 1997 regular verbs. In this section, we show how minimal generalization learns the distribution of the regular allomorphs when the learning data contain no encroaching forms like *burnt* or *learnt*.

Consider first [-əd]: since this suffix occurs only after *t* and *d*, minimal generalization generalizes only enough to encompass these two segments, as in (9). Notice that the specifications that minimal generalization yields are not maximally concise, because the only features eliminated from the specification are those for which the segments differ in their values. The procedure does not yield the most elegant linguistic analysis of the relevant context, but one which is formally equivalent:

[5] http://psyling.psy.cmu.edu/Brian/papers.html

$$\text{(9)}\quad \varnothing \rightarrow [\text{-ə}d] \mathbin{/} X \left[\begin{array}{l}\text{-syllabic}\\ \text{-sonorant}\\ \text{-continuant}\\ \text{-delayed release}\\ \text{-nasal}\\ \text{-labial}\\ \text{-round}\\ \text{+coronal}\\ \text{-strident}\\ \text{-lateral}\\ \text{-dorsal}\\ \text{-high}\\ \text{-low}\\ \text{-back}\\ \text{-tense}\end{array}\right] \_\_]_{\text{word}}$$

This context covers 661 verbs, of which 590 are regular, yielding an adjusted reliability of .88.

The suffix [-d] occurs after voiced segments, including vowels (*tried*), sonorant consonants (*harmed*), voiced fricatives and affricates (*pleased*, *judged*) and voiced stops except for *d* (*rubbed*). For these classes, minimal generalization yields the hypotheses in (10)-(13):

$$\text{(10) Vowels:}\quad \varnothing \rightarrow [\text{-d}] \mathbin{/} X \left[\begin{array}{l}\text{+syllabic}\\ \text{+sonorant}\\ \text{+continuant}\\ \text{-nasal}\\ \text{+voice}\\ \text{-s.g.}\\ \text{-strident}\end{array}\right] \_\_]_{\text{word}}$$

$$\text{(11) Sonorant consonants:}\quad \varnothing \rightarrow [\text{-d}] \mathbin{/} X \left[\begin{array}{l}\text{-syllabic}\\ \text{+sonorant}\\ \text{+voice}\\ \text{-spread glottis}\\ \text{-strident}\end{array}\right] \_\_]_{\text{word}}$$

(12) Voiced fricatives and affricates:

$$\varnothing \rightarrow [\text{-d}] \mathbin{/} X \left[\begin{array}{l}\text{-syllabic}\\ \text{-sonorant}\\ \text{+del rel}\\ \text{-nasal}\\ \text{+voice}\\ \text{-s.g.}\end{array}\right] \_\_]_{\text{word}}$$

$$\text{(13) Voiced stops except } d\text{:}\quad \varnothing \rightarrow [\text{-d}] \mathbin{/} X \left[\begin{array}{l}\text{-syllabic}\\ \text{-sonorant}\\ \text{-continuant}\\ \text{-del. rel.}\\ \text{-nasal}\\ \text{+voice}\\ \text{-s.g.}\\ \text{-strident}\\ \text{-coronal}\end{array}\right] \_\_]_{\text{word}}$$

When the contexts in (10)-(13) are compared to find the most general context for [-d], we see that only two feature values are shared by all:

(14) Most general context for [-d]:

$$\varnothing \rightarrow \text{[-d]} \;/\; X \begin{bmatrix} \text{+voice} \\ \text{-spread glottis} \end{bmatrix} \underline{\quad}]_{\text{word}}$$

The result is a generalization which explains 1156/1270 verbs (adjusted reliability = .90). Interestingly, there is no way to express the notion "all voiced segments except *d*," so the algorithm is forced to include *d*. This creates a potential problem, since for regular verbs like *heed* it derives *[hiːdd] rather than [hiːdəd].[6]

The problem is solved by phonological rules, as described above. When the learner hears [-əd] after *t* and *d*, it also contemplates what would happen if [-d] were attached instead. Since the result would be unpronounceable (*[hiːdd]), it posits a phonological rule inserting schwa between two *d*'s; hence *heeded* [hiːdəd]. This rule brings the success of [-d] suffixation up to 1997/2181 forms (adjusted reliability =.91). In sum, the system expresses the generalization "all voiced sounds except *d*" through a combination of morphological and phonological rules.

