When Nobody Wins

Kyle Gorman and Charles Yang

In memoriam Morris Halle.

Abstract We argue that lexical gaps arise when language learners fail to find productive rules in a morphological domain. Using the Tolerance Principle as a formal model of productivity, we show that lexical gaps can be predicted on purely numerical grounds using lexical statistics, with case studies on Spanish, Polish, and Russian. The learnability approach taken here leads to simpler theories of morphology.

Keywords Productivity · Lexical gap · Tolerance principle · Language acquisition · Corpus statistics · Computational linguistics

1 Introduction

As young children vividly illustrate with their performance on the wug-test (Berko 1958), the ability to generalize linguistic patterns to novel items is a core property of language. It thus comes as a surprise when we stumble into a dark and dusty corner...
of language where our boundless linguistic productivity has unexpectedly failed. For example, the English verb *forgo ‘to go without’ is unexceptional in all ways but one; unlike nearly every other verb in the language, it lacks a well-formed simple past. There’s nothing pragmatically odd about having gone without in the past—the past participle *forgone is unobjectionable—but speakers recoil at the sound of *forwent, and *forgoed is no better (and perhaps a bit worse). It is possible to find occasional examples of the hated *forwent, given a large enough corpus, but very many of these tokens are “mentions” of the gap rather than earnest “uses”. This is not the only English verb exhibiting such an idiosyncrasy; for instance, it has long been noted (e.g., Hill 1976, 668; Pinker 1999, 136f.; Pullum and Wilson 1977, 770, among others) that stride, which has an unobjectionable simple past, strode, lacks a standard past participle; both *stridden and *strided are ill-formed, and few speakers accept *strode in a past participle context.1 We refer to these unexpectedly ineffable forms as lexical gaps.2

In one sense, there is nothing particularly novel about lexical gaps. They are found both in the languages of large and powerful nation-states—those with armies and navies, those with national language academies—and in endangered languages spoken by a few toothless elders; for cross-linguistic samples, see papers in Baerman et al. 2010, Fanselow and Féry 2002, and Rice and Blaho 2009. But while lexical gaps caught the attention of both ancient grammarians (see, e.g., Neue 1866) and the founders of modern linguistic theory (see below), they have played remarkably little role in theories of productivity; they have been “ill incorporated into our theories of language” (Baerman and Corbett 2010, 2). For instance, Optimality Theory (Prince and Smolensky 1993), Distributed Morphology (Halle and Marantz 1993), and dual-route models of word formation (e.g., Pinker 1999) all appear to assume the existence of a default form and must employ additional mechanisms to predict the illformedness of *forwent and *stridden.

In his Prolegomena to a theory of word formation, Halle (1973) draws attention to the existence of lexical gaps, illustrating this with an example from Russian (see Sect. 5 below). Halle proposes two accounts of these gaps, which he regards as equivalent. Under the first—and better-known—proposal, words like *forwent are generated by the word formation component of the grammar, but are marked with a feature [−Lexical Insertion] preventing their actual use. This idea is implicit in accounts of lexical gaps (e.g., Pertsova 2005, 2016) depending upon the notion of lexical conservativism (Steriade 1997, 2008), which holds that learners require explicit positive evidence before they will attempt to use a particular form. This is easily encoded in classical Optimality Theory: assuming the “initial state” is one

1For a quantitative analysis of the *stridden gap and the failings of frequency and other measures of indirect negative evidence as a solution to the gap problem, see Yang (2016), section 5.1.1.

2We avoid the competing term paradigm (or paradigmatic) gap, as we do not wish to make a commitment to the theoretical notion of the inflectional paradigm.

When Nobody Wins

where all markedness constraints dominate all faithfulness constraints (Smolensky 1996), then for any novel input, a phonologically null—yet ineffable—candidate, the Null Parse, is optimal (e.g., Raffelsiefen 1996, 1999; Rebrus and Törkenczy 2009; Rice 2005; Wolf and McCarthy 2009).3 However, this account is difficult to reconcile with the unbounded creativity of word formation: children do pass the wag-test, and the instant a novel verb like google entered the English lexicon, its past tense googled became available.4

As an alternative, Halle (1973, fn 1) suggests that only those items undergoing “nonproductive” word formation rules are eligible for the feature [−Lexical Insertion], all others being marked [+Lexical Insertion] by default. Hetzron (1975), arguing against Halle’s first proposal, comes to the same conclusion: gaps arise in unproductive corners of the grammar. Many later authors have adopted some form of Halle’s second proposal (e.g., Albright 2003, 2009; Anderson 2010; Baronian 2005; Hudson 2000; Maiden and O’Neil 2010; Pullum and Wilson 1977), linking lexical gaps to unproductivity (though putting forth divergent accounts of the factors contributing to productivity). We too endorse Halle’s second proposal. But since words and rules do not wear their (un)productivity on their sleeves, we employ a model of linguistic processing, and more specifically, one of the consequences of this model, the Tolerance Principle (Yang 2005, 2016). This principle predicts, on strictly numerical grounds, whether or not a linguistic generalization will be encoded as productive by the child learner. Under our account, lexical gaps arise in precisely those linguistic contexts for which the Tolerance Principle identifies no productive generalizations.

In the following section, we summarize the key features of the Tolerance Principle and how it is applied to the problem of productivity in language acquisition. Lexical gaps are a logical consequence of the Tolerance Principle in language acquisition, and as we will see, are closely related to the notion of competition in morphology, the central theme of the present volume. We then provide case studies of lexical gaps in three languages. Note that our goal, in each case study, is to predict the linguistic contexts in which gaps arise, defining those contexts using distributional, and where available, behavioral, evidence. However, unlike some prior work (e.g., Albright 2003; Hudson 2000; Pertsova 2016), we do not attempt to predict exactly which complex words will be judged ineffable. We adopt this more modest goal due to our suspicion that ineffability of individual forms is synchronically arbitrary (see, e.g., Baerman 2008 on defectivity in Russian verbs), and in Sect. 3, we observe the failure of an existing, explicit model of gaps making

3Orgun and Sprouse (1999) propose a different, albeit roughly equivalent, mechanism for encoding defectivity in Optimality Theory; see Fanselow and Féry (2002) and Rice and Blaho (2009) for discussion of these mechanisms.

