On productivity*

Charles Yang
Department of Linguistics Yale University

Language learning is a remarkably robust process. The child is incredibly good at recognizing systematic regularities even when faced with lexically and contextually restricted exceptions This paper sketches out a preliminary model that recognizes productive processes and exceptions as such; accordingly, the learner can proceed to internalize each as different kinds of linguistic knowledge. We argue that if a linguistic process is conjectured to be productive, then having exceptions to it can add (surprisingly) significant cost to its online processing. Empirically, we explore these issues in the domain of morphology, which leads to finer-grained analyses of a number of well-known morphological problems. We also briefly discuss how the methodology and results of this work may generalize to syntactic learning.

Keywords: productivity, exceptions, morphology, psycholinguistics, language acquisition, computational linguistics, corpus linguistics, English past tense, German noun pluralization

> It is quite obvious that many of the phonological rules of the language will have certain exceptions which, from the point of view of the synchronic description, will be quite arbitrary. This is no more surprising than the fact that there exist strong verbs or irregular plurals. Phonology, being essentially a finite system, can tolerate some lack of regularity (exceptions can be memorized); being a highly intricate system, resulting (very strikingly, in a language like English) from diverse and interwoven historical processes, it is to be expected that a margin of irregularity will presist in almost every aspect of the phonological description. Clearly, we must design our linguistic theory in such a way that the existence of exceptions does not prevent the systematic formulation of those regularities that remain.
>
> Noam Chomsky & Morris Halle
> *The Sound Pattern of English* (1968: 172)

Linguistic Variation Yearbook 5 (2005), **265–302.**
ISSN 1568–1483 / E-ISSN 1569–9900 © John Benjamins Publishing Company

1. Introduction

As Sapir once famously remarked, all grammars leak (1928: 38). In less colorful terms, all grammars have exceptions that co-exist with general statements about the linguistic system. Grammar leaks, as Sapir continued, do not undermine what he called "the fact of grammar, a universal trait of language, … a generalized expression of the feelings that analogous concepts and relations are most conveniently symbolized in analogous forms." (1928: 37–38). Despite the leaks, there *is* a grammar after all.

Indeed, grammars leak all over the place, and right from the beginning. It is thus even more remarkable that we somehow steer clear of them and still attain the systematic grammar in the end. To wit, one of the earliest utterances an American child hears may be "Baa Baa Black Sleep, have you ___ any wool?", yet she would not go on to talk as if American English inverted the main verb in interrogative questions. A serving of General Tso's Chicken does not entail that /ts/ is a valid onset consonant clusters for English. And every English speaker learns that English is not a pronoun or topic drop language like Italian or Chinese, despite hearing missing subjects in imperatives, recipe constructions, and dairy drops ("seems good to me"). Viewed this way, the human language learner is remarkably sensitive to the *productivity* of linguistics processes. The learner is able to recognize that, in general, English must fill the subject position, does not raise the main in interrogatives, and cannot start a syllable with /ts/. The language faculty, then, must have built in a component that recognizes the productivity of linguistic processes, and responds accordingly: the excpetions don't "count". As the quote from SPE illustrates, exceptions restricted to specific lexical items and contexts can be memorized, and systematic regularities are allowed to generalize over novel instances of language use.

The present paper is an attempt to develop a principled approach to the classic problem of linguistic producitivity. Our strategy is learning theoretic. We believe that the formulation of the problem in SPE is essentially correct. The challenge posed by linguistic processes with various degrees of productivity may be resolved in the "evaluation measure" (cf., Chomsky 1955; Chomsky 1965), i.e., a theory of language learning. Our long term goal is to articulate a learning model that learns both productive processes and unproductive exceptions. Of the immediate concerns in the present paper is the development of a decision procedure–dubbed the *Tolerance Principle* – that recognizes a productive process when it sees one. If a linguistic process is not productive, i.e., it is an exception, the learner can proceed to memorization, which in effect factors it out in the learning of the core grammatical system.

