8

The Origin of Linguistic Irregularity

Charles D. Yang
Yale University

1. The Challenge of Imperfection

The ban on the discussion of language evolution by the Société de Linguistique de Paris in 1866 surely ranks among the most defied gag orders ever issued. While there has never been shortage of evolutionary speculations on the origin of language, recent years have seen an explosive growth of work, emerging from a high-profile biannual international conference, a monograph series at the Oxford University Press, and regular publications in leading journals such as *Nature* and *Science*. The renewed enthusiasm in the evolution of language was made possible by the advances in the study of language and related topics. Fifty years of modern linguistic research has revealed more about the nature of human language than the Société could have ever imagined.

But it is important to keep in mind that deeper understanding of language only makes the game of language evolution harder to play. Everyone can tell a story about free will because we have no faintest idea how free will works, or what it is, for that matter. Games without rules are the easiest kind.

Modern linguistics has given us a good idea of what the *result* of the evolution of language is; we must now reconstruct the *process*

that might have led to it. Just as no one will take seriously a theory of how wings evolved, if it is devoid of an explanation of why wings are the way they are but not of some other imaginable form or structure, no one will take seriously a theory of how language evolved if it is devoid of an explanation of why language is the way it is, but not some other logical, but empirically unattested, possibility. The challenge is the classic problem of form and function in evolution. It is easy to tell post-hoc stories how language is useful for this or that function,[1] it is far harder to understand why language, as we see it in the world around us, has *this* form but not *that*. The tension is particularly acute when the observed form appears to be functionally inferior to some conceivable alternative.

One such feature is the phenomenon of irregularity in the lexicon. The English verb past tense is of course the best-known case. Of all English verbs, about 120 are irregular; the rest are regular, forming past tense by add -d, along the line of *walk-walked*. Another example: the noun plurals in German. German plurals fall into five classes: *Kind-er* ("children"), *Wind-e* ("winds"), *Ochs-en* ("oxen"), *Daumen-ø* ("thumbs", with a null suffix), and finally, *Auto-s* ("cars"), a class which, despite having the fewest members of all, is nevertheless the default (Marcus et al., 1995). The pattern of irregularity also holds for languages without (obvious) overt morphology. Chinese, for example, has a classifier system with irregulars and a default: *yi **tou** zhu* (a pig), *yi **zhi** yang*} (a sheep), *yi **pi** ma* (a horse), each referring to a specific kind of objects/nouns, whereas *ge* is the default, which can be used with novel or nonsense nouns.

Irregulars, by definition, are unpredictable, which means that special attention has been paid to them by language learners and users. Yet one can easily imagine a lexicon without irregulars, in which the morphophonological uses of all words are completely

[1] And it is true, many features of language do seem to be superbly adaptive, e.g., the arbitrary association between sound and meaning, the recursive mechanism that forming words and sentences to express thought, etc.

predictable. For example, we may imagine a language in which nouns ending in a vowel add -t to form plurals, and those nouns ending in a consonant add -o. From every functional perspective, the imaginary system would be far easier to learn, produce, and process, and thus posing considerably less cost to the cognitive and perceptual systems. Moreover, no principle of language, as far as I know, prohibits a regular lexicon like this. But lexical irregularity shows up, in one form or another, in all languages that we know of. Now *this* is a non-trivial fact of language that a non-trivial theory of language evolution would like to explain.

I would like to suggest that the presence of lexical irregularity results from the mechanisms of how words are learned. My argument — it's a long one — takes the following form. First, based on the much-studied problem of English past tense, we will argue for a model of word learning that sharply differs from all previous approaches; specifically, Steven Pinker's dual-route Words and Rule model (1999). Section 2 summarizes the developmental evidence for our alternative model, which consists of two components: (a) how phonological rules are learned, and (b) how such rules are used in word learning. Section 3 explains the algorithmic processes that underlie these two components, which, we shall argue, may involve domain-general abilities. The possibility that the learning mechanism used in word learning is not unique to language suggests that this mechanism might have been present *prior* to the emergence of language, and thus would have served as a constraint that shaped the properties of language and the outcome of language evolution. Finally, the model of word learning is extended to a model of sound change over time, with which we will show that irregularity in words is (almost) an inevitable outcome of how words are learned.

2. The Reality of Phonological Rules

There is probably no problem in cognitive science that occupies more minds than the acquisition of English past tense; witness the 15 years of debate in *Cognition*, and the introduction intended for

the general public (Pinker, 1999). There are three key empirical findings. First, since Berko's classic work (1958), we know that children, like adults, generally inflect novel verbs by adding the -d suffix. Second, Marcus et al. (1992) show that about 10% of English children's irregular verb uses are *overregularization* errors, e.g., instead of *hold-held*, the child may say *hold-holded*. Finally, over-*irregularization* errors, such as *bring-brang* where the child misused an irregular pattern, are extremely rare; only 0.2% of all irregular past tense uses (Xu and Pinker, 1995).

2.1 Two models

According to the Words and Rule (WR) model (Pinker, 1999), the irregular verbs are memorized as associated stem-past pairs, following the connectionist literature (Rumelhart and McClelland, 1986), and the regulars are computed by the rule "add -d", following the tradition of generative linguistics (Halle, 1962; Chomsky and Halle, 1968). This is illustrated in Figure 1. Since words don't carry (ir)regularity tags, the WR model requires the learner to distinguish regulars and irregulars. In other words, the learner must learn that "add -d" is the regular rule, when he is exposed to a mixture of both regulars and irregulars. Here, the case of German plurals becomes relevant (Marcus et al., 1995); how does the learner conclude that the smallest class in fact is the default? We will return to this question in Section 3.

The approach in the generative phonology is different. It asserts that the computation of all verbs, irregular or regular, is rule-based, and this is the approach we shall pursue. According to Lexical Phonology (Kiparsky, 1982), the rules for the irregulars are lexical/special; they are defined over particular words, and only these words. The default rule, in contrast, is general: it is not lexically restricted and can in principle apply to all words. The effect of irregularity is achieved by a principle of ordering, the *Elsewhere Condition*, which states that when multiple rules are applicable, the most specific will be used. Hence, an irregular lexical rule, which

Figure 1
Past tense according to Words and Rule.
$W_0 \ldots W_5$ denote irregular verbs, which are directly associated with their past tense, $O_0 \ldots O_5$.