4See also chapter 5 of Yang (2016) for criticisms of a more subtle, frequency-based, form of lexical conservatism that uses indirect negative evidence.
predictions at the word level. Furthermore, we do not model actuation or change of lexical gaps, or the strategies speakers use to compensate for gaps (e.g., periphrasis), though these represent fruitful topics for future work.

2 Productivity and Gaps

How would one represent a rule \( R \) and the exceptions of that rule (e.g., a set of words \( w_1, w_2, \ldots, w_n \))? If one is committed to a mechanistic account—like a computer programmer, for instance—perhaps the only way to encode rules and exceptions is through a set of conditional statements:

1. If \( w = w_1 \), then …
   If \( w = w_2 \), then …
   …
   If \( w = w_e \), then …
   Apply \( R \)

(1) entails that in order for \( R \)—the productive rule—to apply, one must first scan through a list of competing generalizations to ensure that it is not one of the exceptions \((w_1, w_2, \ldots, w_n)\). This immediately recalls the Elsewhere Condition, an idea dating back to Pāṇini and later adopted by generative grammarians (e.g., Anderson 1969; Kiparsky 1973). But there is something perverse about (1); to produce walked, for example, does one really need to scan a list of irregular verbs to confirm that walk is not among them? A moment of reflection, however, suggests that the only other option—listing thousands of regulars instead of a hundred odd irregulars—is even worse. One could imagine assigning to each regular verb a special feature triggering the application of the “add -ed” rule, but that would imply that the morphological status of every verb would need to be committed to lexical memory: both the regulars, and the irregulars too, since they are by definition unpredictable.

In fact, there is wide-ranging behavioral evidence for the “irregulars first, regulars last” representation of morphology; see Yang 2016, chap. 3, for a review. The key evidence comes from online processing of words in lexical decision and word production studies. When irregulars and regulars are suitably matched for various factors (e.g., stem and surface frequency) known to affect the speed of processing, irregulars are recognized and produced significantly faster than

2.1 The Tolerance Principle

From this we can provide a cost-benefits analysis for any set of competing generalizations. Specifically, an application of a productive rule must “wait” until all the exceptional conditions have been checked, and the more exceptions there are, the longer one will have to wait. Under very general assumptions about word frequencies (Yang 2016, 61f.), we can prove:

2. Tolerance Principle

Suppose a rule \( R \) is applicable to \( N \) items in a learner’s vocabulary, of which \( e \) are exceptions that do not follow \( R \). The necessary and sufficient condition for the productivity of \( R \) is:

\[ e \leq \theta_N \text{ where } \theta_N := \frac{N}{\ln N} \]

The Tolerance Principle requires two input values, \( N \) and \( e \), and returns the productivity status of a rule. Its application requires a well-defined rule such that \( N \) and \( e \) can be measured, whether by the child learner or the researcher. To learn the structural description of a rule, typically in the form of \( "X \rightarrow Y" \), one will need to invoke an inductive learning procedure for extracting generalizations from specific learning instances. For example, suppose two strong baseball batters are described with feature bundles \([+\text{red cap}, +\text{black shirt}, +\text{long socks}]\) and \([+\text{red cap}, +\text{black shirt}, +\text{short socks}]\). Simply by intersecting the two feature bundles we can obtain a minimal, exceptionless generalization \( "[+\text{red cap}, +\text{black shirt}] \rightarrow \text{good hitter}" \). Obviously, an appropriate inductive learner must encode the structural constraints on the human language faculty and other cognitive systems implicated in language acquisition (Chomsky 1965).

Table 1 provides some sample values of \( N \) and the associate threshold value \( \theta_N \).

<table>
<thead>
<tr>
<th>( N )</th>
<th>( \theta_N )</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>40.0</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>35.0</td>
</tr>
<tr>
<td>50</td>
<td>13</td>
<td>26.0</td>
</tr>
<tr>
<td>100</td>
<td>22</td>
<td>22.0</td>
</tr>
<tr>
<td>200</td>
<td>38</td>
<td>19.0</td>
</tr>
<tr>
<td>500</td>
<td>80</td>
<td>16.0</td>
</tr>
<tr>
<td>1,000</td>
<td>145</td>
<td>14.5</td>
</tr>
<tr>
<td>5,000</td>
<td>587</td>
<td>11.7</td>
</tr>
</tbody>
</table>

5We do not claim that the language user will activate the full lexical entries for the irregulars before accessing the regulars. It is conceivable that the computational search process deals with a hashed list, as described in (1): the entries are “keys” indexed by the semantic and phonological properties of words, which point to full lexical entries in long-term memory, rather than lexical entries themselves.
These thresholds are significantly lower than a naïve “majority rules” principle, traditionally viewed as a hallmark of productivity (e.g., “statistical predominance”, Nida 1949, 14), and this has some interesting implications for language acquisition. For instance, the principle predicts that a smaller vocabulary (i.e., smaller values of N) can tolerate a higher percentage of exceptions, all else held equal, so productive rules are more detectable in a learner who has encountered less primary linguistic data. This may explain children’s remarkably early command of the main ingredients of language, and explain why maturational constraints seem to aid, rather than hinder, child language acquisition (Newport 1990; see Yang 2016, chap. 7, for extensive discussion).

Despite its simplicity, the Tolerance Principle has proved surprisingly effective. In Yang (2016), it is applied nearly 100 times, making accurate productivity predictions across numerous languages and domains from corpus statistics. Furthermore, artificial language studies conducted with young children have found near-categorical confirmation for the Tolerance Principle (Schuler et al. 2016). These robust results are somewhat unexpected; the derivation in (2) makes use of numerical approximations that only hold when N is large, but in the artificial language studies, the vocabulary, and thus N, is often quite modest (e.g., 8 or 9).