The empirical domain we have chosen to establish the our model is morphology, where a wide range of linguistic productivity is on full display. The best known case is perhaps the past tense system in English. English irregular verbs are no doubt exceptions: they must be learned on the basis of the input and must be committed to memory on an individual basis. Of courses, not all English past tenses are idiosyncractic: the regular class is fully productive, as evidenced by the classic Wug test in which English speakers reliably generalize the "add -d" rule to novel verbs (Berko 1958). The choice of morphology as the empirical testing ground is two fold. First, as will be clear, our proposal draws from the study of real-time morphological processing, and makes extensive use of corpus statistics of word distributions. At the present time, it is not clear how such studies can be extended to syntactic investigations. Second, we believe that the problem of learning when faced with exceptions is sufficiently general: a proposal that works for morphological productivity may shed light on other learning problems such as syntactic acquisition.

This paper is one long argument. In Section 2, we introduce the problem of productivity through the case of English past tense. We also present a computational model for learning morphological rules that serves as the background for our approach to productivity. Section 3 proposes (a component of) a computational model of morphological processing along with evidence for its plausibility. Section 4 discusses the formal properties of the morphological processing model. One of these consequences is the Tolerance Principle, a decision procedure that recognizes the productivity of morphological rules. Section 5 puts the Tolerance Principle to an empirical test with a re-analysis of the well-known problem of German plural morphology. Section 6 concludes with some brief remarks on the problem of linguistic productivity in the context of syntactic learning.

2. Rules and productivity

2.1 English past tense

If anything useful came out of the so-called past tense debate, it is that regular verbs and irregular verbs are different beasts. This is not the place to review the extensive literature on this topic (see Pinker & Ullman 2002 for a summary, and

Yang 2000 for a general assessment of the intellectual merit of the debate). It suffices for our purposes to state that we agree with the Words and Rules (WAR for short) model (Pinker 1999; Clahsen 1999, 2005) that the irregular verbs are exceptions and must be somehow memorized, while regular verbs are inflected by the use of a productive rule "add -d".

This is not to say that we agree with the details of the WAR model. In Yang (2002), I put forward a reanalysis of the past tense learning data (Marcus et al. 1992) and showed that even the irregulars are inflected by the use of rules as has been assumed all along in generative grammar. Following a useful term introduced by Anderson (1974), these irregular rules can be best described as *morpholexical*: unlike "add -d", these rules apply to a fixed list of lexical items and do not generalize to novel tokens.[1] For instance, the rule "-t & Rime→/a/" is morpholexical in the sense that it applies only to a list of words such as *think*, *catch*, *bring*, *buy*, *teach*, and *seek*, which have no conceivable similarities among them.[2] Irregular inflection still requires memorization, but in this approach, the learner does not memorize the inflected forms directly. Rather, it memorizes the specific list of words to which each morpholexical rule applies. Thus, a frequency-dependent associative memory still plays an important role in morphological computation: association is established between between words and rules, not between words and words (stem and past) as in the WAR model. This alternative view, dubbed "Rules over Words" (ROW), allows us to not only capture the well-known frequency effects in past tense learning but offer a finer-grained interpretation of the acquisition data. We will not dwell on that work here; the interested reader should consult Yang (2002) for empirical evidence in favor of such an "all rules" approach.

Whichever morphological learning model one assumes, the learner must solve the productivity problem. In order for the WAR model to work, the learner must first recognize whether a given verb is regular or not – so it can slot the verb into the appropriate module (memory or rule). This, of course, requires the learner to (a) construct the rule "add -d" on the basis of the input and (b) recognize that the rule "add -d" is actually a productive/default rule. In order for the ROW model to work, the learner must (a) construct the rules – "add -d" as well as the morpholexical rules for the irregulars–on the basis of the input and (b) recognize the "add -d" is productive, while irregular rules such as "-t & Rime → /a/" are not (i.e., they are morpholexical). In a moment, we present a rule-learning model that both the WAR and ROW approaches require.

Empirically, we know that children are exceptionally adept at the productivity problem. In past tense acquisition, almost all errors children make are over-regularizations (such as *hold-holded*); these make up about 10% of all irregular past tense inflections (Marcus et al. 1992; Yang 2002). By contrast–and this point is not widely discussed–over-irregularization errors such as "bring-brang", where the child over-applies an irregular rule, are exceedingly rare; these constitute only 0.2% of irregular past tense (Xu & Pinker 1995).[3] A direct implication of this finding is that the role of "analogy", however it should be properly formulated, is an extremely weak factor in morphological learning.

