A Principle of Word Storage
Draft of July 2002

Charles D. Yang
charles.yang@yale.edu

1 Rules and Exceptions

In generative grammar, the phonology of words is described by rules:

(1) X $\longrightarrow$ Y in the context Z.

In American English, for example, unstressed vowels are reduced to schwa, hence,

(2) X = Vowel, Y = /ə/, and Z = Vowel unstressed

And the familiar flapping rule can be stated as

(3) X = /d, t/, Y = /ɾ/, Z = intervocalic

There are good reasons to believe that words are grouped into classes, each of which is defined by a rule that describes the sound change process shared by these words. Recent work shows that even the English irregular verbs, a component of the lexicon ridden with unpredictability, are also organized in groups that can, and perhaps must, be defined by phonological rules (Yang 1999, 2002). For example, quantitative evidence from language acquisition shows that verbs such as *fly*, *grow*, and *draw*, which have virtually no phonological similarity among them, nevertheless fall under a shared rule that generates their past tense, namely, "Vowel $\longrightarrow$ /u/".

That's not the end of the story, however. A rule, as stated in (1), has two logical components, which I shall call *CAUSE* and *EFFECT*:

(4) a. *CAUSE*: the context Z
b. *EFFECT*: X $\longrightarrow$ Y

CAUSE and EFFECT may form an "if Z then X $\longrightarrow$ Y" relation. CAUSE Z specifies some properties of words, the satisfaction of which will trigger the realization of the EFFECT, $X \longrightarrow Y$. To put differently, EFFECT denotes the shared phonological process among a group of words, whereas CAUSE denotes what properties these words must have to form a group in the first place.

It is clear that CAUSE is, at least sometimes, an operative force. For example, English nouns ending in sibilants form plurals by adding /iz/, specifically,

(5) a. CAUSE: /... [+sibilant]/
b. EFFECT: +/iz/

Presumably, learners of English derive the CAUSE from exposures to input like "chuch-churches", "dish-dishes", etc. As the result of learning, novel nouns with the prescribed final consonants will follow the same rule, e.g., "beamish-beamishes". Yet it is also clear that the force of CAUSE must be curtailed. A rule, (specifically, its CAUSE), very often has exceptions, and exceptions create

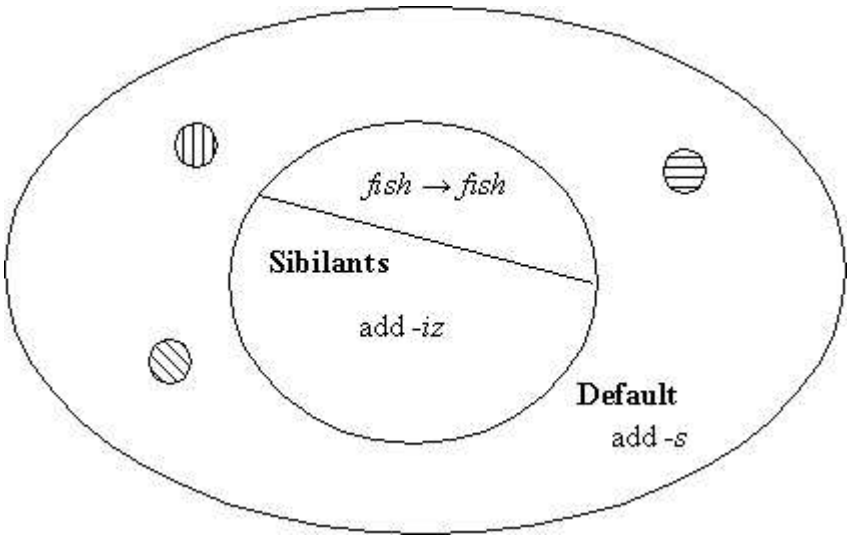

Figure 1: Nested rules whose application follows the Elsewhere Principle.

problems for applicabilities of CAUSE and ultimately, productivities of rules. For instance, suppose the learner hears "grow-grew" and "blow-blew", he may be tempted to form a rule as below:

(6) a. CAUSE: verbs ending in /ow/
b. EFFECT: /ow/ ⟶ /u/

If (6)–clearly no part of the mature English grammar– were ever postulated at some stage of learning, it must be rejected later on, presumably on the basis of *exceptions*, namely, words that contradict this this rule, such as "slow-slowed", "row-rowed", and "tow-towed". Hence, the learner, as speakers of English do, must instead memorize individually the particular instances that fall under the EFFECT: "blow", "grow", "know", ... That is, the CAUSE part of the rule is in effect vacuous, and words are associated with EFFECT by fiat. For this reason, when novel verbs with /ow/ ending are presented, e.g., "smellow", speakers of English do not follow the /ow/ → /u/ pattern.

But the mere presence of exceptions must not categorically prevents the formation of rules, or more precisely, rules with operative CAUSEs. Even the plural rule (5) described earlier has an exception–namely, "fish"–yet there is no doubt the sibilant plural rule is present in the mental lexicon of English speakers. The organization of plural rules is schematically illustrated in Figure 1.