Finally, we note that the Tolerance Principle provides a discrete threshold for productivity, which conforms with the traditional conception discussed by Aronoff (1976) and reviewed by Bauer (2001) but are at odds with the more recent, gradient, perspective on productivity (e.g., Baayen 2009; see also Versloot and Hoekstra 2019). The child language acquisition evidence, however, unambiguously favors the categorical view; see Yang (2016, chap. 2), for a cross-linguistic review. In addition, it is interesting to observe that a threshold such as that provided by the Tolerance Principle, can only be analytically derived and is in principle empirically undiscoverable. Productive processes will lie above the threshold and unproductive processes will lie below, but with arbitrary “distance” from it in both cases: the threshold cannot be determined from data.

2.2 Acquisition of Productive Rules

Before returning to the question of lexical gaps, we first demonstrate the application of the Tolerance Principle to a well-known case, the acquisition of the English past tense. We then briefly discuss how the Tolerance Principle can provide a learning-theoretic account of the German noun plural system, which has several competing but productive rules.

Any reasonable inductive procedure will identify the regular -d suffix (realized as [-d, -t, -ed]) as applying to stems of all phonological shape; thus, its productivity will be determined by the total number of verbs and the number of irregular verbs—N and e, respectively—in the learner’s vocabulary. The same considerations apply to the various irregular rules. For instance, the irregular verbs bring, buy, catch, fight, seek, teach, and think all undergo a stem change replacing the rime with [i]. But this “rime = o” rule will fare terribly when applied to the remainder of the lexicon; its seven undergoes are easily swamped by thousands of exceptions, far exceeding the tolerance threshold. As a result, the rule must be lexicalized. Other irregular patterns can be analyzed similarly: as shown elsewhere (Yang 2016, chap. 4), all rules except the regular “add -d” are unproductive by this principle, accounting for the near-total absence of over-regularization errors in the speech of children acquiring English (Xu and Pinker 1995; see Yang 2016, 34 for discussion).

Following the same logic, we can see that the emergence of the “add -d” rule will require a long gestation period. Although children can quickly induce its structural description—perhaps using no more than a few dozen verbs (e.g., Yip and Sussman 1997)—irregulars are overrepresented in baby’s first verbs. For instance, in a corpus of five million words of child-directed English drawn from the CHILDES (MacWhinney 2000) database, 76 of the 200 most common past tense verbs are irregular. As θ_{200} is only 37, we predict that children who only know the top 200 verbs cannot recognize the productivity of “add -d” despite the fact that it is the majority rule. For a learner with such a vocabulary, we posit that verbs produced with the -d suffix are in effect encoded as similarly to (“true”) irregulars, so that the generalization does not extend beyond regular verbs encountered thus far; see Marcus et al. (1992) for review and discussion.

Telltale evidence for productivity comes from the first attested overregularization errors, and when the developmental record is sufficiently rich, the Tolerance Principle helps us predict when this will occur with some precision. It certainly works for “Adam” (Brown 1973), one of the most-studied language learners in psycholinguistics, who produced his first recorded overregularization error at 2:11: “What dat feel like?” Adam’s earlier transcripts, going back nearly a year, do not contain a single instance of an incorrect irregular verb past tense. Tolerance predicts that by 2:11, Adam had acquired a sufficiently large number of regular verbs to overwhelm the irregulars. In Adam’s transcripts up to age 2:11, there are N = 300 unique verbs in all, of which e = 57 are irregular. This is quite close to the predicted θ_{300} = 53, and the discrepancy may be due to undersampling of the regulars, which are less frequent on average and thus more likely to be left out of a sample. In summary, Adam appears to have acquired a productive “add -d” only once he acquired a filibuster-proof majority of regular verbs, consistent with the predictions of the Tolerance Principle.

The formulation of the Tolerance Principle suggests that a rule cannot be productive unless it has relatively few exceptions. On the face of it, this is inconsistent with the fact that sometimes minority rules can still be productive. For instance, the much-studied problem of German noun plural formation concerns the status of five suffixes (-s, -(e)n, -e, -er, and -ö). Virtually all scholars hold that -s as in Auto-Autos ‘car(s)’ is productive. The other four suffixes are controversial: the dual-route model holds them to be unproductive and lexical (Marcus et al. 1995; Clahsen

---

\(^6\)Baayen’s formulation unfortunately, and tautologically, quantifies productivity using an inherently continuous measure, a ratio.
Table 2 Distribution of noun plural suffixes for frequent nouns in child-directed German

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Types</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
<td>87</td>
<td>18.9</td>
</tr>
<tr>
<td>-e</td>
<td>156</td>
<td>34.1</td>
</tr>
<tr>
<td>-er</td>
<td>30</td>
<td>6.5</td>
</tr>
<tr>
<td>-(e)n</td>
<td>172</td>
<td>37.5</td>
</tr>
<tr>
<td>-s</td>
<td>13</td>
<td>2.8</td>
</tr>
</tbody>
</table>

1999) but many other scholars consider them to be largely productive (in terms of predictability) albeit with exceptions (e.g., Wunderlich 1999; Dressler 1999; Hahn and Nakisa 2000; Wiese 2000). Notably, the -s suffix covers only a small minority of nouns. Table 2 above provides the statistics based on some 450 highly frequent noun plurals in child-directed German in a recent study (Yang 2016, Sect. 4.4): the distribution is very similar to those based on larger corpora (Clahsen et al. 1992; Szagun 2001; Elsen 2002; Sonnenstuhl and Huth 2002).

An application of the Tolerance Principle clearly will not produce a productive suffix over all the nouns, since no suffix is anywhere near the requisite tolerance threshold. Yang (2016) reports several detailed case studies of this type: when children fail to discover a productive rule over a set of lexical items, they will subdivide the set along some suitable dimension and apply the Tolerance Principle recursively. For the German plural system, the relevant dimensions are gender and the phonological properties of the final syllable, which have been discussed in the previous literature on German morphology and phonology (e.g., Wiese 2000) and supported by the acquisition finding that young German-learning children have good command of noun gender and phonotactics (e.g., Mills 1986; Szagun 2004). The analysis is too complex to be summarized here but a recursive application of the Tolerance Principle readily establishes -(e)n as the productive suffix for feminine nouns, -Ø for (non-feminine) nouns ending in a schwa followed by l, r, or n, -n for (non-feminine) nouns with a final schwa, and -er for monosyllabic neuter nouns with a back vowel. The net effect is that almost all nouns are predictably accounted for by the four suffixes: each suffix still has exceptions but the number of them fall under their respective tolerance thresholds. This removes almost all nouns from consideration when it comes to the -s suffix, which imposes no structural restrictions on the noun and thus becomes the default.