Children's acute sense of productivity in past tense accords well with the research on morphological acquisition in general. Crosslinguistically, children's morphological errors almost always involve omissions or over-generalizations of default forms, while the use of incorrect morphological forms (substitutions) are very sparse; see Phillips (1995) and Guasti (2002) for reviews. It is clear, then, children are very good at learning the appropriate use of rules: they recognize and generalize productive rules while memorizing the restricted use of unproductive ones.

The question is, How do they do that?

2.2 Learning and productivity

Children can assess the productivity of rules only after they learn what the rules are. Hence we sketch out a computational model of rule learning before diving into the productivity problem.

Twenty years of the past tense debate have produced no shortage of morphological learning models: too many, in fact, for us to give a review here. Moreover, since we are concerned with the productivity of rules, we will not discuss models such as connectionist networks that reject the use of rules on a priori ground.

The question of rule learning arises for models–such as WAR and ROW–that make explicit use of rule(s). It has been suggested that the default rule "add -d" can be learned by attending to the process that applies the majority of word types (Pinker 1999). This idea runs into problems in cases where the default class is not statistically dominant. An example is the much studied German noun plural system, which consist of five classes: *Kind-er* (children), *Wind-e* (winds), *Frau-en* (women), *Daumen-ø*(thumbs), and *Auto-s* (cars). Marcus et al. (1995) demonstrate that, despite the numerical minority (about 7% of all noun), the -s class is the default; an algorithm that counts type frequencies will not be able to learn the default status of such a rule.

A most promising approach to rule learning comes from a number of models developed by computer scientists (Mooney & Califf 1995; Sussman & Yip 1997), following an even earlier approach to learning in Artificial Intelligence (Mitchell 1982). These models are not well known in linguistics and cognitive science; we therefore take this opportunity to present an implementation of the Sussman-Yip model by Molnar (2001). The direct relevance of rule learning to our study of productivity will be made clear in Section 3.

To begin, we assume that the context of a morphological rule is inductively determined by the properties of the words it applies to. Stated in the classic form of generative grammar, rules are of the form:

(1) R: A→B/C___D

where A→B is the *structural change* of R and C___D is the context or *structural description* of R. Learning is viewed a search problem that incrementally discovers the commonalities among the properties of structures that form natural classes. Adapting it to morphological learning, the algorithm seeks the shared structural descriptions (C___D) among words that undergo the same structural change (A→B). It does so by making conservative generalizations, an idea that goes back to at least Chomsky (1955) and has been used in models of language learning in recent years as the Subset Principle (Berwick 1985). The learner constructs rules on an item-by-item basis, replacing older and more specific rules with new and more general ones as more items are processed. Figure 1 schematically illustrates how the "add -d" rule is learned.

When learning starts, there are no rules. Suppose now the first word *walk-walked* comes in. We are assuming, by no means obviously, that the underlying form and the derived form as well as their relations are available to the learner. The model identifies the structural change to be "add -d".[4] No generalization is possible on a single piece of data, and a trivial rule is learned by rote: **IF** *walk* **THEN** "add -d". Suppose the next word is *talk-talked*. As before, the learner learns: **IF** *talk* **THEN** "add -d". Generalization is now possible: the two rules constructed so far can be collapsed into one as both involve identical structure change ("add -d"). The learner proceeds to discover the (partial) commonalities between *walk* and *talk*: a conservative generalization yields that, for instance, they differ only in the first consonant, which then must not restrict the applicability of "add -d". An intermediate rule is formed: **IF** ⋆*alk* **THEN** "add -d", where ⋆ stands for "irrelevant", and the previous two separate statements