Rules are "nested" according to their specificities, words that are exceptions are marked by shaded dots. The application of rules follows the Elsewhere Principle, and thus, from inside out. For example, "church", which has a sibilant final consonant, will go through the sibilant rule. Within that rule, however, it will have to be matched against "fish", the dot – again required by the Elsewhere Condition – to make ensure it actually isn't "fish". Since "church" is not fish, it will be picked by the rule that add /iz/ to pluralize. If the input word is "car", then the sibilant box will not even be

activated, for "car" fails to meet the CAUSE of the sibilant rule. Rather, it will go straight to the most specific box that matches its phonology properties. Here, that rule is the default, which says "add -s" regardless of the word's properties: the reason that sibilant-ending words never add -s is due to the specificity of the sibilant box over the completely general "add -s" box. However, within that default box, there are also words explicitly marked as exceptions, e.g., "foot-feet": only failing to match those will cause the suffixation of "-s", as in the case of "cars". And finally, "fish" will go to the sibilant box, and will match the exception, and hence will take "fish" as plural.

Hence, the storage of words hinges on a balance between rules and exceptions. At first glance, it seems that if exceptions are relatively minor, then a rule is postulated such that every word meeting the specifications of CAUSE, unless expressly marked as exceptions, will automatically triggers EFFECT. Plural formation, the add -d rule for past tense, etc. are examples of this kind. On the other hand, it seems that if exceptions are abundant, the learner resorts to listing *everything* as exceptions, ignoring any degree of generality that may hold among a small number of words. The "know-knew" past tense class, the umlaut plural class (e.g., /u/ → /i/, as in "foot-feet" and "tooth-teeth") seem to fall in this category. Here, there is no evidence that speakers ever generalize the /ow/ → /u/ pattern, or the /u/ → /i/ pattern, to novel words.

These observations in turn call for a principled solution to the rule vs. exception problem, and more specifically, a principled criterion that alerts the learner when the postulation of a rule is warranteed, or when a list of exceptions, rather, are the prevailing pattern of the data to be captured.

What's the point of all this? For one thing, the computation of rules is clearly an important part of phonological studies in general. There is currently a large amount of behavioral data, ranging from child acquisition, to adult processing, and to neurological studies, that concerns the nature of the mental lexicon, and they call for a theoretically informed examination. This is particularly important when opinions exist that deny the reality of phonological rules. An empirically confirmed theory on the construction of the phonological rules, which includes the rule vs. exception problem as an integral part, will pose serious challenges to such views.

The other important reason for studying the rules vs. exceptions problem concerns the oft-discussed notion of *analogy* in phonology. On the one hand, numerous theories appeal to analogy, also known as *family resemblance*, or *prototype effects*, to explain the organization of the lexicon, specifically the exception or irregular words. Yet concrete formulations of what analogy is are generally lacking, and when one does take the vague suggestions in the literature seriously (e.g., Pinker & Prince 1994), they fail to account for the very data they are designed to explain (Yang 2002). On the other hand, the facts attributed to analogy are nevertheless real: young children do sometimes say "bring-brang", on the analogy to the "sing-sang" class of irregular verbs, and historically, words have changed by shifting from one class of phonological rules to another: "strive-strove" became "strive-strived" along the line of adding "-d"(*analogical levelling*), "wear-weared" became "wear-wore" along the line of "bear-bore" (*analogical extension*). These facts, then, demand a principled and precisely formulated answer. In section 4, we show that, with much empirical merit, one ought to dispense with the notion of analogy altogether, and replace what is considered analogy with phonological rules in disguse. This, we shall see, builds on a principled solution to the rule vs. exception problem.

2 Storage and Computation

We propose that the storage of words (with rules and exceptions) be studied with the formalisms of data structures and computational algorithms, integrated with independent linguistic principles. Specifically, we conjecture that the balance of rules and exceptions is maintained by a principle of *economy*, which seeks to minimize the time required to retrieve words from the lexicon. In a nutshell, we will claim that a rule will be postulated by the human learner in face of exceptions when if the expected time to retrieve a word by doing so is less than storing all words as exceptions, i.e., a list of specific cases.

Consider first the is the strategy that all words are simply stored a list of exceptions:[1]

(7) IF $V = c_1$ THEN ...
IF $V = c_2$ THEN ...
...
IF $V = c_N$ THEN ...

For an input word V, one need to find a matching clause in (7). Suppose that the set of words in $\mathcal{N}$, and its cardinality of N. To retrieve a particular word x, one needs to do a search through a database of the size N. We assume that the expected time to find a match is a function of N: the longer the list is, the longer it takes to locate the target clause. Write this function as $T(N, N)$.

Imagine what one must do to write rules that describe a set of patterns with exceptions. Again, assume that the set of words is $\mathcal{N}$ A subset, $\mathcal{M}$, with a cardinality of M, consists of the exceptions.

(8) If a word V satisfies CAUSE Z, then EFFECT X $\longrightarrow$ Y
except if $V = c_1, c_2, ...c_M$

In phonology, the effect of exceptions in (8) can be given by the Elsewhere Principle, which requires the application of phonological rules in an order of from specific to general. Hence, (7) will be stated as:

(9)

If $V = c_1$ then ... (1)

If $V = c_2$ then ... (2)

...

If $V = c_M$ then ... (M)

If Z then X $\longrightarrow$ Y (*)

The Elsewhere Principle commands that the more specific clauses be evaluated first to match the identity of the input word V. If the special cases of c_1, c_2, ..., and c_M all fail, the general rule (*) will be used. (9) clearly duplicates the effect of (7).

[1]The ... in these clauses represent phonological rules that an input word is associated with.