How does [-t] suffixation compare to [-d] suffixation? Since the learner does not know a priori that phonological rules will allow [-d] to work in all environments, the contexts for [-t] must also be explored. Comparing verbs like *ripped* and *passed*, minimal generalization yields the hypothesis that [-t] can attach after voiceless segments. In addition, since sequences of voiced obstruents (*b, d, g, z,* etc.) plus *t* are not pronounceable in English, it posits phonological rules changing *t* to *d* in these cases. This will not work for all cases, however, since [t] is perfectly pronounceable after sonorants (*ant*, *pelt*, *part*) Therefore, the [-t] environment is never generalized beyond obstruents, accounting for 1229/1347 cases (adjusted reliability =.91).

3. Distributional encroachment

We have shown how minimal generalization, with some rudimentary phonology, can learn the distribution of [-t], [-d], and [-əd] as long as [-t] never occurs in voiced contexts. In this section, we consider how distributional encroachment can lead minimal generalization astray.

[6] In fact, the confidence score for *[hiːdd] comes out slightly higher (.90) than that for the correct form [hiːdəd] (.88).

For many English speakers, the past tense of *burn* is *burnt*.[7] *Burn* ends in a voiced (sonorant) consonant, but takes [-t]. Thus, it is a case of distributional encroachment, as defined above. We previously saw that [-t] can be added to any obstruent, given the proper phonology. Now we consider what minimal generalization does when confronted with *burnt*; generalizing, it spawns a new context:

(15) $\varnothing \rightarrow$ [-t] / X [-syllabic] ___ $]_{\text{word}}$

= “Attach [-t] to any final consonant”

How does this hypothesis perform? Adding [-t] after obstruents worked in 1229/1347 cases. Expanding this to include sonorant consonants explains one more form and adds 519 new exceptions (including all of the other verbs, such as *planned*, which end in sonorants but take [-d]). Nevertheless, the fact that [-t] worked so well for obstruents means that this new hypothesis works in 1230/1867 cases (adjusted reliability =.65). Although this hypothesis does not perform as well as the others, it does predict that a novel verb *flan* should be at least moderately acceptable with past tense *flant*! We believe that this prediction is wrong; *flant* is absurd.

The problem at hand, then, is to provide a way to detect cases of distributional encroachment as such, without letting them lead to overgeneralization.

4. Detecting distributional encroachment

4.1. Diagnosis

It is intuitively clear where minimal generalization goes wrong with *burnt*: the overambitious generalization “Attach [-t] after consonants” is internally heterogeneous. The vast majority of stems that it covers end specifically in obstruents, not just any consonant.

We can characterize heterogeneity more precisely. Assume a dialect in which the only verb of the *burnt* class is *burnt* itself. We compare the scope and hits of the two generalizations, “Attach [-t] after consonants” and “Attach [-t] after obstruents”:

[7] A few similar forms occur variably according to dialect: *learnt*, *spelt*, *smelt*, *dwelt*, *spilt*.

(16)	Scope	Hits
“Attach [-t] after consonants”	1867	1230
“Attach [-t] after obstruents”	1347	1229

In comparison to “Attach [-t] after obstruents,” “Attach [-t] after consonants” adds a large number to the scope—and just one case to the hits. Thus, although “Attach [-t] after consonants” looks good, this is because a major subset does almost all the work.

This comparison suggests a way to locate internally heterogeneous contexts: if we are to take a context seriously, it must offer a substantial improvement over the performance of its best subset.

4.2. “Impugnment”

Pursuing this idea, we propose to revise the evaluation metric for contexts. Suppose we have some context C, affiliated with some structural change A→B. We must consider every other context C′ affiliated with A→B, asking: Does C′ cover a subset of the cases of C? If so, it is a candidate for the role of the context that is “doing most of the work.”

To find out if this is so, we calculate how well C performs in cases not also covered by C′. The raw reliability of this residue set (C – C′) is:

$$\text{(17) Raw reliability}(C - C') = \frac{\text{hits}(C) - \text{hits}(C')}{\text{scope}(C) - \text{scope}(C')}$$

From Raw reliability(C – C′), we can calculate Adjusted reliability(C – C′) by taking the 75% confidence limit. However, in this case, since we are trying to estimate the *sparseness* of cases in the residue rather than the *denseness* of cases within the generalization, we must use the upper confidence limit of Raw reliability(C – C′), rather than the lower confidence limit.