At this point it is important to stress that the formulation of rules, and their subsequent evaluation under the Tolerance Principle, are statements about the psychological state of language learners—the counts of N and e, specifically—concern the vocabulary composition of young children. We must also keep in mind that children learn languages very early and accurately, and the terminal state of language acquisition across individuals in a speech community is remarkably uniform, as shown in detailed studies of language variation (e.g., Labov 1972; Labov et al. 2006). As a logical consequence, the core morphological rules of language must be learnable with a very small vocabulary of fairly common words. For instance, a three-year-old’s knowledge about the word order and inflectional morphology of their native language is generally perfect: this is age by which even the most fortunate have only just over 1,000 words in their lexicon (Hart and Risley 1995). There is converging evidence that lexical frequency can help to provide a reasonable approximation of vocabulary. Nagy and Anderson (1984), for instance, estimate that most English speakers know words above a certain frequency threshold (about once per million). Developmentally, it is also known that children’s vocabulary acquisition correlates with word frequencies in child-directed speech (e.g., Goodman et al. 2008), especially for open-class words, the primary arena of morphological productivity; see Merkuur et al. (2019) for an application of the Tolerance Principle to words from various frequency ranges to approximate the vocabulary growth of language learners. Thus the calibration of productivity in a psychological setting, one which aims to characterize the internal grammar of human speakers, should only use vocabulary items that can plausibly factor into language acquisition. As stressed by Dressler et al. (2019), one needs to distinguish true productivity from occasionalism and other highly context-dependent neologisms that often involve a high degree of language awareness.

This means that mere attestation in very large corpora, which are now increasingly available for linguistic research, may have no bearing at all to the acquisition of that form: no one learns morphology from the OED, probably not even linguists.

### 2.3 A Condition on Lexical Gaps

The recursive application of the Tolerance Principle, then, goes hand in hand with its smaller—is-better property (Table 1). The failure to establish a productive rule for a set of lexical items triggers the subdivision of the set, and the learner will seek to discover productive rules within: a smaller N can tolerate a proportionally higher number of exceptions. In other words, productive rules can be established on non-overlapping domains. But the subdivision of the lexical items must be empirically motivated, and can only take place along the dimensions that child learner can attend to: for instance, it’s easy to imagine morphological marking being sensitive to the animacy of nouns (e.g., cat vs. chair) but probably not the furriness of nouns (e.g., a Siamese vs. a Sphinx cat). Thus, there remains a possibility that the learner, or the linguist, will fail to identify a productive rule under any (humanly) possible scrutiny of the data. That, we argue, is where gaps arise.

(3) Condition on lexical gaps (Yang 2016, 142):

Let there be S rules, each affecting N_i lexical items (1 ≤ i ≤ S), and

\[ N = \sum_{i} N_i \]

Gaps arise if and only:

\[ \forall i : \sum_{j \neq i} N_j > \theta_N \]
That is, gaps arise when no rule \( (N_i) \) in \( N \) applies to sufficiently many items to tolerate the remaining \( (\sum_{j \neq i} N_j) \) items as exceptions. The following three case studies, focusing on previously-documented lexical gaps in Spanish, Polish, and Russian, illustrate this prediction in action. To reiterate: in each case, our goal is simply to predict the contexts in which lexical gaps arise, defining those contexts via independently-motivated linguistic analyses.

This is a good place to review the Tolerance Principle and its application to gaps in the context of competition, the main theme of the present volume. The centrality of competition can be observed in three aspects of the present work. First, and as long recognized, the Elsewhere Condition is a direct embodiment of competition between the lexicalization of exceptions and the automatic application of rule-based computation. Second, the derivation of the Tolerance Principle is predicated on the assumption that the language learner favors efficiency: the grammar that lists the exceptions ahead of a rule and the grammar that lists all lexical items are in competition, and the faster one is chosen. Third, the condition on gaps claims that gaps arise from competition over a morphological category: when no one generalization is sufficiently dominant, nobody wins.

3 Spanish Verb Stems

Harris (1969, 114) observes that the Spanish verbs agredir ‘to attack’ and aguerrir ‘to harden’ are defective (i.e., have unpredictable lexical gaps) in certain inflectional forms. Below we review this pattern of defectivity, identify several other verbs with the same pattern of defectivity, and show that these patterns are consistent with the above condition on lexical gaps. But first, we review the competing generalizations.

3.1 Stem Changes and Defectivity

The majority of Spanish verbs conjugate without any stem change save for a shift of primary stress between (the final syllable of) the stem and the various syllables of the desinence (i.e., the inflectional suffixes). However, many verbs undergo stem changes targeting the vowel of the final syllable of the stem. All such verbs have a mid vowel—\( e \) or \( o \)—in the final syllable of their stem in the citation form (the infinitive), and henceforth we refer to these as mid vowel stems. The most frequent such change is seen in those mid vowel stems in which the stem’s final syllable vowel is \( e \) or \( o \) when primary stress falls on the desinence, as in the infinitive (e.g., contar ‘to count’), but is realized as a diphthong—\( ie \) [\( je \)] or \( ue \) [\( we \)], respectively—when stress falls on the stem, as in the 1st person singular indicative (cuento ‘I count’). Near-minimal pairs like negar-niego ‘deny’ and pegar-pegó ‘stick onto’ show that one cannot easily predict whether or not a mid vowel stem will diphthongize or not.\(^7\) Following Brane and Bordelois (1973, 132f.) and Harris (1969, 74f., 1985), we assume a lexically-specific rule which maps underlying /e/ to /\( je\)/ and /\( o\)/ to /\( we\)/ in certain stems.