Consider what happens when one wants to retrieve a word V under organizations like (9). If V is one of the exceptions, then one will search through the M clauses: the complexity is similar to that in (7), only with a small vocabulary of M items. If, however, V falls under the description of (*), then it will be processed until all the exceptional clauses (1—M) are checked, and fail. That is, for every word that falls under the rule, there is an accumulative cost of ruling out *all* exceptions. Write the expected time to find a match in a list of size N with M exceptions as $T(N, M)$. We have

(10) $T(N, M) = \text{Pr}(V \in \mathcal{M}) \times T(M, M) + \text{Pr}(V \notin \mathcal{M}) \times \sum_{i=1}^{M} T(N, M)$

In comparison to the complexity of listing all of $\mathcal{N}$ as exceptions, we propose:

(11) **Conjecture**

a. If $T(N, N) < T(N, M)$ then $\mathcal{N}$ will be stored as a list of N exceptions altogether.

b. Else, $\mathcal{N}$ will be stored as a rule with a list of M exceptions.

That is, we claim that the verbs with /ow/-endings fall under (11a), that is, "know", "grow", etc. are individually associated to the process that replaces their vowels with /u/, and no generalization is drawn from the fact that they do share a same final vowel. And we claim that the class that adds /iz/ to sibilant falls under (11b)–a generalization *is* drawn from the shared characteristics of these nouns–with ("fish") individually listed as an exception.

In what follows, we will discuss (11) under a few further (and we believe, reasonable) assumptions about word storage, and examine several empirical cases in the study of rules and exceptions.

3 Formal Issues

Suppose that $\mathcal{M}$, the set of exception words, is a random sample of the words in $\mathcal{N}$. Write $t(i, M)$ as the time required to match the word $v_i \in \mathcal{M}$, in a list consisting of M clauses. Again, recall that $T(N, N)$ denotes the expected time of storing all N words as a list of exceptions, and that $T(N, M)$ denotes the expected time of storing the words in $\mathcal{M}$ as a list of exceptions and the rest of $\mathcal{N}$ as a rule.

The expected time to match words under the two approaches are:

(12) a. All-exceptions: $T(N, N) = \sum_{V_i \in \mathcal{N}} E[t(i, N)]$

b. Rule-exceptions: $T(N, M) = \sum_{V_i \in \mathcal{M}} E[t(i, M)] + \sum_{V_j \notin \mathcal{M}} E[t(j, M)]$

The Elsewhere Condition commands that the exception clauses be examined by the general rule being reached. Hence, when a rule-following word is presented, there is a cumulative cost of searching through *all* the clauses in the list of exceptions to ensure that the word is *not* one of the exceptions. That is, in (12b), for $V_t \in \mathcal{M}$, $E[t(j, M)] = M$. Because of this peculiar way in which rules and exceptions are handled in phonology, which follows from the Elsewhere Principle, postulating a rule may a formidable cost if the number of exceptions is large.

Since there are M exceptions, (12b) becomes:

(13) Rule-exceptions: $T(N, M) = eT(M, M) + (1 - e)M,$ where $e = \frac{M}{N}$

What's the computational complexity of $T(M, M)$, the expected time of retrieving a clause out of M clauses? It is easy to see that it cannot be worse than (a linear function of) of M: the worse one can do is to go through the list exhaustively. But humans can perhaps do better than that, and we can get more accurate estimates on $T(M, M)$. There's good evidence that word retrieval (and by logic, storage) is sensitive to word frequencies.

To a first approximation, for exception words such as the English irregular verbs, frequent ones are generally associated with the appropriate phonological processes faster and more reliably than rare ones.[2]. There is a natural computational implementation for this effect. One can place the most frequent word at the first clause, the 2nd most frequent at the second clause, and so on. So the most frequent word costs one clause to match, the 2nd most frequent costs two, and son. It is easy to see that using this strategy, the averaged retrieval time is minimized.

Yet this optimized storage algorithm may not be psychologically plausible, for it requires the learner to keep track of, and then rank, the frequencies of all these words. So some plausible alternative ought to be considered.[3] However there are many online, or "self-organizing" (Knuth 1998: 401) algorithms for list searching that are perfectly plausible and whose efficiency comes very close to to that of the optimized one; see Bachrach & El-Yaniv (1997) for a survey of these so-called list accessing algorithms. One simple instantiation is the MOVE-FRONT algorithm: whenever a clause is successfully located, it is moved to the beginning of the list.

Assume that the words to be stored follow a Zipfian distribution. That is, the rank r_i of a word V_i is inversely proportional to its frequency p_i.

(14)

$$\forall i, V_i \in \mathcal{M}, r_i \times p_i = C, \text{ where } C \text{ is some constant}$$

The frequencies of the words then are:

(15)

$$\forall i, V_i \in \mathcal{M}, f_i = \frac{1}{iH_M}, \text{ where } H_M = \sum_{j=1}^{M} \frac{1}{j}$$

The expected retrieval time under the optimal algorithm is:

(16)

$$T_{\mathrm{O}}(M, M) = \sum_{i=1}^{M} f_i i = M/H_M$$

[2]This does not mean that the past tense of an irregular verb is learned faster if it is more frequent. We have only been talking of the CAUSE component of the rules, i.e., the word-rule association, which does seem to correlate strongly with frequency. The actual derivation of the past tense must also go through the application of the EFFECT component of a rule, which cannot be isolated from the organization of the overall phonology in the language. See Yang (2002) for details.

[3]Unless, of course, if one is inclined to declare a constraint, " MIN RETRIEVAL TIME" , rank it very high, and call it a day.