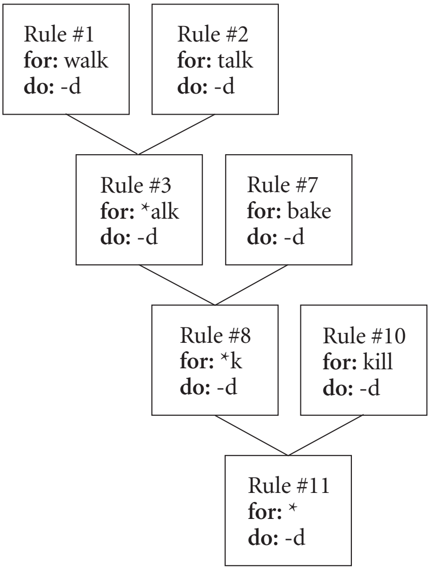


Figure 1. The learning of the regular rule "add -d" based on Molnar (2001).

will be eliminated. As more verbs are learned incrementally, the condition for the application of "add -d" becomes ever more general. Eventually, the learner will conclude that "add -d" has no restrictions (⋆'s all around as in Rule 11 in Figure 1). This corresponds to the notion of a default rule, which applies without restrictions.In the present learning model, the generality of a rule is directly related to the diversity of words it applies to.

The Sussman-Yip model has a number of attractive features. First, it is completely general and is applicable to any inductive inference problem that can be stated in the form of A→B/C___D. Its use could be extended to models of word learning, syntactic learning, and other inference reasoning problems. The implemented models (see Sussman & Yip 1997; Molnar 2001 for details) learn English morphology with far higher accuracy and efficiency than any other learning model reported in the literature. They learn the "add -d" rule in addition to the morpholexical rules for the irregulars as assumed in the ROW model.

Second, the model does not count any frequencies, and is thus a plausible candidate for solving the minority default problem in German plurals. As is well known, the "add -s" in German plurals consists of mostly foreign words, which are quite heterogeneous in their morpho-phonological properties. (In Section 5.2, we return to the German noun system in great details.) A model that identifies the generality of rules with the diversity of its application is perfectly suited for extracting this pattern.

The Sussman-Yip model is not without problems. Perhaps the most significant one is the assumption that morphological rules create direct mappings between the underlying form and the surface form, which closely resembles the two-level approach to computational morphology (Koskenniemi 1983). As noted by Barton et al. (1987), Anderson (1988), and recently Karttunen & Beesley (2003), the two-level approach seems inadequate for morphological phenomena that require cascades of ordered derivations (cf. Kiparsky 2000), whose effects can be directly observed in morphological learning (Yang 2002). We refer the reader to Ristad (1994) for a formalization of how ordered rules can be acquired, and some related complexity results (in the worst case).

In the rest of the paper, we will simply assume that the learner is capable to extracting morphological generalizations in the learning data, perhaps along the lines of the Sussman-Yip model in combination with some currently unknown mechanism that extracts generalizations over ordered morphological derivations. It is important to recognize that the learning model produces the *form* of rules, but it does not say anything about the *productivity* of rules. And productivity is the problem we are after.

2.3 Some terminological remarks

Before we proceed, it is useful to clarify a few terminologies that are essential to the study of productivity. We would like to call attention to the precise meanings of "default", "productive", and "(ir)regular". In many writings, those terms are used interchangably, yet we believe that finer distinctions ought to be made.

We take *default* to mean as a backup option, or "when all fails". That is, when more specific rules fail to apply, the default is used. (Note that this definition tacitly assumes the Elsewhere Condition in rule application; more on this later.) Thus, the default is the maximally general rule possible and its application has no restrictions.

We take *productive* to mean "predictable" and "generalizable". A rule is productive if it automatically applies to a set of lexical items characterized by certain properties, producing predictable derived forms, and can extend to others, including novel items that have the same properties. For instance, we may imagine a rule that says:

(2) **IF** a verb ends in a vowel, **THEN** add -n to form past tense.

Rule 2 is productive because one could make up a verb that ends in a vowel, and a speaker that knows 2 would be able to inflect the nonce word by adding -n. However, the productivity of rule 2 is more restricted than that of a default rule: the former applies to a subset of verbs characterized by specific properties (supposing that some verbs end in a vowel), while the latter applies to the set of all verbs (i.e., no restrictions on their properties other than being verbs). Thus, a productive rule needn't be the default, while the default rule is, by definition, necessarily productive.