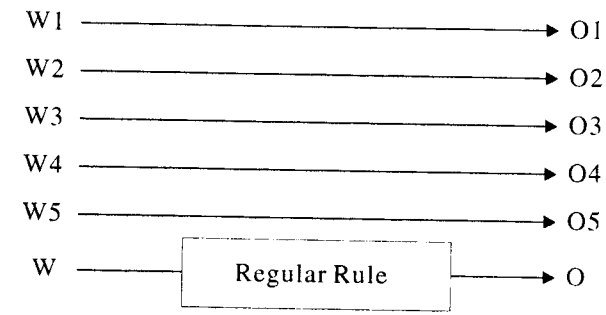
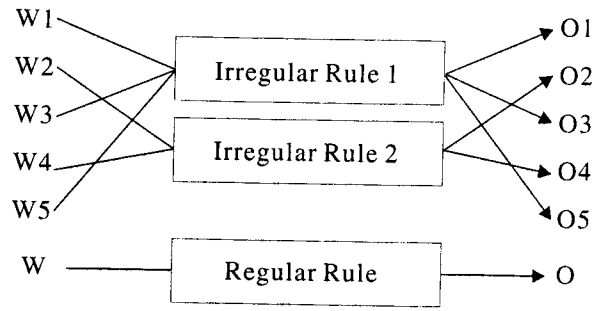


Figure 2
Past tense according to Rules over Words.
The irregular rules take irregular verbs $W_0 \ldots W_5$ as input, and produce output $O_1 \ldots O_5$, the irregular past tense.



specifically refers to particular (irregular) words, will override the default; hence, we have *hold-held* rather than *hold-holded*. Because both irregular and regular words are computed by rules in this approach, let's call it the *Rule over Words* (RW) model. It is illustrated in Figure 2.

For the purposes of this section, we will show that irregular verbs are indeed learned in classes, which are defined by rules that multiple verbs share, just as suggested in the traditional approach to phonology; we will turn to how such rules are learned in Section 3. The rules are, in fact, not so different from those stated in pedagogical grammars, along the lines of:

(1) a. bring, buy, catch, seek, teach, think: add –t & Rhyme → /a/
b. cost, cut, hurt, let, put, quit, set, . . . : add -ø & no change
c. blow, draw, grow, fly, know: add -ø & Rhyme → /u/
d. feed, shot, lose, leave, shoot, . . . : suffixation (-d, -t, and -ø) & Vowel Shortening
e. . . .

In what follows, we will present acquisition evidence in favor of the RW model over the WR model. For a detailed discussion of past tense learning and related issues, see Yang (2002a: Chapter 3). The acquisition data is taken from the longitudinal study of four American children in Marcus et al. (1992). The children and the percentage of their correct past tense use are: Adam (98.2% = 2446/2491), Eve (92.2% = 285/309), Sarah (96.5% = 1717/1780), and Abe (76% = 1786/2350).[2] To evaluate the two models, and to quantify the linguistic data during past tense acquisition, we have obtained the frequencies of past tense irregulars from the adult sentences (transcribed in CHILDES) that these four children were exposed to.

[2] Although Abe's performance is considerably worse than the other children, it is not because a few high-frequency verbs were used badly. Rather, Abe's past tense is problematic across the board; cf. Maratsos (2000).

2.2 The failure of frequency

In the WR model, the association between an irregular verb and its past tense is established by association. Thus, association follows a simple principle of memory (Pinker, 1995, 1999): the more you hear, the better you remember. Indeed, Marcus et al. (1992) found strong correlation between adult frequency and child overregularization errors –.33.

However, having Bill Gates in a bar, though raises the average income of the patrons, does not make everybody rich. Similarly, collapsing all the irregular verbs together will very likely obscure important differences among these verbs. Figure 3 shows the frequency-performance correlation of the irregular verbs that appeared in the children's production significantly often (> 25 times). Despite the fact that we are only looking at a subset of children's irregular verbs, the frequency-overregularization correlation is still –0.32, comparable to that (–0.37) reported in Marcus et al. (1992) for all irregular verbs. We plot the (logarithm of) adult frequencies along the X-axis, and children's performance along the Y-axis. If the WR model is correct, then we would at least expect a more or less monotonically upward-moving curve, following Pinker's memory principle; that this is untrue you can see for yourself.

Some of the verbs, such as *bite-bit* and *shoot-shot*, were heard 20–30 times less than *get-got* and *put-put*, yet they were used nearly perfectly, 89.2% and 93.8%. Some, such as *threw* and *knew*, are a few times more frequent that *bit* and *shot*, but were used far worse: 32.4% and 73.9%. Such examples are abundant, as Figure 3 makes clear.

Let's see how such frequency-performance disparity can be explained. In the RW model, learning an irregular verb consists of two parts: the learner will have to associate the verb with a specific lexical rule, which also has to be constructed as part of learning. The verb-rule association is established upon exposure to the particular verb in past tense; this can be quantified by its token frequency. The rule, however, can be established whenever *any* verb

that falls under it is encountered. In other words, by the virtue of being shared by multiple words, the learner's experience with a rule is actually the *sum* of the token frequencies of all verbs that fall under it. This two-step process implies that the performance of a

Figure 3. Frequency effects under the WR model

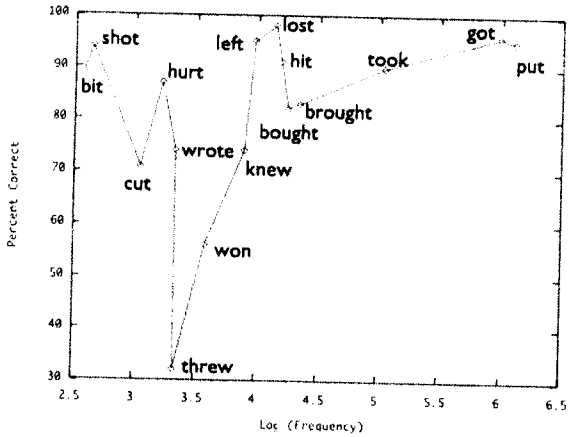
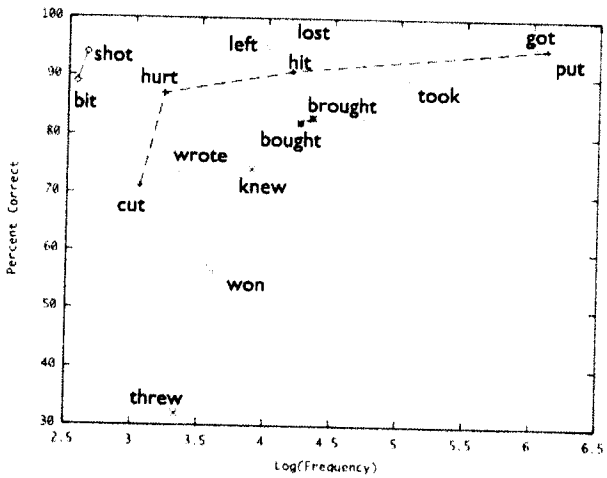


Figure 4. Frequency effects within irregular classes



particular irregular verb, W, will be correlated with the *product* of two frequencies: that of W, and that of the rule R of which W is a member, or $f_W \sum_{i \in R} f_i$. This leads to two quantitative predictions:

(2) a. for verbs within the same class, token frequency will determine relative learning performance.
b. for verbs with comparable frequencies, the size of their respective *classes*, i.e., the sum of token frequencies, determines relative learning performance.