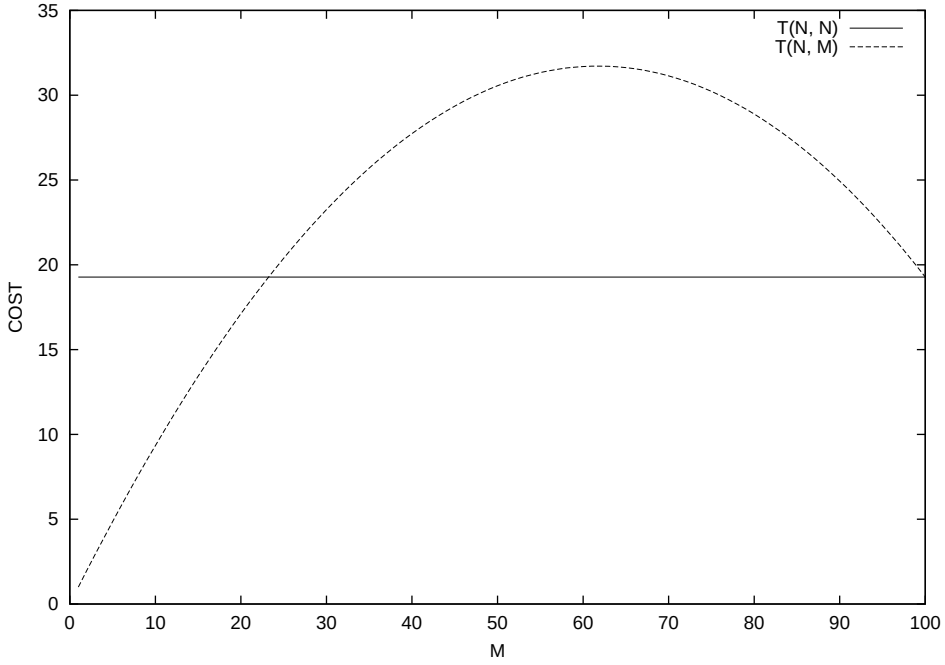

Figure 2: The cost of storing M words as exceptions and the rest as a rule, compared with storing all N words as exceptions.

The MOVE-FRONT algorithm can be shown, though we will not dwell on the mathematics here (see Knuth ibid and the references cited therein) that under the Zipfian assumption of word frequencies, that expected search time is approximately the MOVE-FRONT algorithm is only worse than the optimal algorithm by a factor of less than 2. An even more suitable algorithm–call it ONE-UP–moves a successfully located clause one position up in the list, swapping with the preceding clause. The efficiency of ONE-UP is even closer to the optimal one (Rivest 1976); since it requires even less computational effort on the part of the human processor and is hence more plausible. We conjecture that the arrangement of the list of clauses is close to optimal, and the expected retrieval time is $T(M, M)$ is approximately M/H_M, where H_M is the Mth Harmonic number.

These assumptions enable us to compare $T(N, N)$ and $T(N, M)$. Figure 2 plots the case for $N = 100$. When M, the number of exceptions, is less than about 25, it is more economical to employ a rule in addition to these exceptions. And when M is greater than 25, it becomes more economical to store every words as exceptions.

For the sake of completeness, there is a technical problem that requires some comments. We have been assuming a Zipfian frequency distribution of the words in $\mathcal{N}$, which both seems to accord with reality (for whatever, perhaps uninteresting, reason), and makes it possible to obtain complexity results for the class of storage algorithms discussed earlier. We have also been assuming that the words in the exception subset $\mathcal{M}$ also follow a Zipfian distribution. This is false, if, for example,

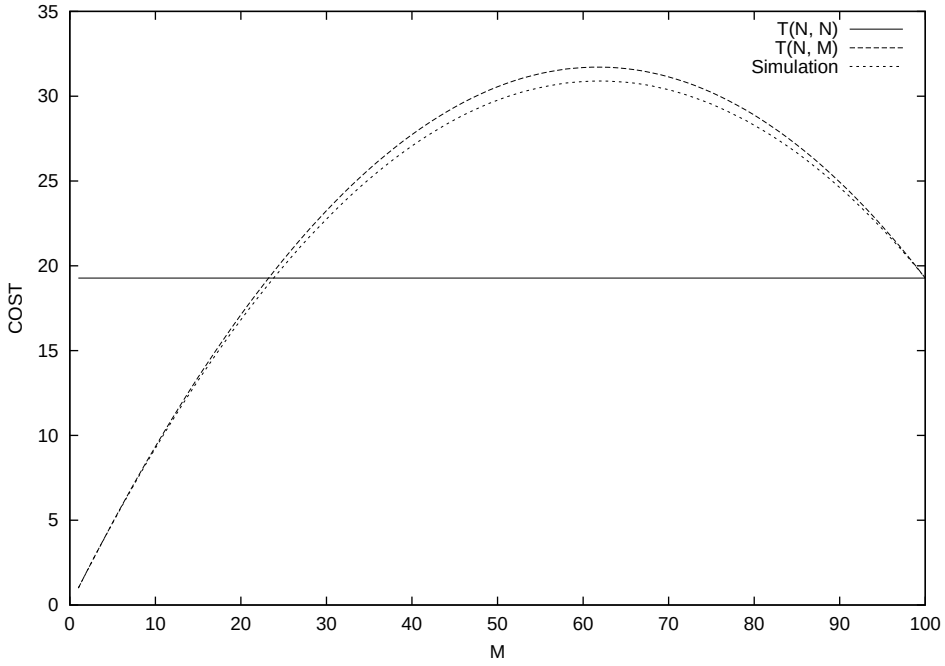

Figure 3: The expected retrieval time from simulation ("Simulation"), compared to the theoretical results ("T(N, M)"). The results are averaged our 10,000 trials for each value of M, $1 \leq M \leq 100$.