If Adjusted reliability$_{\text{upper}}$(C – C′) is lower than Adjusted reliability$_{\text{lower}}$(C), then we know C was taking credit for work really done by C′. Therefore, we **impugn** C: the evaluation metric rates C at Adjusted reliability$_{\text{upper}}$(C – C′), rather than Adjusted reliability$_{\text{lower}}$(C). Thus, C receives credit only for work that it does on its own, outside the domain of C′. Impugnment is carried out for all contexts of all rules.

This algorithm is similar to the “pruning” algorithm proposed by Anthony and Frisch (1997). However, their algorithm requires the subset generalization to cover at least as many positive forms as the superset. For *burnt*, the superset generalization covers one more case than the subset, and would not be eligible for pruning.

4.3. Results for English

We tested this modification by rerunning our algorithm on the English database, this time including the exceptional form *burnt*. We used the resulting grammar to project possible past tenses for the made-up word *flan*. The results in (18) show that impugnment essentially eliminates the hypothesis that [-t] can attach to *flan*.

(18) Guess	Rule/Context	Adjusted Reliability
flanned	$\varnothing \rightarrow d\ /\ X\begin{bmatrix}+\text{syl}\\-\text{hi}\end{bmatrix} n \_\_\_]_{\text{word}}$	.953**not impugned**
flant	$\varnothing \rightarrow t\ /\ X\begin{bmatrix}+\text{cor}\\-\text{cont}\\+\text{ant}\end{bmatrix} n \_\_\_]_{\text{word}}$ = *t,d,n*	.383 **impugned** **to .006**

5. Conclusion

This paper describes our effort to develop an algorithm that can learn phonological and morphological patterns in data that include exceptions and conflicting generalizations. Unsurprisingly, the task is not as straightforward as was anticipated in Pinker and Prince (1988). In particular, distributional encroachments like *burnt* require a mechanism specifically designed to detect them.

Although we encountered the problem of distributional encroachment in the context of a bottom-up generalization algorithm, we believe that the phenomenon is a problem for top-down approaches as well. Furthermore, distributional encroachment is common in the world's languages,[8] and so must be addressed by any morphological learner.

[8] Some other cases: French "*h*-aspiré" nouns, *-a*/*-o* mismatched to gender in Romance languages, and languages in which some roots take exceptional vowel harmony, such as Hungarian (Vago 1976), Turkish (Clements and Sezer 1983), and Chi-Mwi:ni (Kenstowicz and Kisseberth 1977).

REFERENCES

ALBRIGHT, ADAM. 1998. Phonological subregularities in productive and nonproductive inflectional classes: evidence from Italian. M.A. thesis, Dept. of Linguistics, UCLA.

ANTHONY, SIMON and ALAN FRISCH. 1997. Cautious induction in inductive logic programming. In *ILP97, The Seventh International Workshop on Inductive Logic Programming*, 45-60.

CLEMENTS, GEORGE N. and ENGIN SEZER. 1982. Vowel and consonant disharmony in Turkish. In Harry van der Hulst and Norval Smith, eds., *The Structure of Phonological Representations, Part II*, Foris, Dordrecht.

DZEROSKI, SASO and TOMAZ ERJAVEC. 1997. Learning Slovene declensions with FOIDL.
http://alibaba.ijs.si/tomaz/Bib/ECML97/

KENSTOWICZ and KISSEBERTH. 1977. *Topics in Phonological Theory*, Academic Press, New York.

MIKHEEV, ANDREI. 1995. Automatic rule induction for unknown-word guessing. *Computational Linguistics* 23, 405-423.

MOONEY, RAYMOND and MARY ELAINE CALIFF. 1996. Learning the past tense of English verbs using Inductive Logic Programming. In *Symbolic, Connectionist, and Statistical Approaches to Learning for Natural Language Processing*, Springer Verlag.

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.

PRASADA, SANDEEP and STEVEN PINKER. 1993. Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* 8, 1-56.

VAGO, ROBERT. 1976. Theoretical implications of Hungarian vowel harmony. *Linguistic Inquiry* 7, 243-263.