Figure 4 examines the prediction in (2a); we see that once verbs are examined in classes, frequency-performance correlation is perfect, with the exception of one verb (*win-won*) in one class. The reader is referred to Yang (2002a) for a detailed comparison of the individual verbs, including raw statistics.

2.3 The Free-rider effect

The WR model fares far worse once we compare verbs that belong to *different* classes; here the frequency-performance correlation completely breaks down.

First, there are verbs for which adult frequencies are comparable (about 20 out of 110,000 sentences), but children's performance differs significantly: *hurt* and *cut* were used 80% correctly, and *draw*, *blow*, *grow*, and *fly* were used only 35% correctly.

Second, some low-frequency verbs are used much better than higher-frequency ones. Again, *hurt* and *cut* were used about 20 times by adults but have a correct usage rate of 80%; in contrast, *knew* and *threw* were used by adults 58 and 31 times respectively, but were used only 49% correctly.

Finally, for Abe, whose past tense learning is quite delayed compared to Adam, Eve, and Sarah, some of the most frequent irregular verbs were used worse than some with very low frequencies. For example, Abe used *hurt* and *cut* correctly 66% of time: that's better than *go-went* (64% correct), which was used by adults 557 times, and *come-came*, which was used by adults 262 times.

These frequency-performance, or input-output, disparities have a straightforward explanation in the RW model; see (2b) above. Verbs may have high performance if they belong to large classes; recall that exposure to every member of a class contributes to the learning the shared rule. Once the effect of rules is taken into account, we see that *hurt* and *cut*, while rare in the input, nevertheless belong to a very large class: the "no change" class, which include very frequent items such as *let*, *set*, *put*, etc., which tally up to over 3,000 occurrences in the input — that's even more frequent than *went* and *came*, two of the most frequent irregular verbs, which nevertheless act alone. The "Rhyme→/u/" class, which contains the problematic verbs *drew*, *blew*, *grew*, *flew*, *knew*, and *threw*, only have 125 tokens altogether. Hence, lower or comparable frequency can be enhanced by class frequency — a Free-rider Effect — confirming the predictions of the RW model.[3]

2.4 General process in special words

Finally, there is a class of irregular verbs that was used very well by all children, almost irrespective of their frequencies. They are shown in Table 1.

Table 1. Vowel shortening irregular verbs

Word	Input Frequency	Percent Correct
lose-lost	63	98%
leave-left	53	96%
say-said	544	99%
shoot-shot	14	94%
bite-bit	13	90%

[3] Also worth noting is the verb *caught*. It was used by adults 36 times — compared to 58 times for *knew* and 31 times for *blew* — but used 96% correctly by children. The reason is that the Rhyme→/a/ class is also large: it includes *thought*, which alone appears 363 times in the input.

All these verbs form past tense by adding a suffix (-t, -d, or -ø), which is followed by the process of Vowel Shortening (Halle and Mohanan, 1985). Vowel Shortening under suffixation happens to be a very general process in the English language, and can be observed in following examples (Myers, 1987):

(3) a. [ay]-[ɪ]: divine-divinity
b. [i]-[ɛ]: deep-depth
c. [e]-[æ]: nation-national
d. [o]-[a]: cone-conic
e. [u]-[ʌ]: deduce-deduction

It has been argued that Vowel Shortening falls out of the interaction between universal phonological principles and the way in which syllabification works in English (Myers, ibid; Halle, 1998). If this is the case, then the rule of Vowel Shortening is essentially given for free. Consequently, the task of learning Vowel Shortening verbs is greatly simplified, having been reduced to learning the particular suffixes.[4] Given the fact that cross-linguistically, acquisition of affixal morphology is near perfect (Phillips, 1995; Guasti, 2002), children's impressive performance on vowel shortening verbs is expected. It is important to note that, in order to explain the acquisition data of the Vowel Shortening verbs, one must appeal to the overall sound patterns of the language. This is something that the WR model is in principle incapable of.

Note that both RW and WR models make use of memorization to explain irregular acquisition. Irregulars are, by definition, unpredictable, and must be memorized, somehow, on an individual basis. The two models differ in how the verbs are memorized. The WR model memorizes the past tense of verbs directly, whereas the RW model memorizes word-rule associations, after which rules

[4] It is certainly not the case that merely a dozen repetitions, as in the case of *shoot-shot* and *bite-bit*, suffice for near perfect learning; recall from earlier discussion that verbs in other classes are used far worse despite higher frequencies.

apply to stems to generate past tense forms. In both models, no memorization is used for regular verbs, and in both models, when a verb isn't one of the irregulars, the default rule will pick it up. Much of the empirical work in the WR framework, ranging from language acquisition to online processing to cognitive neuroscience to aphasiology (see Pinker, 1999 for a review), focuses on the irregular vs. regular dichotomy, that is, the unpredictability of the irregulars, which calls for special memorization; on this score, both models fare equally well.

Summarizing the acquisition findings, we see that irregular verbs are learned and organized in groups, by the means of irregular and regular rules. The word-rule association is statistical in nature, and this is an amendment to the classical conception of phonology, where rule application to words is categorical. Section 3 outlines a computational model that learns these rules and establishes word-rule associations.

3. The Computation of Rules

3.1 What makes a rule default

As noted earlier, in order for the WR model to work, the child must be able to identify, amongst the various sound change patterns in past tense, that "add -d" is special. Only after the -d rule is learned can the child sort verbs into the regular and irregular bins, and proceed to memorize the irregulars individually.

Clearly, the default cannot be identified with the rule covering most verb tokens. Among verbs with highest frequencies in English, most are irregulars; indeed, irregular verbs make up 60% of the probability mass in English verbs (Grabowski and Mindt, 1995). It is then suggested (Pinker, 1999) that the default is one that covers most verb *types*; since there are only about 120 irregular verbs, this idea surely works for English.

But serious problems arise for German noun plural acquisition. Marcus et al. (1995) have established, using the Wug test and others on German speakers, that the "add –s" rule is the default. However, the -s class is the smallest among the five plural classes, four of which are irregular. Looking for the default as the rule with the dominant type frequency will get the German learner nowhere.