the words in $\mathcal{N}$ have equal probabilities of being listed as an exception. However, as is well known, exceptions tend to be high frequency items. We may then assume that the selection of $\mathcal{M}$ out of $\mathcal{N}$ is also a frequency dependent process. Specifically, word v_i being drawn as the first word is f_i, so clearly more frequent words are most likely to be an exception. After each draw, the selected word is placed into $\mathcal{M}$ and removed from $\mathcal{N}$: this results in the renormalization of the probabilities for the remaining words in the next round of drawing. The analytical results will be dealt with in the next incarnation of this paper, but for now, we can simulate the drawing of $\mathcal{M}$, and explicitly compute the expected retrieval time of $\mathcal{M}$ under the MOVE-FRONT or ONE-UP algorithm introduced earlier. Averaging these results over many trials, we see in Figure 3 that our theoretical results match the simulation nearly perfectly.

Finally, let's examine the critical values of M_c where $T(N, N)$ and $T(N, M)$ meet. If $M < M_c$, then it's better off to postulate a rule for the words in $\mathcal{N} - \mathcal{M}$, and store $\mathcal{M}$ as a list of exceptions. If $M > M_c$, then it's more economical to store the entire $\mathcal{N}$ as a lists, ignoring any (partial) regularities among the words. Figures 2 and 3 show that M_c is in fact quite small relative to N, indicating the significant cost of searching an entire list of exceptions–which follows the Elsewhere Condition–in computation of word storage and retrieval. In section 4, we will look at some empirical consequences for this conjecture.

Solving the equation $T(N, N) = T(N, M)$ will yield the value of M_c for a given N. We have:

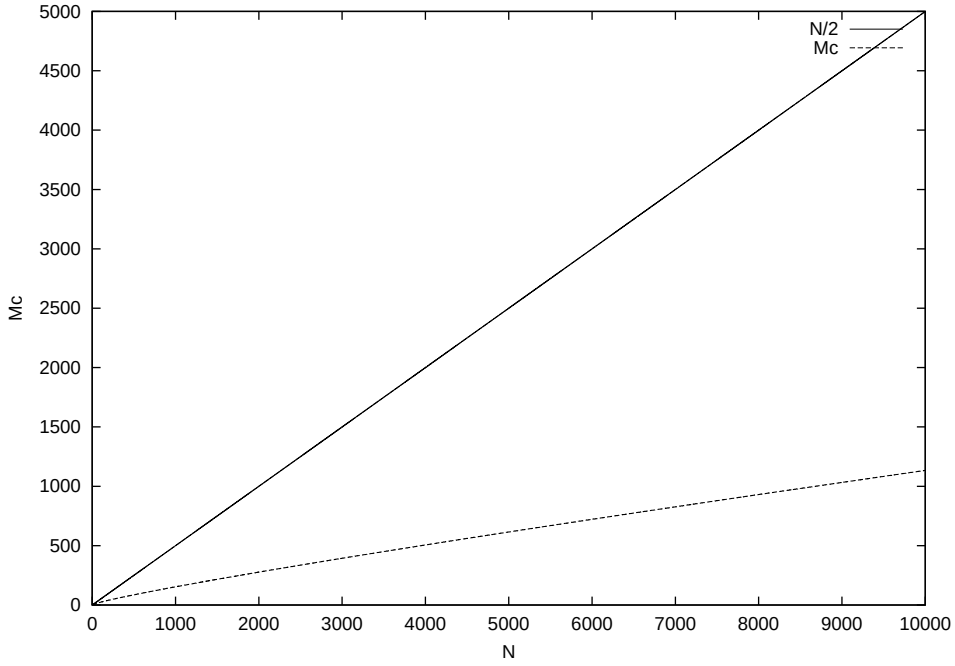

Figure 4: The critical value M_c as a function of N.

(17)

$$\frac{N}{H_N} = e\frac{M_c}{H_{M_c}} + (1-e)M, \text{ where } e = \frac{M}{N}$$

(17) is difficult to solve even if we approximate H_N with $\ln N$. Figure 4 gives the solutions of M_c for $N = 1, 2, ...10000$ found by a computer. Compared to the growth of M_c with the $N/2$, an alternative strategy that says a rule will be postulated if there are more rule-following words than exceptions. When N is small, M_c is comparable to $N/2$. When N is large, $M_c(N)$ becomes a *near* linear sublinear growth with a slope of about 0.1, considerably smaller than 1/2.

This leads some predictions. If our conjecture of word storage economy is correct, it predicts an upper bound of how many exceptions rules can tolerate, before everything is better off stored as a list of exceptions. It is reassuring to know that English irregular verbs have not reached that limit yet. According to corpus studies (Grabowski & Mindt 1995), English has about 4,500 verbs, of which 160 are irregular, considerably below than the theoretical bound of approximately 560: the -d rule would disappear when the number of irregular verbs exceed that bound.

4 Analogy as Rules

Section 3 pursues the consequences of our conjecture and the discussion is highly abstract, and admittedly speculative. Along the way we have made some assumptions of how words and rules are

stored and used in the mental lexicon, based on a limited range of behaviorial data. In this section, we will attempt to justify our theory through a variety of case studies.