In the case of English past tense, the only productive rule happens to be the default: the morpholexical rules for irregular verbs do not generalize to novel items (Pinker & Prasada 1995). But this needn't be so for languages in general.[5] One can imagine a morphological system that divides lexical items into several mutually exclusive subsets, each of which uses a perfectly productive rule–but there is no default rule that works for all words. Indeed, the Polish genitive case system has been argued to not have a default (Dabrowska 2001). It appears that the presence of the default follows from the composition of the linguistic data: if there is an unrestricted rule in the data, it will be learned, and if there isn't, there will be no default. The existence of a default, therefore, does not appear to be a hard requirement imposed by Universal Grammar: this is also the implication of the Sussman-Yip model of rule learning. In Yang (2005), we develop a historical explanation for why languages tend to have default rules in general.

Neither productive nor default rules need to be exceptionless. The case of default rules is obvious: "add -d" in English has to put up with some 150 exceptions. The same is true for productive (but non-default) rule. An example is the epenthesis rule in English plurals for nouns that end in sibilants (*church-churches*, *box-boxes*). This rule is completely productive as shown by nonce nouns (beamish-beamishes), and yet it is not 100% reliable: *fish*, *perch* and *tooth*, for example, must be marked as exceptions and memorized as such.

For the remainder of this paper, we will be using the terms "regular" and "irregular" to refer to not rules, but *words*, as in "English irregular verbs", "German regular nouns", and so on. The current literature on the mental lexicon regards regulars are those inflected by the regular rule, and irregulars are often taken to be all the rest, but this is really an unfortunate consequence of the supposed dichotomy inherent in the WAR model. Such a view is too simplistic, for it fails to consider the logical possibility that there may be perfectly productive rules, whose application does not require explicit memorization but are yet not as general as the default. In the morphological literature, words that follow productive but non-default rules have also been called "regular" (Wiese 1996; Wunderlich 1999), which is a practice that we shall follow. (In this sense, English nouns that end in sibilants are "regular" as well.) Note that (ir)regularity of words is dependent on the productivity of rules they fall under: as we understand the nature of productivity better, words that have been assumed to be "irregular" in the literature may turn out to be "regular" after all.

Once again, we are interested in productive rules in general, which include, but are not limited to, default rules. Specifically, we are interested in the problem that a learner must solve, apparently correctly and rapidly: what makes a rule productive and what makes it morpholexical?

3. Toward a computational model of productivity

3.1 The cost of exceptions

Imagine a child learning the past tense rules of English, following the schematic description of Sussman-Yip model in Figure 1. If the child has only learned two words *ring-rang* and *sing-sang*, then she might be tempted to conjecture a rule: /ing→ang/, which reads "**IF** /ing/ ending **THEN** change to /ang/". This would be a productive rule, at this point, for it is completely consistent with the data gathered so far. And, at this point, if the child learns a third word "bring", she would say "brang" for past tense. However, as her vocabulary increases, the /ing→ang/ rule will run into more and more exceptions: e.g., *bring-brought, sting-stung, swing-swung, wing-winged*, among others. If she still wishes to maintain the /ing→ang/ as a productive rule, then "bring", "sting", "swing", "wing", and so on must be memorized as exceptions. Now she may decide that the rule once thought to be productive doesn't work so well after all: more often than not, she couldn't use the rule productively and must resort to lexically marked exceptions (which are presumably inflected with other rules). Now she may decide that she's made a mistake. The rule /ing→ang/ is not productive: it is morpholexical and applies to a lexically restricted set of *sing* and *ring*, period.

By contrast, the rule "add -d" would pay off more handsomely: it works for a great deal more words than the some 150 exceptions. Making "add -d" productive only requires a relative small number of exceptions committed to the memory. Viewed this way, the productivity of a rule hinges on an evaluation procedure that does the right kind of cost-benefit analysis on the basis of both well-behaving items and exceptions for a rule that may be productive.

The general flavor of this problem is not new. There is a large literature, mostly in psychology and AI, that is dedicated to categorization, inductive generalization, and decision making under inconsistent data, and a variety of techniques could in principle be applied to the problem of assessing rule productivity. However, as Chomsky (1965) stresses, the choice of the evaluation procedure for a linguistic theory "is not given a priori ... Rather, an proposal concerning such a measure is an empirical hypothesis about the nature of language." (p. 37); see also Sober (1975). The present work introduces an evaluation procedure from the perspective of *computational complexity* but its suitability must be rigorously defended as an empirical hypothesis (as we do in Section 3.2 and 3.3.)