In addition to the German challenge, Pinker's view of default learning faces two other problems. First, it is unclear how much computational power will be needed for the child to keep track of frequencies of multiple classes to determine the statistically dominant default class. Second, and empirically, there are morphological systems with no defaults. For example, the genitives in Polish (Dabrowska, 2001) have three case markers, each restricted to a subset of nouns, and none is the default. That is, none of three markers passes the standard suits of benchmarks including the Wug test. This is an awkward problem for a learning model that *looks for* a default rule/class based on type frequency (or anything else, for that matter).

So, a successful model for rule learning will have to meet the following conditions:

(4) a. It must handle the statistical minority of the default class in German plurals, and
b. it must *not* require the presence of a default rule to work and yet it must be equipped to learn it if present in the learning data.

3.2 Rule learning by induction

In an important series of papers, Sussman and Yip (1996, 1997) provide, as far as I know, the only model that fits the bill. The following discussion draws from Molnar (2001), which is an extension of that work. In the S&Y model, the learner constructs phonological rules as mapping relations between input (e.g., stem) and output (e.g., past tense, plural) forms. Both input and output

are represented as a linear sequence of phonemes. Each phoneme is represented by a universal set of distinctive features (Jakobson, Fant, and Halle, 1951; Chomsky and Halle, 1968; Halle, 1983). For instance, /ae p l z/ ("apples") is represented as follows:

Table 2
The feature representation of "apples" in the Sussman-Yip model.

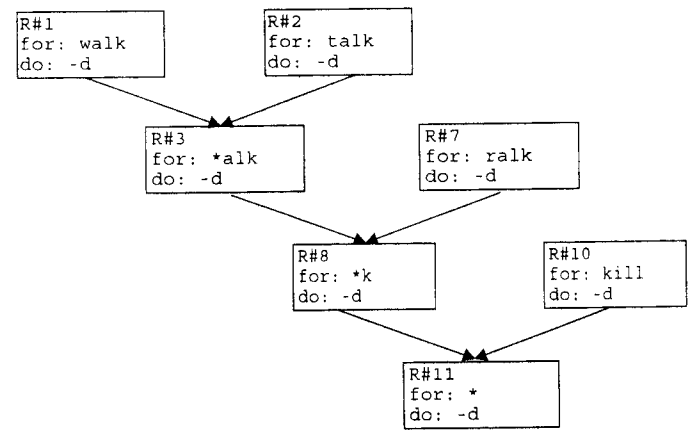
	[æ]	[p]	[l]	[z]	"apples"
syllabic	1	0	0	0	
consonantal	0	1	1	1	
sonorant	1	0	1	0	
high	0	0	0	0	
back	0	0	0	0	
low	1	0	0	0	
round	0	0	0	0	
tense	0	1	0	1	
anterior	0	1	1	1	
coronal	0	0	1	1	
voice	1	0	1	1	
continuant	1	0	1	1	
nasal	0	0	0	0	
strident	0	0	0	1	

The learning algorithm is simple, and its intuition is based on how phonological rules are represented. Traditionally, rules take the following form (Halle, 1962; Chomsky and Halle, 1968):

(5) $A \rightarrow B \,/\, C\_\_\_D$

That is, a computational process (A changes to B) takes place in a specific context (between C and D). In the present model, the learner constructs phonological rules inductively: for words that undergo identical sound change, i.e., $A \rightarrow B$, it tries to find what they have in common, i.e., $C\_\_\_D$. An example in Figure 5, adapted from Molnar (2001), illustrates how the -d rule is learned.

Figure 5. The induction of the default -d rule



When learning starts, there are no words or rules. Suppose now the first word, *walk-walked*, comes in. The model identifies the phonological change from stem to past by comparing the representation of *walk* and that of *walked*, and establishes that the relevant change is "add -d".[5] Since this is the only piece of learning data available, the learner can draw no generalization but store it by route as a trivial rule: **if** *walk* **then** "add -d".

Suppose that the next word is *talk-talked*. The learner will follow same procedure to obtain: **if** *talk* **then** "add -d". Now the learner is ready to make inductive generalizations. The two statements thus far constructed share the **then** clause — they undergo identical phonological change in past tense. The learner then tries to discover what they have in common in their respective

5 In the implementation, the "ed" suffix in "walked" is actually a /t/, and the whole word is represented as a feature matrix.

phonological descriptions; namely, what can be generalized from *walk* and *talk*.

The induction algorithm works maximally conservatively; conservatively follows both logical (Berwick, 1985) and empirical (Clark, 1993) principles of acquisition. It compares two phonological representations, and for phonemes that have conflicting phonological features (+ and –), it places a* ("don't care") in the generalization. In our example, /walk/ and /talk/, which only differ in the initial phoneme, are generalized to /*alk/. The learner now considers all words that fit the description /*alk/, that is, all verbs that end in /alk/, to undergo the change "add -d" in past tense.

As illustrated in Figure 5, as more words come in, the learner will continue to carry out the procedure described above. It is clear that, after a few regular verbs are presented, the condition for "add -d" will get very general: conflicting feature values, due to the diverse phonological shapes of regular verbs, will lead to more *'s in the generalization. Eventually, the learner will determine that *anything* can take -d in past tense, where "anything" is represented by *'s across the board in the if clause. And this is what phonologists call the default rule.

The computer simulation actually returns three rules for regular verbs:

(6) a. Verbs that end in a voiced phoneme but not a d:
```
[*.*.[+voice,+sonorant].d]
[*.*.[+voice,-coronal].d]
[*.*.[-low,-round,-tense,+continuant],d]
```

b. Verbs that end in an unvoiced phoneme but not a t:
```
[*.*.[-voice,+strident].t]
[*.*.[-voice,-coronal,-continuant].t]
```

c. Verbs that end in (d, t):
```
[*.(d,t).I.d]
```

They in fact match exactly the phonological rules that linguists would use to describe regular past tense.

Rule learning is very efficient in the S&Y model. Using the training data in MacWhinney and Leinbach (1993) and Ling and Marinov (1993), the S&Y model outperforms all previous implementations of past tense learning. When trained on regular verbs only, the model achieves 99.8% accuracy on prediction of regular past tense with only 30 examples. In contrast, the Ling-Marinov model and MacWhinney-Leinbach model only have a 90% prediction accuracy after 500 training examples, with the former learning faster than the latter. When trained and tested on both regular and irregular verbs, the S&Y model achieves 95% accuracy in prediction after merely 60 examples, the Ling-Marinov model, 76% with 500 examples, and the MacWhinney-Leinbach model, 57% with 500 examples. In addition, the Sussman-Yip model is able to learn verb past tense and noun pluralization at the same time, again, producing rules that are perfectly acceptable to phonologists.