Because the problem concerns the postulation of rules in the face of exceptions, an experimental approach is, at least in principle, tenable. The classic test for the presence of rules is the Wug test: if nonce words (with sufficiently diverse phonological properties) are derived by the subject following a particular pattern, then there must be a rule/process responsible for the computation of such regularity. So established are the the plural suffixation rule "add -s", and the past tense rule "add -d", through Berko's classic experiments (1958). Hence, it is possible to explore the balance between rule and exceptions by studying the ellicitted response from (adult or child) subjects when novel words are presented, after teaching them the phonology of a list of words with varying number of exceptions. This, however, must be left for future work.

There is suggestive evidence from, again, the acquisition of past tense that supports our theory. As noted in the previous section, the presence of the -d rule is predicted by our theory, for the number of irregular verbs is far below the tolerable critical value for the rule, though we will have a little more to say about its acquisition and development momentarily. Right now we focus on the regularities among irregular verbs.

The developmental evidence for the -d rule comes from *over-regularization*, that children sometimes (about 10% of the time) inflect irregular verbs using the the -d rule, such as "hold-holded". Similiarly, the developmental evidence for rules among the irregulars would be over-*irregularization*, that a verb is inflected following an incorrect but *irregular* pattern. These errors in fact are quite rare, accounting only 0.2% of all past tense uses (39 out of 20,000, Xu & Pinker 1995), but they are nevertheless revealing.

Significantly, of the 39 over-irregularized verbs in children's production in Xu & Pinker (1995), nearly half (16) are of the -ing ending: they are "bring", "swing", and "fling". This is also the *only* systematic class among all the over-irregularized verbs (not many to begin with, to be sure). It is often said that such errors are due to "analogy" (Xu & Pinker 1995). True, "bring" and "swing" sound a lot like "sing" or "ring", but such an approach is problematic, upon further reflection. Analogy is a notion that has rarely been made explicit in the literature, as remarked earlier. Recently, it has been suggested that analogy can be modeled as activating the stored "prototype" that most closely matches a new word (Hare et al. 1995, Eddington, 2000, among others). But "Snow" sounds like "know", but why doesn't anyone say "snew"? "Ride" sounds like "hide", but why not "rid"? "Hatch" sounds like "catch", but why not "haught"? The list goes on and on. The trouble with analogy is, closest match *compels* a word to be associated with a class that bears most phonological similarities with it, yet phonological similarity is not a reliable predictor of what words would go, as these examples illustrate. In contrast, the *ing-ang* change does seem to indicate a (some) productive class.

There is no mystery in these data if one dispenses with analogy altogether. Just like the Wug test and "hold-holded" proving the existence of the add -d rule, the productivity of a particular pattern, including over-application errors by children, prove the existence of a rule that institutes that pattern. Hence, the only source of productivity are rules: whenever there is "analogy" , the culprit must be a rule, and hence analogy has no independent empirical status.

The productive *ing-ang* change, at least during the early stage of language acquisition, is due

to a rule that changes /i/ to /æ/ for words that end in /iŋ/, where no similar rule exists for any other irregular class. Why a rule exists for this particular class but not others is a direct consequence of our principle of economy in word storage.

To fully understand the issue, we need to say a few words about how rules are learned from data; for details, see Yip & Sussman (1996, 1997), Molnar (2001), and Yang (in preparation). In a nutshell, the learner derives generalizations in the phonological properties of words that undergo identical, or similar, sound changes in their phonological derivation. In other words, the learners tries to construct rules that describe the commonalities–the CAUSE–for words that share an EFFECT.

The precise algorithm for doing so need not concern us here. Briefly, the algorithm is a two-step procedure. Words in this model are representation as a sequence of phonemes, and each phoneme is represented as a set of distinctive features. The input word, e.g., "walk", and the output, "walked", which is the product of the rule being learned, are lined up together. First, the difference between their featural representations can easily be identified, and that is taken to be the EFFECT of the phonological rule. second, words with identical EFFECTS are then grouped together to identify the CAUSE: the feature values they share are retained in the generalization, CAUSE, and the conflicting feature values among them are considered as "don't care" cases in the CAUSE. The learning model is completely general, applicable to all cases of phonological rules, and in fact, can be s extended to include morphosyntactic features as well. Its application to the past tense data thoroughly outperforms all previous models of learning, both in efficiency and in accuracy (Yip & Sussman 1996, 1997).

Any rule can in principle be learned. Upon seeing "sing-sang" and "ring-rang", a rule is learned that converts all words with /iŋ/ endings to /æŋ/ in past tense. Upon seeing "know-knew", "blow-blew", and "grow-grew", a rule is learned that converts word ending from /ou/ to /u/. However, that a rule can be learned does not mean it will be incorporated into the lexicon. The learner will soon run into exceptions to rules. Rules are retained only if the cost of maintaining the rules with a few exceptions is lower than storing everything as exceptions, as suggested by our conjecture. For this reason, as we show with some corpus statistics that the /iŋ-æŋ/ rule can (at least temporarily) retained, and the /ow-u/ rule must be rejected.

We take the adult sentences from the Brown and Suppes databases in the CHILDES corpora as a sample of the data the model learns from. There are over 110,000 sentences. All the /iŋ/ ending and /ow/ endings verbs are listed below, along with their frequencies in past tense:

(18) a. -ing ending: *brought* (80), rang (1), sang (6), *stung* (4)

b. -ow ending: knew (49), *showed* (11), threw (28), blew (5), *followed* (6), *slowed* (1), *snowed* 5, *towed* (1).