4 Diphthongization:

\[
\begin{align*}
\{ e \} \\
\{ o \}
\end{align*} \rightarrow \begin{align*}
\{ \text{je} \} \\
\{ \text{we} \}
\end{align*} \quad \text{(Condition:} \sqrt{\text{CONT}}, \sqrt{\text{NEG}}, \ldots \text{)}
\]

Some verbs of the third (-i-) conjugation exhibit a superficially-similar stem change, an alternation between \( e \) and \( i \) in the stem-final syllable (e.g., pedir-pido ‘ask for’, gemir-gimo ‘groan’).\(^8\) Unlike diphthongization, however, this alternation is not directly conditioned by primary stress, and indeed, there does not seem to be any one positive condition for raising /\( e\)/ to \( i\). Rather, we follow Harris (1969, 111) and Lerner (2011) in modeling this alternation with a rule of high-vowel dissimilation lowering /\( e\)/ to \( e\) when followed by another \( i\).\(^9\) This rule must also be lexically conditioned so as to prevent overapplication to non-alternating stems such as vivir-vivo ‘live’.

5 Lowering:

\[
i \rightarrow e / C_0 i \quad \text{(Condition:} \sqrt{\text{FID}}, \sqrt{\text{GIM}}, \ldots \text{)}
\]

Albright (2003) and Maiden and O’Neill (2010) claim that, with rare exceptions, defective verbs are mid vowel stems of the third conjugation, like the examples noted by Harris (1969). Maiden and O’Neill further observe that gaps are present in exactly those inflectional forms alternating due to the diphthongization and/or lowering rules. For example, the 1st person singular present indicative of agredir could be *agredo, *agriedo, or *agrido, but speakers are reluctant to produce or to accept any of the three. Examples of the four patterns are given below.\(^{10}\)

\(^7\) Albright et al. (2001) argue that there is some weak segmental conditioning on the application of diphthongization. Even if this is correct—and the evidence is not overwhelming—it does not obviate the need for lexical diacritics.

\(^8\) These \( e-i \) alternations are also present in deverbal derivatives (Harris 1969, 110; Lerner 2011).

\(^9\) This correctly predicts that this pattern will be restricted to the third conjugation, since only this conjugation has an -i- theme vowel needed to condition the lowering rule.

\(^{10}\) Note that the \( e-i \) alternation in the conjugation of sumergir is purely orthographic; both are [\( x\)] throughout.
generalizations is productive. More concretely, we predict that the competing rules satisfy the condition on lexical gaps. We now show this prediction is supported both by numerical and behavioral evidence.

3.2 Productivity of Stem Changes

Table 3 gives the frequencies of the competing mid vowel stem generalizations in the LEXESP database (Sebastián et al. 2001) by conjugation. In both the first (–a-) and second (–e-) conjugations, the “no change” pattern emerges as productive: it has fewer exceptions (105 and 23, respectively) than the Tolerance Principle thresholds (181 and 32). In the third conjugation, lowering represents a slim majority, but there are too many exceptions to the three competing generalizations for any of the three to satisfy the Tolerance threshold (10), consistent with our predictions above.

These generalizations are supported by results from two studies of child speech errors. Calhoun et al. (2002) analyze verb inflection errors in a sample of 15 children acquiring Spanish, constituting roughly 6,000 tokens. They report that nearly all errors involving verb stems (116 out of 120) involved underapplication of diphthongization: that is, the child failed to apply the rule to a diphthongizing verb. Mayol (2007) analyzes verb stem errors in a sample of six children acquiring Spanish (constituting roughly 2,000 tokens). Mayol identifies several hundred instances where diphthongization was underapplied, but not a single case where it

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Counts of stem changes in Spanish mid vowel verb stems by conjugation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st (–a-)</td>
<td></td>
</tr>
<tr>
<td>No change</td>
<td>1,200</td>
</tr>
<tr>
<td>Diphthongizing</td>
<td>105</td>
</tr>
<tr>
<td>2nd (–e-)</td>
<td></td>
</tr>
<tr>
<td>No change</td>
<td>144</td>
</tr>
<tr>
<td>Diphthongizing</td>
<td>23</td>
</tr>
<tr>
<td>3rd (–i-)</td>
<td></td>
</tr>
<tr>
<td>No change</td>
<td>3</td>
</tr>
<tr>
<td>Diphthongizing</td>
<td>13</td>
</tr>
<tr>
<td>Lowering</td>
<td>20</td>
</tr>
</tbody>
</table>

11 Many Spanish verb stems have multiple prefixal variants; for instance, destresar ‘to undress’, investir ‘to invest’, revestir ‘to decorate’, and traer(se) ‘to cross-dress’ are all plausibly derived from vestir ‘to dress’. Without exception, all verbs derived from the same stem undergo the same set of stem changes, and therefore we collect counts over verb stems rather than verbs.

Verbs were manually grouped by stem, with an etymological dictionary (Roberts 2014) used to adjudicate unclear cases.

12 We analyze the three conjugations separately because it has long been theorized that the productivity of the various stem changes may be a function of conjugation. Note, however, that the Tolerance predictions are the same if the first two conjugations are grouped together.
was overapplied to a no-change verb. Statistical analyses find a significant effect of conjugation on correct usage of irregular verbs, with the highest error rates occurring in the first and second conjugations. Of the 345 errors identified by Mayol, the only errors in the third conjugation were two overregularizations of the diphthongizing stems dormir "to sleep" and venir "to go", respectively. This suggests not only that children acquiring Spanish are sensitive to conjugation, but it is consistent with our predictions that “no change” is the default in the first and second conjugations, and this results in underapplication of diphthongization in these conjugations. Somewhat counterintuitively—but as predicted by our model—the absence of a productive default for mid vowel stems of the third conjugation is responsible in part for the rarity of errors in that conjugation.

In summary, a straightforward application of the Tolerance Principle predicts the context in which Spanish lexical gaps are found, as well as patterns of inflection errors made by children acquiring Spanish.