Irregular rules can be similarly learned; some of the results from computer simulation are given below, with the rules and irregular verbs that follow them.

(7) a. `[*.*.i->ae.ng]` rang, sang
b. `[*.*.E->a.t]` forgot, got, shot
c. `[*.E.n.->t]` bent, lent, meant
d. `[*.*.(r,l),*->u]` blew, drew, grew, fly
e. `[*.*.*->a.t]` bought, brought, caught, taught, thought
f. `[*.*.*->o.z]` chose, froze, rose

The current implementation has a number of problems. First, it does not distinguish special and general rules, and this gives the impression that the irregular rules are also productive. For example, the statement in (7a) `[*.*.i->ae.ng]` for the *sing-sang* indicates that *all* verbs with an /ing/ ending will change /i/ to /ae/, which is obviously incorrect ("bring" and "wing" are counterexamples).

The solution to this problem is beyond the scope of the present paper. It boils down to this: when is a rule lexically restricted to certain words that the child has to identify in the learning data, as opposed to becoming generally applicable (to novel items as well)?

As far as I know, the productivity problem of phonological rules has been addressed as a problem of learning. In other work (Yang, 2002b), I suggest that the learner strives to maintain a balance between the productivity of a rule and the number of exceptions it has to explicitly maintain. For example, if the rule [*.*.i->ae.ng] were productive, it must mark *bring*, *wing*, etc. as exceptions, because they do not follow under the pattern it describes. Under reasonable assumptions about how words are stored and accessed, it is possible to derive formal results on the mechanism the child learner may use to draw the special/general distinction. Roughly, if a rule has too many exceptions, the learner will regard it as lexically marked. Such, I suggest, is the fate of the irregular rules listed in (7). With an independent model of what makes a rule lexical, we can maintain the current model of rule learning. And interestingly, children do occasionally say "*bring-brang*" at a younger age, which is the only somewhat robust pattern of the very rare over-irregularizations (Xu and Pinker, 1995). This suggests that the [i->ae.ng] rule was productive early on in past tense acquisition. Presumably, if the only /ing/-ending words the child knows are *sing*, *ring*, and *sting*, and they both follow the [i->ae.ng] pattern, the child may well assume the rule to be productive for words with /ing/ endings. It is the accumulation of exceptions (*bring*, *wing*) to the rule that demotes it to special/lexical status.

The other problem with the learning model is that it amounts to a two-level model of phonology, with direct mapping between the underlying form (the stem) and the surface form (the past tense). It has no notion of rule ordering. Hence, all the problems associated with two-level phonology, e.g., KIMMO (Koskenniemi, 1983), will be implicated here as well (Anderson, 1988). Moreover, the representation of words as a linear sequence of phonemes does not reflect the nonlinear representations adopted in modern phonological theories. And although the model can behaviorally replicate its effect, Vowel Shortening is not learned as a general and unified process in the language. Nevertheless, we believe that the basic algorithm of finding generalization through diversity of phonological representation provides a basic framework suitable for rule learning and can be augmented with richer phonological principles and constraints.[6]

We can now address the two problematic lexical systems for the WR model, namely, Polish genitives and German plurals. First, the Polish problem is not a problem. The emergence of the default rule is facilitated by the *learning data*, not as a necessary part of the learning algorithm. If three rules provide a complete (and disjunctive) coverage of words, as in the case of Polish genitives, so be it; there is nothing in the learning model that forces the existence of a default.

The German plural challenge is also straightforwardly resolved. The default, according to the rule learning model, is identified with the general rule that has *'s across the board, one which imposes no restrictions whatever on the word. Learning the default has nothing to do with statistics. In German, the default -s class consists of largely loan words, many from English, e.g., *Auto-Autos*, *Radio-Radios*, etc. Therefore, the nouns for the "add -s" class, being sufficiently diverse phonologically (like the regular nouns in English), will quickly lead the learner to recognize that the "add -s" rule has no phonological restrictions on the noun.[7] The irregular nouns, in contrast, are associated with four other irregular rules, which are more restricted.[8]

The acquisition evidence reviewed in Section 2 strongly suggests that the mental lexicon is structured by morpho-phonological rules,

---

6 See Ristad (1994) for a theoretical formulation of learning ordered rules.

7 Indeed, in the Y&S model, the -s rule can be learned after on average only 10 English noun plurals.

8 However, they do seem to be productive, but only with respect to nouns with particular morpho-phonoligical properties (and are hence less general than the default -s rule). The claim that the -s rule is the default is correct, but slightly misleading. Such claims are established on the fact that when German speakers are presented with novel nouns, the -s rule is often used (Marcus et al., 1995). However, if the novel noun in fact obeys the general morpho-phonotactics of German, an irregular rule is used. See Pouplier and Yang (2003) for discussion.

much as suggested by classical generative phonology (Halle, 1962; Chomsky and Halle, 1968). The present section provides a companion model for learning rule-based phonology. It remains to be seen whether non-derivational approaches such as the Optimality Theory can provide an answer to the developmental and learnability problems posed by past tense, a most basic fragment of the English language.

3.3 Possible origins of rule learning

We now turn to the cognitive basis of the ability to learn and use rules in phonological learning, with the suggestion that this ability may be due to general mechanisms of learning that apply to other, non-linguistic, domains of knowledge.

Consider the inductive learning algorithm that finds commonality among words through their phonological feature descriptions. There are strongly parallel findings between past tense acquisition and classifier acquisition in Chinese (Hu, 1993; Myers and Tsay, 2000) and Japanese (Yanamoto and Keil, 2000). These classifier systems[9] also have defaults and irregulars, and both Chinese and Japanese children overregularize the default. Given that

[9] Eds.: Note that the term *classifier system* here refers to the noun classifiers used in such languages as Chinese. Classifiers are words that are used obligatorily to identify the category of a noun that is to be quantified; LaPolla briefly discusses classifiers with regard to the complexity of language in Part IV.
Fortuitously, the term *classifier system* is also used to refer to the computationally complete, rule-based, message-parsing system described by Holland in Part IV. Yang's *Rule over Words* framework and the classifier system have much in common: in particular, both are rule-based systems consisting of a set of rules that encode the default behavior of the agent, with exceptions to those rules acquired to improve the efficiency of the agent; furthermore, both Yang and Holland assume that specific, exceptional rules override general, default rules — a principle that Yang calls the Elsewhere Condition (see main text) — to generate a *default hierarchy* (see Holland, Part IV).

the use of classifiers is largely determined by semantics and conceptual categorization of nouns, we may interpret classifier acquisition as the same algorithm at work: conservative generalization for nouns sharing classifiers, and probabilistic association between nouns and classifiers. However, in classifier acquisition, the algorithm will have to operate on different (semantic) feature representations.