Were the two rules postulated, there are certain words that would count as be exceptions; they are highlighted in italics. For (18a), there are 2 exceptions out of 4 words, and for (18b), there are 6 exceptions out of 8 words. Solving the equation in (17) we see that for 4 words, the critical value $M_c(4) = 3$, but for 8 words, $M_c(8)$ is also 3. Hence, the /iŋ→ æŋ/ will be used in the /iŋ/ ending words, but /ow → u/ will not be used in the /ow/ ending words. Rule (18a) will be postulated by the learner and rule (18b) will not. And thus, over-irregularization errors for (18a), but no such thing

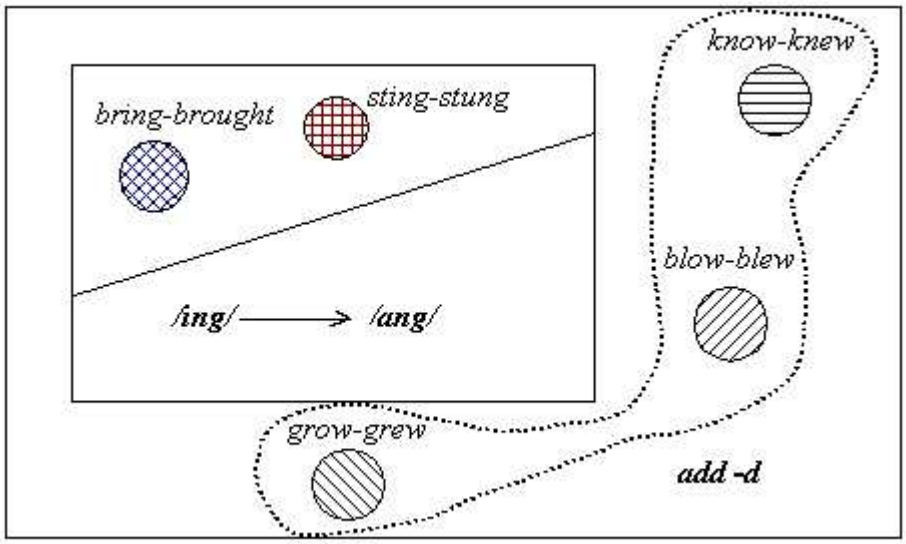

Figure 5: The organization of words with /ing/ endings and /ow/ endings. The dotted line indicates what would have been a rule, if there were so many exceptions.

for (18b). Figure 5 shows the structure of the lexicons with the relevant rules.

Quite possibly, even the /iŋ-æŋ/ rule will disappear as more words (with /iŋ/ endings) are encountered by the learner. Here complications arise because for the class of verbs with /iŋ/ endings, in addition to the /iŋ-æŋ/ change, also contains the /iŋ-ʌŋ/ change, such as "cling-clung", "sting-stung", and "swing-swung". This is not a problem: if the /iŋ-æŋ/ change is attempted as a rule, then all the /iŋ-ʌŋ/ verbs would be listed as exception, and vice versa. Furthermore, "wing", whose past tense is the regular "winged", will count as an exception for either. In both cases, though, there are more exceptions than the economy principle can handle, so all /iŋ/ endings words are listed as exceptions, and "wing" is picked up by the default rule. The reason for an active CAUSE in the /ing-ang/ rule is due to a sampling effect: the /iŋ-ʌŋ/ words are quite rare, and are unlikely to be attested robustly in early language acquisition.

Note that this conclusion is confirmed by behavioral data from ellicitation studies. Prasada & Pinker (1993), for example, found that when presented with novel verbs, adult subjects consistently prefer the default inflection, if when the novel verb is minimally different from existing irregulars, such as "spling", "skring", "sprink", "cleed", etc.

Our theory also provides a direct explanation of the well-known U-shaped development in the acquisition of English past tense. The phenomenon refers to the observation that children's very early past tense consists of almost exclusively irregulars. When the default form starts being used, there was a dip in the irregular performance, as overregularization becomes a possibility. Only late, after more exposure to irregulars do they revert to adult level performance. This pattern is

entirely predicted. The early verbs that children learn are bound to be mostly irregular: irregular verbs, let's recall, make up 29 of the top 30 most frequent verbs, and constitute nearly 60% of the probability mass (Grabowski & Mindt 1995). Therefore, the small verb vocabulary of young children must contain a good deal of irregular verbs, and it is easily the case that having a (default) rule with many exceptions is prohibitive. As the voabulary expands, more regular verbs will be added than the irregular ones: according to our theoretical calculation, a 500 word vocabulary can roughly accommodate 80 irregular verbs. Sometime during the expansion of the child's vocabulary, the addition of the default -d rule would be justified now that it is more economical to have a rule with exceptions than all exceptions. After the default rule is postulated,[4] the learner will then have to combat the force of overregularization upon further exposure to irregular verbs (Pinker & Prince 1994, Yang 2002).

More, and more robust, evidence for our theory may come from languages with less degree of irregularity than English irregular verbs. German pluralization is such a case. Despite its numerical rarity (only 7% of all nouns), the rule, "add -s", is the default, for it passes the equivalent of the Wug test (Marcus et al. 1995).[5] In the "irregular part", plurals largely follow some general rules that are conditioned upon both the morphophonology of the noun. It is found (Köpcke 1998) that German plurals fall into eight, distinct classes of suffixation and/or vowel changes, including the -s class, and in fact 85% of all plural forms are predictable from the stem. The rest, of course, has to be individually memorized as exceptions.

If our theory is correct, then rules must exists for these "irregular" classes of German nouns. And if such rules exist, then German speakers, particularly children, will over apply them. This then forms a strong contrast with the very rare over-irregularizations in English, i.e. over-applications of irregular rules in English, which we have argued to be non-existent, except for the /iŋ/ → /æŋ/ class as discussed earlier, and only for a short period of time.