3.3 Evaluation of the Minimum Generalization Model

Albright (2003) proposes a computational model intended to predict exactly which Spanish verbs will be defective. In an earlier study, Albright et al. (2001) uses a probabilistic rule induction system, the Minimal Generalization Learner (MGL; Albright et al. 2001; Albright and Hayes 2003) to model stem changes in Spanish verbs. This model takes as input pairs of strings and induces a set of rules to map from the former to the latter. Albright et al. report that this model accurately predicts human performance on a wug-test conducted with adult Spanish speakers, though they do not evaluate their model on real-world data. In the MGL model, each induced rule is associated with a confidence score, a function of the number of undergoers and exceptions to that rule. Albright (2003) extends the wug-task by asking native speakers to rate, on a seven-point Likert scale, their “uncertainty” as to the correct inflection of various Spanish verbs. This measure of uncertainty, averaged across participants, is negatively correlated with (subjective) frequency and with model confidence scores, and Albright implies these measures predict defactivity of specific verbs.

We share with Albright the intuition that the productivity of a rule is a function of the number of that rule’s undergoers and exceptions. However, our theory makes no predictions about the effect of frequency, nor does it make any predictions about the defactivity of individual verbs. Furthermore, it is not obvious to us that Albright’s dependent variable—average speaker “uncertainty”—is an appropriate operationalization of lexical gaps, particularly since Albright notes it is not strongly correlated with dictionary data on defactivity. We therefore attempt to estimate the degree to which the MGL model confidence score and frequency could act as predictors of the lexical gaps listed in (7). Following Albright (2003), we train the MGL model on pairs of candidate words—non-defective third-conjugation mid-vowel verbs—from the LEXESP database (Sebastián et al. 2001), "holding out" the defective verbs (7) for evaluation.13 Each input pair consists of the infinitive minus -ir and the 1st person singular indicative minus -o. This latter form is chosen because it exhibits a diphthong in diphthongizing stems, and the high vowel alternant in lowering stems. For example, for the lowering verb pedir-pido, the input form is [ped] and the output is [pid]. As in Albright 2003, the confidence score associated with each pair is that of the highest-scoring applicable rule.

In Fig. 1, the confidence score is plotted on the x-axis and lemma frequency on the y-axis; defective verbs are indicated with circles. While some defective verbs have a relatively low confidence score, and while they are of relatively low frequency on average, there are also many non-defective verbs which have even lower confidence scores and/or frequencies.14 In other words, neither of the features proposed by Albright appear to correctly predict gaps, at least without further information. This result confirms our suspicion that while Albright’s model may produce scores closely correlated with speakers’ intuitions about “uncertainty”, these intuitions do not adequately operationalize defactivity. We note that similar results are reached by Boyé and Cabredo Hofherr (2010) in an evaluation of Albright’s model on certain French present tense first and second person plural indicatives and imperatives (e.g., clore ‘to close’). We also note that in other

13Whereas in Table 3 we collapse prefixed variants of a stem into a single entry, Albright does not mention any such practice, so inputs to the MGL system in this experiment are verb stems plus any prefix(es), rather than stems.
14However, it may still be the case that subjective frequencies used by Albright are better predictors of defactivity than the corpus-based frequency norms used here.
languages, lexical gaps are found even in high-frequency words (e.g., Greek kópela ‘girl’; Sims 2007), further undermining the hypothesis that low frequency contributes to lexical gaps.

4 Polish Masculine Genitives

Polish declension has been discussed in the language acquisition literature as a serious challenge to the dual-route model of morphology, according to which learners must identify a default rule. Although we are not aware of this case in the literature on inflectional gaps, the empirical issue is clearly the same.

In Polish, masculine nouns in the genitive singular (gen.sg.) either take the -a or -u suffix, but Dąbrowska (2001) shows that neither is the default according to conventional criteria (Marcus 1995). For instance, unlike the English past tense -d, which is always extended to novel verbs (e.g., googled), some novel Polish masculine nouns take -a whereas others take -u. The distribution of these two suffixes is widely studied but has consistently defied philologists’ attempts to provide a systematic classification; the choice appears to be largely arbitrary (Maunssch 2003). While some masculine nouns show -a and -u in apparent free variation (e.g., deser ‘design’), for others the genitive singular is infeasible:

A Pole is often uncertain as to the correct genitive form of cities like Dublin ‘Dublin’ and Göteborg ‘Gothenburg’ [...]. My informants were uncertain as to the correct genitive form of Tarnobrzeg ‘Tarnobrzeg’ which, according to the dictionaries, takes -a. (Kotum 1981, 182f.)

We add to this list the following masculine nouns, identified as defective in the genitive singular by our Polish informants, which clearly shows that neither -a nor -u is productive.

(8) 
- drut ‘wire’
- rower ‘bike’
- balon ‘balloon’
- karabin ‘rifle’
- autobus ‘bus’
- lotos ‘lotus flower’

These facts collectively suggest the absence of a productive rule; which suffix a noun selects must be learned by rote, a process which continues well into the teenage years (Dąbrowska 2001, 2005). In contrast, the genitive plural (gen.pl.) for masculines is unproblematic: the default suffix is -ów with a small number of exceptional nouns taking -ů’s. Children acquiring Polish make very few errors in the genitive singular (gen.sg.), and frequently overextend -ów in the gen.pl., just as this description leads us to expect (Dąbrowska 2001).

Table 4: Distinctives of genitive suffixes on Polish masculine nouns, the productivity prediction of the Tolerance Principle, average frequency (mean tokens per million words), and children’s error rates

<table>
<thead>
<tr>
<th>Suffix</th>
<th>N</th>
<th>Avg. freq.</th>
<th>Child error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gen.sg.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-a</td>
<td>837</td>
<td>7.2</td>
<td>1.28</td>
</tr>
<tr>
<td>-u</td>
<td>516</td>
<td>8.8</td>
<td>0.24</td>
</tr>
<tr>
<td>gen.pl.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-ów</td>
<td>551</td>
<td>6.5</td>
<td>0.41</td>
</tr>
<tr>
<td>-ů’s</td>
<td>61</td>
<td>11.4</td>
<td>15.53</td>
</tr>
</tbody>
</table>

Adapted from Dąbrowska (2001)

Applied to noun stems found in child-directed Polish, the Tolerance Principle provides a straightforward account of both patterns. The data comes from a word frequency list compiled by Ewa Haman from CHILDES transcripts (Smoczyńska 1985; Weist et al. 1984; Weist and Witkowska-Stadniki 1986). Masculine nouns were manually classified into four classes according to their genitive case endings (e.g.: -a or -u; pl.: -ów or -ů’s). These results are summarized in Table 4.