More generally, if there is inductive learning at all in human perception and cognition, or in other species for that matter, it must be carried out conservatively, for nothing useful can be learned otherwise. Suppose one observes that both Bush Sr. and Jr. are ready to start a war against Iraq; the rational, and conservative, conclusion is that only families representing the oil industry are inclined to conjure up B-52s for profit, not all father-son pairs.

Just learning rules is not enough; the learner also has to know how to use rules to organize words. Here the fundamental principle is the Elsewhere Condition, which asserts that more specific rules override the application of more general rules. Again I would like to suggest that the Elsewhere Condition has counterparts in other domains of human cognition.

The most relevant example can be found in the Gricean conversational Maxim of Quantity, Be Informative. If I were to tell a fellow linguist about the ACE meeting, I would say "this is a conference about language acquisition, change, and evolution," rather than "this is a conference about language." When alternative ways of speaking are available, we pick the most specific, which is of the same character of the Elsewhere Condition.[10]

Specificity over generality may also be seen in non-linguistic tasks. According to the FIFA official rules (2002), if a player deliberately handles the ball, it is a freekick; but if one deliberately handles the ball in the box, it's a spot kick. Interestingly, the rule for

[10] Gregory Ward once lost a bet (to Larry Horn) when he tried to get college students to call a square a rectangle; they wouldn't, even in contrived situations.

freekicks does not have the clause " . . . except in the penalty area". It is unwritten, but tacitly assumed, that human referees will adhere to the specificity over generality principle; it follows that handballs in the area are awarded with penalty kicks.

Or consider a poker hand: ♣K ♦K ♠K ♥Q ♦Q. It is a full house, although "two pairs" is also a possible, but not the most specific, description. Again, poker rules do not explicitly state that a hand with a pair and three of a kind *cannot* be two pairs — it is simply assumed that a more specific rule, if applicable, does apply.

These examples open up the possibility that the Elsewhere Condition was a derivative of a more general cognitive ability. This ability would have evolved before language, assuming that language was the latest major event in the evolution of cognition. It was then co-opted for learning and organization of words with rules, *after* the emergence of the faculty of language. The combination of this domain-general principle and the domain-specific knowledge of language gave us words and rules, and how they interact in the lexicon. Of course, the argument might just go the other way; it is also possible that the Elsewhere Condition was phylogenetically linguistic, and its use in other cognitive domains was a by-product of a fundamentally linguistic principle. A possible way to tease these alternatives apart is to see whether there is evidence for the principle of specificity over generality in other (non-linguistic) species.

4. The Evolution of Rules

4.1 Word drifts

The preceding sections suggest that word learning has two components: (a) rule construction, and (b) word-rule association, which is governed by the Elsewhere Condition. Under this view, an irregular verb V may be simultaneously pulled by two rules: (a) the lexical rule R that it is associated to, and (b) a productive rule R′ that may apply to V because V falls under R's phonological description yet doesn't, following the Elsewhere Condition, because the presence of the more specific R. Figure 6 illustrates.

Figure 6
A word with competing rules.

For example, V = "catch", R = "-t suffixation & Rhyme → /u/", and R′ = default. The V-R association is established by fiat, purely on the basis of repeated exposure to "caught", whereas the V-R′ association is automatic given the unrestricted applicability of R′.

Throughout the history of the English languages, many words have drifted from rule to rule. The so-called *analogical leveling* typically refers to an irregular verb becoming regular: for example, *cleave/clove/cloven* is now *cleave/cleaved/cleaved*, and for many speakers, *strive/strove/striven* has become *strive/strived/strived*; see Campbell (1998).

Another kind of word drift, *analogical extension*, refers to a regular verb becoming irregular. For example, the past tense of *wear* was *werede*, which would have been regular if survived to modern English, but in fact it took on the *bear-bore*, *swear-swore* class.

What would attract a word W from the rule it currently falls under, say A, to a different rule, B?[11] There are two logical possibilities.

First, if A is general and productive, and if B is more specific match for W than A, then W-B association is assured by the Elsewhere Condition. That accounts for the pattern of analogical extension: for example, W = *wear*, A = default, and B = [er->or], as discussed above.[12]

---

[11] The fact that W *does* drift from A to B clearly means that B is productive and applies to novel items; here it means that B matches the phonological description of W.

[12] Which means that when *wear* made the shift, [er->or] was a productive rule that would automatically convert /er/ to /or/; this is certified by consulting the OED.

Second, if A is special and hence unproductive, then the W-A association is established by fiat, on the basis of quantitative linguistic data during learning. If there is not "enough" evidence, W would escape the bounding of A and succumb to the attraction of B. This accounts for the pattern of analogical leveling: for example, W = *strive*, A = [i->o]], B = default.

The RW model provides a novel, and precise, quantification of how much evidence is "enough" to bind a word to a lexical rule. Recall that the success of learning an irregular verb W is positively correlated to the *product* of its token frequency, f_w and the token frequency of its class R, that is, $f_w \Sigma f_i$, where i and W belong to the same class. Hence, a verb can stay irregular due to either its own frequency or the high frequency of its class; the acquisition data analyzed in Section 2 provide strong evidence for this view, which can be called "Salvation by Volume". It contradicts the claim of Bybee and Slobin (1983) and Pinker (1999) that irregularity is maintained by high frequency alone, which can be called "Salvation by Height". While the frequency-based theory is largely correct for English past tense, it cannot be correct for German plurals. Recall that only about 8% of nouns in German are regular, which means the majority of nouns are irregular. According to the Salvation by Height theory, the majority of the irregular nouns, failing to match English irregular verbs in frequency, will have to drift to the default class; that is not what we see in German, however.

4.2 Evolutionary model of words and rules

We are now equipped to develop a model of word and rule change over time. The algorithm runs as follows:

(8) a. Start with a random set of words, each with a "phonological" description (e.g., 100 bits of 0's and 1's). The words are grouped into arbitrary classes. Assume that the frequencies of the words follow a Zipfian distribution.

b. For each generation

i. use the rule learning algorithm to derive a generalized rule based on the phonological description of words that belong to a same class.

ii. for each word W and $W \in R$, with a probability $\left(1 - \exp[-V_w/\tau]\right)$, where $V_w = f_w \sum_{i \in R} f_i$, and τ is a decay constant

- find a rule R′ which match the phonological description of W most closely without conflicting bits (Elsewhere Condition)
- associate W with R′

c. Repeat (8b).