In general, there are two measures, space and time, that may enter into complexity calculation. Both metrics are in principle valid, and the choice must be based on their empirical status. Recently, there has been interest in the induction of linguistic structures, often crouched in the Minimum Description Length (MDL) framework (de Marcken 1996; Brent & Cartwright 1996; cf. Clark 2001). This framework views the grammar as a data compression device: it strives to minimize the structural descriptions of the relevant linguistic data, an idea that goes back to Chomsky (1955). At some level, this approach seems correct. A leading idea of generative linguistics is to eliminate redundancies to arrive at ever more economical representations of the grammar. To the extent that the ROW model (Yang 2002) is correct, the human learner does appear to construct space-saving rules rather than direct storage of inflected forms, even when the saving achieved for the 150 or so irregular verbs is quite limited. Yet questions remain on the applicability of the MDL approach to grammars, which requires independent motivations. Our approach here, by contrast, makes use of time complexity in the evaluation of morphological productivity. As we shall see, there is a good deal of empirical evidence that bears on the data structure and algorithms that may be used by the mental lexicon: a non-trivial evaluation procedure based on time complexity can be developed as a result.

Our approach complements but differs from the existing literature on morphological productivity. One could, for instance, conduct a series of Wug-like tests on specific morphological rules/processes, gather the relevant corpus statistics, and in some cases, develop a quantitative formula of morphological productivity (Aronoff 1976; Baayen & Lieber 1991; Prasada & Pinker 1995;

Albright 2002; cf. Bauer 2001). While this approach has unearthed many subtleties in the morphological system, it is, at best, a statistical summarization of data that "accord nicely with [linguist's] intuitive estimates of productivity" (Baayen & Lieber 1991: 01), rather than an independently motivated theory of morphological computation. Even if the actual criterion for productivity established from empirical tests turns out to be true – e.g., a productive rule can tolerate 25% of exceptions – we still would like to know why the magic number is 25%, rather than 18% or 32%.

In the remainder of this section, we shall first develop an algorithmic model of how rules are applied to form morphologically complex words. This is followed by some evidence for the model as a psychologically plausible theory of morphological processing.

3.2 Morphological knowledge and use

The computation of morphologically derived words involves at least three processes (Caramazza 1997; Levelt et al. 1999, etc.):

(3) a. **word search**: lexical lookup of the stem (e.g., *walk*)
b. **rule selection**: search for the appropriate morphological rule(s) for the stem (e.g., "add -d" for past tense)
c. **rule application**: use the rule to generate the inflected form (e.g., *walked*)

In principle, these three components are executed independently and sequentially, and in general, they must. To process morphologically complex languages such as Turkish and Finnish, holistic storage of derived forms, which numbers in the billions (Hankamer 1989; Niemi, Laine, & Tuominen 1994), cannot in principle work.

Overall, morphological processing appears to be geared toward real-time processing efficiency. A telling source of evidence comes from frequency effects. One of the earliest and most robust findings in lexical processing is that high-frequency words are recognized faster and more reliably than low-frequency words (Forster & Chambers 1973; Whaley 1978; Balota & Chumbley 1984). Similar frequency effects also obtain in the selection (3b) and application (3c) of morphological rules (Taft 1979; Burani, Salmaso & Caramazza 1984), which can be assessed after the frequency effects from word search (3a) are neutralized.

In recent years, gradient effects such as frequency in the lexicon are widely considered to be problematic for abstract representations of linguistic structures in generative grammar; see Seidenberg & Gonnerman (2000), Bybee (2001), Baayen (2003), among others. Like the earlier controversies regarding the "psychological reality" of linguistic theories (Chomsky 1980), these reactions run the risk of throwing the baby out with the bath water. More empirically, there are at least two possibilities that may explain the perceived irrelevance of linguistic theories to linguistic processing.