For the singulars, there are 837 + 516 = 1353 stems. By the Tolerance Principle, a productive suffix can tolerate no more than 0.1353 ≈ 188 exceptions: -a and -u are both too numerous to meet this threshold, and the absence of a productive suffix is thus predicted. For the plurals, however, -ów attaches to 551 stems whereas -ů’s attaches only to 61: the productivity threshold is easily met (61/551 ≈ 95). These predictions accord well with errors made by children acquiring Polish. For the singulars, the lack of a productive process offers no opportunity for overregularization and children’s performance on both suffixes is consistently high. For the plural, however, the existence of the -ów productive default serves as the attractor for overregularization: to wit, -ů’s nouns have the highest error rates by far, even though they have a higher average frequency in child-directed Polish. To summarize, the presence of a productive rule for generating the gen.pl. leads to overregularization errors, whereas the lack of such a rule in the gen.sg. leads to defectivity.

5 Russian Verb Stems

We now return to lexical gaps in Russian verbs, the phenomenon which occasioned some of the earliest considerations of productivity in generative grammar. Halle (1973) estimates that there are roughly 100 defective Russian verbs, but does not
provide a list; Sims (2006) develops such a list by inspecting multiple Russian dictionaries annotating defectivity, and identifies 70 defective verbs in all; Pertsova (2016) further refines this list to 63 verbs. A few examples are given below.

- oščutit’ ‘sense’
- šerstit’ ‘search vehemently’
- pobedit’ ‘win’
- čudit’ ‘act strangely’
- forsit’ ‘swagger’
- pylesosit’ ‘vacuum’
- derzit’ ‘be rude’
- lazit’ ‘climb’

With one exception, all defective verbs are dental stems of the 2nd (-t-) conjugation, and are gapped in the first person singular (1sg.) of the so-called “non-past” tense. Yang (2016, 152f.) provides a brief Tolerance-based account of this gap, as follows. Second conjugation verb stems ending in t undergo one of two mutations in the 1sg. nonpast (in addition to the -u suffix); they may either become the affricate č [t̪] (e.g., mettis’-metču ‘mark’) or an alveopalatal fricative šč [ʃ] (sokraitis’-sokraitču ‘reduce in number’). While the former is the majority pattern, neither emerges as a productive default under the Tolerance Principle, and thus t-stems of the second conjugation appear to satisfy the condition on gaps.

Pertsova (2016) notes that this account is incomplete, for it fails to account for gaps in d, s, z-stems of the second conjugation. For these verbs, it is more difficult to construct a competition-based account simply because all three are generally thought to undergo exactly one mutation: d and z to ẓ [t̪] (e.g., sidet’-sido ‘sit’, vonzit’-vonču ‘pierce, stab’) and s to š [ʃ] (kositi’-kosč ‘scythe’). However, we will argue there is some evidence for residual competing generalizations.

Sims (2006, 197f.) and Baerman (2008) review the history of the 1sg. nonpast mutations. The modern mutations to š and ẓ represent the expected Russian reflexes of Common Slavic *št, *dšt, respectively. Christianization, beginning at the close of the first millennium, brought about a period of substantial contact with southern Slavic speakers, and their liturgical language, Old Church Slavonic (OCS), contributed novel reflexes of *št, *dšt, namely č [t̪] and ẓd [t̪zd]. The OCS reflexes were found in, among other contexts, the 1sg. non-past—where they competed with the native mutations—and the past passive participle, where they were largely entrenched. Ultimately, č persisted in the 1sg. non-past but ẓd was driven out sometime in the early twentieth century (Baerman 2008, 85); however, the latter persists in past passive participles (e.g., rodit’-rožděnny ‘give birth’).

We propose, however, that the OCS ẓd is still a grammatical competitor for the 1sg. non-past mutation, and thus there is a morphophonological competition in ẓ-stems just as in t-stems. Beyond the aforementioned fact that this mutation is present elsewhere in the grammar (i.e., in certain past passive participles), there is some behavioral evidence that it is synchronically active, despite the fact it is rarely if ever found in the written standard. Sims (2006, chap. 6) administers a cloze task in which native speakers are asked to produce the 1sg. of a (defective) verb shown in the infinitive: several participants select the supposedly-extinct OCS mutation (e.g., ubediti’-ubeždu ‘convice’). Sloussar and Kholodilova (2013) find that this OCS mutation is used with verbs borrowed into the 2nd conjugation in a corpus of web text. For instance, the English borrowing zafrenditi ‘include on one’s friend list’ exhibits variation between the expected 1sg. zafrendžu and zafrendu. Together, these suggest that the OCS alternation is synchronically active in the 1sg. non-past, even if it is not prescribed for any native verbs. We propose the following two rules for t, ẓ-stem mutations:17

(10) \[
\begin{align*}
\{t\} & \longrightarrow \{č\} / -[+1, -PL, -PAST] \\
\{d\} & \longrightarrow \{zd\}
\end{align*}
\]

(Condition: $\sqrt{sokratin}, \sqrt{sidi}, \ldots$)

(11) \[
\begin{align*}
\{t\} & \longrightarrow \{č\} / -[+1, -PL, -PAST] \\
\{d\} & \longrightarrow \{zd\}
\end{align*}
\]

(Condition: $\sqrt{metčin}, \sqrt{lidi}, \ldots$)

The condition on lexical gaps can then be directly applied to the t, ẓ-stems. Table 5, adapted from table 3 of Pertsova 2016, gives frequencies of the competing dental stop mutations.18 The first mutation (10) is the majority pattern but it cannot tolerate the 39 exceptions represented by OCS-mutation stems ($\Theta_{172} \approx 33$).