An example will help the reader to understand how words and rules change. Suppose we have five words, and their phonological descriptions are W1 = [00001], W2 = [01101], W3 = [01001], W4 = [11110], W5 = [01100]. Their evolution over their generations may look like the following table:

Suppose that initially, words 1, 2, and 5 are in the same class, and the learning algorithm collapses conflicting features to derive a Rule [0 * * 0 *]. Similarly, words 3 and 4 lead to another Rule [* 1 * * *]. In generation 2, suppose word 2 drifts to class 1. Now rules have to be relearned from scratch, and notice that the rules learned in generation 2 do not happen to differ from those in generation 1. However, generation 3 sees the drift of word 3 to class 1; consequently two brand new rules emerge and the old rules disappear from the lexicon.

Figure 7 shows the result from a typical simulation with 500 words. Initially, they were randomly assigned into 50 classes. When the number of classes is stabilized, there are 6 classes.

In Table 4, the simulation converged with a very large class, with completely general descriptions; this we can interpret as the default rule. We also have 5 small classes with specific phonological restrictions. This strongly resembles the organization of the English past tense system.

Figure 7. A history of 500 words

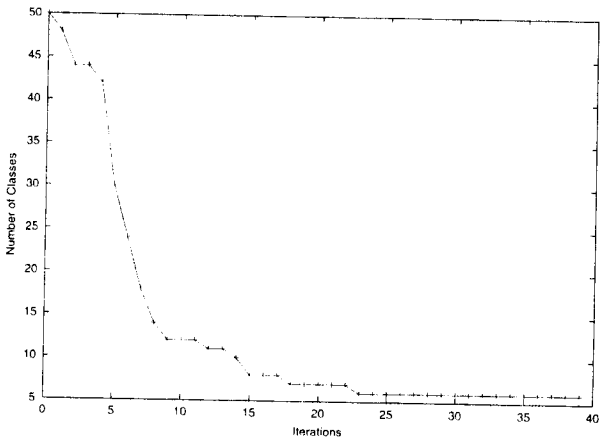


Table 3. A history of 5 words

Generation	Class 1	Class 2	Change
1	(1, 2, 5) [0 * * 0 *]	(3, 4) [* 1 * * *]	
2	(1, 5) [0 * * 0 *]	(2, 3, 4) [* 1 * * *]	W2 -> R2
3	(1, 3, 5) [0 * * * 0]	(2, 4) [* 1 1 * *]	W3 -> R1

Table 4. Six stable rules and their sizes

Rule	Size
**********	456
**********1**0**********0********	12
**********0**********1***	9
****1**********1**	5
*******0**********	9
**********1**0**********	9

It is interesting to note that the sharp reduction in the number of phonological classes in Figure 7 corresponds to the emergence of the default rule. Suddenly a completely general rule emerged, and it very quickly assimilated words from other classes.

Varying certain conditions in the evolutionary model can lead to results similar to German plurals and Polish genitives. In both cases, we may start with several very large irregular classes, i.e., classes that have many members that share part of their phonological descriptions. These classes may remain stable (though it is unlikely, as we shall see Section 5). In the case of German, we simulate the effect of foreign imports by introducing a small class of random words that differ substantially from those already in the rule system. Because of the phonological diversity of words in this new class, a default rule can quickly emerge that covers only a small number of words. It is unable to attract words from the larger irregular classes, because the class size may prevent the drift of even low-frequency words — Salvation by Volume, as noted earlier.

The evolutionary model is a logical extension of the independently motivated word learning model. It is thus of considerable relevance to the study of historical phonology. Our study shows that words can drift on an individual basis while leaving phonological rules intact. Also, for words that fall under a shared rule, some may drift away and some may not; see Table 3. This is consistent with the theory of lexical diffusion (Wang, 1969), and the reconciliation of lexical diffusion with lexical phonology (Kiparsky, 1986), but not consistent with the Neogrammarian regularity principle.

5. Interface Conditions and Language Evolution

Finally, we can address the imperfection of lexical irregularity. Again, imagine a logically possible morphological system, where all nouns ending in a vowel add **-t** to form plurals, and all those ending in a consonant add **-o**; it is systematic, regular, and neat. But one of the trademarks of human language is its "leakiness": in the lexicon

of the world's languages irregularity and exceptions abound, as noted at the beginning of this paper.

If the rule learning and evolution models provide an accurate description of reality, then we may have an explanation for the prevalence of irregularity. Suppose that a language did have a regular system just described, where two disjunctive rules give a complete and predictable coverage of the nouns. Suppose, as the result of language contact, a few foreign nouns entered into the native lexicon. This is a highly plausible assumption, given that language is used and transmitted by humans, and humans are mobile and social animals.

Suppose that, for example, all the foreign nouns add -s in plural forms regardless of their phonological properties. As we saw in the simulation, a small number of diverse words suffice to yield a general default rule: add -s no matter what a (native or foreign) word looks likes. Now the two existing rules have a competitor, and native nouns may start drifting to the default. The -t and -o rules will get more and more specific, and smaller and smaller. Irregularity follows.

As long as the learning data is diversified, through whatever means, there is a very good chance for a default to emerge, which subsequently may assimilate words from other rules and leave behind irregulars, a vestige of once regular rules. Hence, the "real" origin of linguistic irregularity may be historical and unpredictable:[13] the model of learning and change presented here helps us pin down the predictable effects of such unpredictable causes.

Our approach to language evolution can be seen as an execution of Chomsky's Minimalist Program (1995); see Hauser et al. (2002) for elaboration in an evolutionary framework. In the Minimalist Program, the faculty of language is viewed as a cognitive module that interfaces with the rest of human cognition; informally, the

[13] Innovation may also, unpredictably, introduce novel patterns into the lexicon, which can then lead to defaults and irregulars.

"meaning" module and the "sound" module, the so-called *interface conditions*.

We suggest that an additional interface condition lies in the ability to learn. Clearly, for a language to be usable, it must be learnable by children under normal conditions. Here, and at other places (for the learning of syntax, see Yang, 1998, 2002a), I have suggested that the ability to learn language may be due to a learning/growth mechanism in other cognitive and perceptual domains.[14] It then renders plausible the hypothesis that the learning mechanisms were earlier evolutionary products.

In order for language to be usable at all, it must satisfy these interface conditions. To make an analogy, imagine the mind/brain as the motherboard of a computer. Many parts are old, and shared with other species. Language, a recent arrival, would have to work with these old parts. Fortunately, these interface conditions are directly accessible for empirical study. They may be taken as the design specifications or restrictions on the brand new combinatorial linguistic system, from which one may infer some properties of language that might have inevitably followed. This presents a novel and possibly fruitful approach to the study of language evolution: by studying the present, we might learn something about the past. Doing so may help us to understand why language is exactly the way it is, rather than what it might have been.