On the one hand, frequency effects are a hallmark of the mental lexicon– and a hallmark of human learning and memory in general. I recognize my son's face faster than my cousin's presumably because I see my son far more often, and it says nothing about how faces are represented and recognized, or how I can recognize a face, any face, in the first place. Similarly, frequency effects in morphology say nothing about the linguistic representation of morphology unless one can tease apart the linguistic representation of morphological knowledge from the general, and frequency-sensitive principles of memory and learning that may be implicated in linguistic performance. For example, the ROW model is a theory that amends the classic conception of morphophonological rules (and parameters in syntax) with a stochastic component of probabilistic learning that may be independent of language (Yang 2002, 2004).[6] On the other hand, perhaps one needs to look harder at the morphological theory from the perspective of morphological processing. It is possible, as this paper contends, that a performance theory of morphology directly follow from a competence theory of morphology; cf. Miller & Chomsky (1963), Berwick & Weinberg (1984), Fodor & Crain (1985).

I propose to follow the time-honored tradition of *Lexical Search Theory* in morphological processing (Rubenstein, Garfield, & Millikan 1970; Forster 1976, etc.). The Lexical Search Theory takes the strong position of viewing lexical processes as information retrieval algorithms which, as we show in this paper, are amendable to formal analysis. More specifically, we follow Forster (1976) in suggesting that lexical processing involves *serial search* that is sensitive to the token frequency of words.[7] In the simplest form, a list of words that is arranged in decreasing frequency straightforwardly accounts for lexical frequency effects; see the Appendix for a more precise formulation. Moreover, and crucial to the present work, we suggest that the **rule selection** component (3b) of morphological processing also follows a search search algorithm.

Traditionally, linguists hold that the organization of morphology – and perhaps other arenas of linguistic representations and processes – is governed by the Elsewhere Condition (Anderson 1969; Kiparsky 1973; Halle 1990; Halle & Marantz 1993), an insight that dates back to Pāṇini. The Elsewhere Condition provides a mechanism for representing exceptions along with a productive rule. For example, suppose a rule R is defined over a set of N words $\mathcal{N}$, and out of these, a subset $\mathcal{M} = \{w_1, w_2, \ldots, w_M\}$ of M words are exceptions that do not follow R. The Elsewhere Condition directly translates into a serial search algorithm for inflecting word $w \in \mathcal{N}$:[8]

(4) Elsewhere Condition Serial Search (ECSS):

```
    IF $w = w_1$ THEN …
    ELSE IF $w = w_2$ THEN …
    …
    ELSE IF $w = w_M$ THEN …
    ELSE  / apply $R$
```

This is how the ECSS model works. If $w \in \mathcal{M}$, then one of the M exception clauses will be matched, and w will be inflected accordingly (i.e., not with R).[9] If w is not one of the exceptions, rule R will be triggered but only *after* all the exception clauses have been evaluated (and rejected). Whenever a match is found, the search halts.

Suppose now that the exception clauses in the ECSS model are ordered with respect to the frequencies of their access (as measured by their token frequencies). Our proposal, then, is very similar to the "bin" model of lexical access (Forster 1976, 1992). For Forster, the bins correspond to a set of words with specific orthographic properties; for us, a bin is defined by a rule, which is also defined over a set of words (characterized by certain structural properties). A productive rule defines a bin like 4, which contains words with specific properties, some of which may be exceptions the rule.

3.3 Evidence for serial search

There is several strands of evidence that suggests the ECSS model is on the right track.

First, the serial search model predicts that the lexical access time is correlated with the *ranks* of words, i.e., their positions on the search list. Rank is of course related to frequency but it is a relative notion rather than an absolute one. These two notions are rather difficult to distinguish and have not been explored extensively but recent evidence (Murray & Forster 2004) suggests that rank provides a better fit with reaction time data in lexical decision than frequency. Although ranking a list by frequency in a strictly decreasing, and thus optimal, order appears to be computationally taxing, there are a number of so-called randomized algorithms that are computationally trivial yet guarantee near optimality of the sorted list. So far as we know, the connection between these algorithms and psycholinguistic modeling has never been explored; see the Appendix for some observations and suggestions.