The prediction of high rate of over-irregularization is confirmed by the large longitudinal study by Behrens et al. (2001); see references cited therein for similar work. They show that at about 2:2, about 10-20% of nouns are pluralized incorrectly, and errors involving all eight rules are attested. This is in sharp contrast with the ratio of similar errors in English, which is about 0.2%.

A simple (and informal) experiment can demonstrate the presence of "irregular" plural rules in adult German.[6] We present adult German speakers with madeup nouns that resemble native German words; cf., footnote 5. The subjects are instructed to fill in the blanks in sentences that force a plural form, as in (19), and the results are summarized in Table 1.

(19) Ich brauche Deine blöden $\underline{1cm}$ nicht.
I need your stupid $\underline{1cm}$ nicht.

[4]The actual induction of the default rule follows the same algorithm for two-step algorithm discussed earlier. Words that undergo -d change in past tense are grouped together, and the algorithm attempts to find commonality among them in terms of their feature specifications. Because regular verbs are of arbitrary sound shapes, and thus consists of conflicting distinctive features in their representations, very quickly the algorithm derives that words going into the -d rule can be a sequence of arbitrary feature matrixes, that is,, anything goes. And that is precisely the characterization of the default rule. In computer simulation, the -d rule can be derived with 100-200 verbs (Yip & Sussman).

[5]But it is important to bear in mind that most of German nouns that fall under the -s rule are foreign words, and the tests Marcus et al. (1995) administrate are based on words *very* different from German.

[6]I am *entirely* grateful to Marianne Pouplier for her help.

Noun	Subject 1	Subject 2
Blug	Blugen	Blüge
Pfasse	Pfassen	Pfassen
Brangsorde	Brangsorden	Brangsorde
Büxte	Büxten	Büxte
Leup	Leupen	Leupe
Brahmt	Bramhten	Brahmten
Klieb	Klieben	Kliebe
Rongst	Rongsten	Röngste

Table 1: Wug test for native-like German nouns

"I don't need your stupid $\underline{1cm}$.

We find that adults are far more likely to employ the "irregular" rules to pluralize. Subject one seems to prefer the -en suffixation, and subject two, the -e suffix followed by umlaut, but both subjects deem the other's response acceptable. This experiment, together with the acquisition study of Behren et al., suggest that the "irregulars" in German plurals are in fact "regular", in the sense that they are described by productive rules conditioned upon the phonological properties of words, along with exceptions. Hence, they are qualitatively identical to the regular -s rule: the only difference is, the default rule has the least restrictive CAUSE on what words it can apply to. This constitutes further evidence that the dual-route mechanism of Prince & Pinker (1994) and their associates, that irregulars are individually memorized by association and only the regulars are handled by a rule. According to this view, productivity along the "irregular" forms should not be possible.[7] Future work will be dedicated to a detailed corpus study of German nouns and plurals to see for each productive rule, how many nouns must be listed as exceptions, against the rest, rule-following words. The consequence of the theorem in (11) can then be fully studied.

Based on Yip & Sussman's work, we know *how* rules may be learned. With this rule vs. exception problem out of way, we know *when* rules are learned. So we are ready for the question: *what* happens when rules are learned, as a developmental problem, and also as a historical problem, namely, how rules change as they are manifested in the learning data passed on from one generation to another. We turn to these questions in Yang (in preparation).

References

Bachrach, R. & El-Yaniv, R. (1997). Online list accessing algorithms and their applications: Recent empirical evidence. In *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 53-62, New Orleans, Louisiana, 5-7 January 1997.

[7] In the past few years, evidence against the Pinker & Prince model has been amounting, and it has mostly come from the study of German plurals. See Penke & Krause (2002) for a recent example for aphasic studies and adult ellicitation.

Behrens, H. & Kiekhoefer, K. (2000). Identification of inflectional paradigms: the acquisition of German plurals. Paper presented at the 9th International Morphology Meeting. Vienna.

Eddington, D. (2000). Analogy and the dual-route model of morphology. *Lingua* 110: 281-298.

Grabowski, E. & Mindt, D. (1995). A corpus-based learning of irregular vebs in English. *Computers in English Linguistics*, 19: 5-22.

Hare, M., Elman, J. & Daugherty, K.G. (1995). Default Categorization in Connectionist Networks. *Language and Cognitive Processes* 10: 601-630.

Knuth, D. (1998). *The Art of Computer Programming. Vol III: Sorting and Search*. Second Edition. Addison-Wiley,

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

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

Penke, M. & Krause, M. (2002). German noun plurals: A challenge to the Dual-Mechanism model. *Brain and Language* 81: 303-311.

Pinker, S. & Prince, A. (1994). Regular and Irregular Morphology and the Psychological Status of Rules of Grammar. In S. Lima, R. Corrigan & G. Iverson (eds.) *The Reality of Linguistic Rules*. Amsterdam: John Benjamins, 321-351.

Prasada, S., & Pinker, S. (1993). Generalization of Regular and Irregular Morphology. *Language and Cognitive Processes*, 8:1-56.

Rivest, R. (1976). On self-organizing sequential search heuristics. *Communications of the ACM*. 2: 63-67.

Xu, F. & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language* 22, 531-556.

Yang, C. (1999). Words, Rules and Competitions. Ms., MIT.

Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

Yang, C. (in preparation). The Computation of Phonological Rules: Learning, Change, and Origins.

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

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