Acknowledgments

My thanks to Noam Chomsky, John Frampton, Sam Gutmann, Morris Halle, Julie Legate, and Morgan Sonderegger, Bill Wang for comments and discussion on this work. The audiences at ACE II (City University of Hong Kong), Johns Hopkins University, The Haskins Laboratory, University of Arizona, and Northwestern University have also been helpful.

[14] This view of learning in no way marginalizes the innate and domain-specific knowledge of language, the Universal Grammar; in the present context, it is the human phonological system with features and rules that the learning mechanism has to work with.

References

Anderson, S. R. (1988). Morphology as a Parsing Problem. *Linguistics* 26: 521–544.

Berko, J. (1958). The Child's Learning of English Morphology. *Word*, 14:150–177.

Berwick, R. (1985). *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Bybee, J., & Slobin, D. (1983). Rules and Schemas in the Developmental and Use of the English Past Tense. *Language*, 58:265–289.

Campbell, L. (1998). *Historical Linguistics*. Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, N, & Halle, M. (1968). *The Sound Patterns of English*. Cambridge, MA: MIT Press.

Clark, E. (1993). *The lexicon in Acquisition*. Cambridge: Cambridge University Press.

Dabrowska, E. (2001). Learning a Morphological System without a Default: the Polish Genitive. *Journal of Child Language*. 28: 545–574.

Grabowski, E., & Mindt, D. (1995). A Corpus-based Learning List of Irregular Verbs in English. *International Computer Archive of Modern and Medieval English Journal* 19: 5–22.

Guasti, M. T. (2002). *Language Acquisition: The Growth of Grammar*. Cambridge, MA: MIT Press.

Halle, M. (1962). Phonology in Generative Grammar. *Word* 18: 54–72.

——. (1998). The Stress of English Words 1968-1998. *Linguistic Inquiry*, 29:539–568.

Halle, M., & Mohanan, K.-P. (1985). Segmental Phonology of Modern English. *Linguistic Inquiry*, 16:57–116.

Hauser, M., Chomsky, N., & Fitch, T. (2002). The Faculty of Language: What is it, Who has it, and How did it Evolve. *Science*, 298: 1569–1579.

Hu, Q. (1993). The Acquisition of Chinese Classifers by Young Mandarin-speaking Children. Ph.D. Dissertation, Boston University.

Jakobson, R., Fant, G., & Halle, M. (1951). *Preliminaries to Speech Analysis: The Distinctive Features and their Correlates*. Cambridge, MA: MIT Press.

Kiparsky, P. (1982). From Cyclic Phonology to Lexical Phonology. In van der Hulst, H., & Smith, N. (eds.) *The Structure of Phonological Representations. I*. 131–175.

Kiparsky, P. (1988). Phonological Change. In Newmeyer, F. (ed.) *The Cambridge Survey of Linguistics. I*. Cambridge: Cambridge University Press. 363–415.

Koskenniemi, K. (1983). Two-Level Morphology: A General Computational Model for Word-form Recognition and Production. Publication No. 11. University of Helsinki: Department of General Linguistics.

Lightfoot, D. (1999). *The Development of Language: Acquisition, Change, and Evolution*. Oxford: Blackwell.

Ling, C. & Marinov, M. (1993). Answering the Connectionist Challenge: a Symbolic Model of Learning the Past Tense of English Verbs. *Cognition*, 49:235–290.

MacWhinney, B., & Leinbach, J. (1991). Implementation are not Conceptualizations: Revising the Verb Learning Model. *Cognition*, 29:121–157.

Maratsos, M. (2000). More Overregularizations after All: New Data and Discussion on Marcus, Pinker, Ullman, Hollander, Rosen, & Xu. *Journal of Child Language*, 27:183–212.

Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German Inflection: the Exception that Proves the Rule. *Cognitive Psychology*, 29:189–256.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, J., & Xu, F. (1992). *Overregularization in Language Acquisition. Monographs of the Society for Research in Child Development*, No. 57.

Molnar, R. (2001). "Generalize and Sift" as a Model of Inflection Acquisition. Master's thesis. Massachusetts Institute of Technology.

Myers, S. (1987). Vowel Shortening in English. *Natural Language and Linguistic Theory*, 5:485–518.

Myers, J., & Tsay, J. (2000). The Acquisition of the Default Classifier in Taiwanese. In *Proceedings of the 7th International Symposium on Chinese Languages and Linguistics*. Chia-Yi: National Chung Cheng University. 87–106.

Ristad, E. (1994). Complexity of Morpheme Acquisition. In Ristad, E. (ed.) *Language Computation*. Philadelphia: American Mathematical Society. 185-198.

Phillips, C. (1995). Syntax At Age 2: Cross-Linguistic Differences. In *MIT Working Papers In Linguistics* 26. Cambridge, MA: MITWPL, 325–382.

Pinker, S. (1995). Why the Child Holded the Baby Rabbit: a Case Study in Language Acquisition. In L. Gleitman & M. Liberman (eds.) *An Invitation to Cognitive Science: Language*. Cambridge, MA: MIT Press, 107–133.

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

Pouplier, M., & Yang, C. D. (2003). Regulars within Irregulars: The Finer Structure of German Plurals. Manuscript in progress, Yale University.

Rumelhart, D., & Mcclelland, J. (1986). On Learning the Past Tenses of English Verbs: Implicit Rules or Parallel Distributed Processing? In J. McCelland, D. Rumelhart, & the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure Of Cognition*. Cambridge, MA: MIT Press, 216–271.

Sussman, G., & Yip, K. (1996). A Computational Model for the Acquisition and Use of Phonological Knowledge. MIT Artificial Intelligence Laboratory, Memo 1575.

——. (1997). Sparse Representations for Fast, One-Shot Learning. Paper presented at the National Conference on Artificial Intelligence. Orlando, Florida.

Wang, W. S.-Y. (1969). Competing Changes as a Cause of Residue. *Language*, 45: 9–25.

Xu, F., & Pinker, S. (1995). Weird Past Tense Forms. *Journal of Child Language*, 22:531–556.

Yanamoto, K., & Keil, F. (2000). The Acquisition of Japanese Numeral Classifiers: Linkage between Grammatical Forms and Conceptual Categories. *Journal of East Asian Linguistics*, 9: 379–409.

Yang, C. D. (1998). Toward a Variational Theory of Language Acquisition. Manuscript, Massachusetts Institute of Technology.

——. (2002a). *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

——. (2002b). A Principle of Word Storage. Manuscript, Yale University.