This account cannot be stretched any further to cover 2nd conjugation stems ending in the dental fricatives s and z, which also exhibit gaps, simply because there is no inherited OCS competitor. However, there is some evidence that a competitor

| (s) | \[ | \rightarrow \{č\} | 39 \\
| d | \[ | \rightarrow \{čd\} | 58 \\
| \[ | \rightarrow \{zd\} | 75 \\
| \[ | \rightarrow \{zd\} | 0 |

---

16 The only clear exception is the labial stem zaimit’ ‘to eclipse’. Moskvin (2015, cited in Pertsova 2016) argues this gap is due to phonotactic illformedness.

17 Pertsova (2016) distinguishes between t-stems and st-stems, presumably because the s of the latter is never present in the 1sg. non-past; e.g., vorrustiti’-voručču ‘raise, cultivate’, but *vyroz[a][fel]. However, this can be handled with an additional—surface-true—phonological rule simplifying the resulting sibilant cluster, as [stʃ] is not a valid onset in Russian (Vitaly Nikolaev, p.c.).

18 As in Spanish, many Russian verbs are derived from prefixation to stems, but with rare exceptions, all derivatives of a stem undergo the same mutation in the 1sg. non-past. Therefore, Pertsova (2016) collects counts over stems rather than full verbs.
has arisen. Slioussar and Kholodilova (2013) identify a substantial number of English verbs nativized as s, z-stem 2nd conjugation verbs; a few examples are given below:

(12) fiksáti ‘fix’
kapsáti ‘use capital letters’
brouwáti ‘browse’
rejzita ‘raise’

Slioussar and Kholodilova find that for many of these borrowed verbs (including those examples above), the expected mutation (e.g., kapsáti) is rarely applied in a corpus of Russian web text; instead, speakers apply no mutation at all (kapsáti). We do not have any explanation for the emergence of this pattern (though see Sims 2006, 200 for the suggestion that non-alternation in dental stems was also present in pre-revolutionary Russian), and we do not have comparable lexical counts for this pattern. However, we do note that the conventional 1sg. non-past mutations ($s \rightarrow \tilde{s}$, $z \rightarrow \tilde{z}$) are themselves sufficiently rare that it would not take many non-alternating borrowings for this context to satisfy the condition on lexical gaps.\footnote{Petrušova (ibid.) counts 38 non-defective $s \rightarrow \tilde{s}$ stems and 36 non-defective $z \rightarrow \tilde{z}$ stems. Assuming these are related by a single rule of retroflexion, as seems likely, they can tolerate no more than 20 non-alternating exceptions.} We tentatively propose that may be exactly what has occurred.

To summarize, we argued that gaps in the 1sg. non-past of $t, d$-stems are predicted by the condition on lexical gaps. While it is more difficult to relate this to well-known synchronic processes—at least those documented in standard registers—we also suggest that gaps in $s, z$-stems may also result from competition between two morphological rules. However, these conclusions must remain tentative until confirmed by other sources of evidence, e.g., errors in child language acquisition.

6 Conclusions

Above we have shown how the Tolerance Principle can, from a numerical basis, detect productive generalizations as well as predict the existence of lexical gaps in those cases for which no productive generalizations can be found. However, we note that the lack of a productive generalization is only one of many pre-existing conditions with which a word can be afflicted; the principle has nothing to say about, e.g., taboo avoidance, phonotactic illformedness, or why one language realizes in a single phonological word that what requires periphrasis in another. It can only be invoked in cases where gaps are associated with competing, lexically-specific linguistic generalizations. While the cases above are all essentially morphophonological, the principle has also been applied in syntax (Yang 2016, 191f.) and phonology (Yang 2017).

The Tolerance Principle is extremely general, but the principle, and the child language learner if models, need precise statements of these generalizations and their structural properties so as to tabulate the sets of target lexical items (N) and exceptions (e). Both depend upon a well-defined hypothesis space to determine what is to be counted. Furthermore, the quantitative analyses above depend in part on the lexical resources used. Had we operationalized other analyses, or used other lexical resources, we might have reached slightly different conclusions. Therefore we encourage others to reproduce our results and to apply the principle to novel cases and new languages.

We also stress that the Tolerance Principle is simply the consequence of a theoretically and empirically motivated processing model, and is itself not a model of how productivity is encoded in the grammar. While the mere existence of lexical gaps is seen as an uncomfortable fact for many theories of grammar, the proposal here is applicable to any grammatical architecture in which the application of generalizations conforms to the Elsewhere Condition. Morphological productivity has, for example, often been linked to the distinction between inflectional and derivational morphology. Under one account (see Spencer 1991 passim for a review) the former is the result of a productive “narrow” syntactic computation whereas the latter is the result of a less-productive pre-syntactic (“lexical”) computation. But as Halle (1973) already notes, the very existence of gaps in inflection shows that productivity cannot be a definitive characteristic of the inflection/derivation dichotomy once we recognize the connection between unproductivity and gaps. Under the present account, productivity is a property inferred from the primary linguistic data by the child language learner. While the grammar induced by the child must encode this distinction, a theory of this inference can be developed independently from theories of the encoding of productivity.

Learning under the Tolerance Principle proceeds conservatively, requiring considerable evidence before a generalization is enounced. But by no means is this the only trajectory in language acquisition. In syntactic development, in particular, the learner may spontaneously, and productively, access non-target but linguistically possible options such as those in the parametric space of variation, which are then gradually eliminated as disambiguating evidence from the input accumulates over time (Yang 2002, 2010). The learning of morphology decidedly does not proceed this way: the hypothesis space for past tense formation is not one which allows the child to pick from either the English -ed or the French -é. These different patterns of acquisition may reflect familiar dichotomies such as functional vs. lexical heads, and innate parameter spaces vs. induced language-specific generalizations.

The Tolerance Principle, while applied narrowly here, is fundamentally an evaluation metric using processing time to rank competing grammars. Thus it represents a concrete proposal for the role of computational efficiency—the “third factor” (Chomsky 2005)—in the design of language.
References


When Nobody Wins