Second, a serial search model such as 4 predicts frequency effects in the inflection of the exceptional forms. Note that the proper test for this prediction must hold the stem-cluster frequencies of words constant:[10] this neutralizes the influence from the frequency effects in the word search process (3a). Indeed, frequency effects among the exceptions – e.g., English irregular verbs – are among the most relibale foundings in morphological processing (Prasada et al. 1990; Seidenberg 1992; Ullman 1999). This has typically been taken as evidence for a frequency-sensitive memory retrieval process responsible for the storage of irregulars (Pinker 1999), but it is clearly also compatible with the serial search model suggested here.

A final claim of the ECSS model in 4, and an important one, is that during the rule search procedure, words that follow *R* will not be accessed until all rule-defying exceptions are examined and rejected. In this sense, the present model remains true to Forster's bin search model, in which non-words are not recognized until all actual words in the bin are evaluated and rejected. On this matter, it seems that the jury is still out. The prediction is difficult to evaluate directly. Note that we do not predict, for example, regulars are necessarly processed slower than irregulars. To inflect a word, as 3 (repeated below) illustrates, requires three distinct stages:

(5) a. **word search**: lexical lookup of the stem (e.g., "walk")
b. **rule selection**: search for the appropriate morphological rule(s) for the stem (e.g., "add -d" for past tense)
c. **rule application**: use the rule to generate the inflected form (e.g., "walked")

The ECSS model is only concerned with step (5b), the selection of the appropriate morphological rule. While the effect of word search can be controlled via stem-cluster frequency, we still end up with reaction time data that are composite of rule selection (5b) and rule application (5c). Only after we factor out the time course of rule application can we accurately test the predictions of the ECSS model regarding rule selection.

There are a number of studies suggesting that regular inflections are produced somewhat faster than irregulars (Prasada et al. 1990; Seidenberg 1992),[11] which seem to contradict the claim of the present model. However, following the discussion immediately above, the faster processing time of the regulars may be due to the speedup in the application of the default rule "add -d", which has a greater presence (via token frequency) than any specific irregular rules. Although irregulars altogether make up for over 60% of past tense tokens (Grabowski & Mindt 1995), no single irregular class can match the 40% probability mass of the "add -d" rule. Thus, regulars would seem to have an advantage over irregular verbs in the rule application stage of inflection.

The claim that rule application is frequency dependent is supported by evidence from both language acquisition and processing. As shown in Yang (2002), children produce more errors when learning the rule "-ø & Rime→/u/" (*blow*, *fly*, *draw*) than learning the equally arbitrary rule "-&̑ Rime→/a/" (*think*, *catch*, *buy*). Some children do better at vowel-shortening irregulars such as *bite-bit* and *shoot-shot* than even *go-went*: the latter surely ranks among the most frequent irregular verbs. These disparities are attributed to rule frequencies. For example, the "-t & Rime→/a/" rule is used by verbs with higher frequencies, and the vowel-shortening verbs take advantage of, well, vowel-shortening, a robust process in the English language (Myers 1987; Halle 1997). These factors contribute to the success of irregular past tense inflection in language learning.

On the processing side, Sereno & Jongman (1997) find that English nouns are generally inflected faster than verbs. This is argued to derive from the fact that there are more tokens of plural nouns than past tense verbs, and thus speakers of English have "more" experience with the "add -s" rule than with the "add -d" rule and can use the former faster than the latter.[12] If so, more frequent rules are faster in online application than less frequent ones.

Thus, to test the time course of rule selection in the ECSS model, one must eliminate the confounding factor of rule frequency. We are not aware of any study that directly address this issue: much of the recent literature on morphological processing has focused on the rule vs. memory dichotomy postulated by the WAR model, and no serious attention has been paid to the possibility that irregulars may be inflected by rules of a more limited kind. However, a potentially relevant experiment is a production latency study by Clahsen et al. (2004) on the inflection of German past participles. They found that, at least for high frequency verbs, irregular past participles are formed slightly faster than regulars. Irregulars (-t) and and regulars (-n) in German participles are imperfectly but reasonably well matched in frequencies (Clahsen 1999): at least they are not as lopsided as in the case of English past tense. Therefore, the influence from the rule application stage would be more or less neutralized in the case of German participles, and Clahsen et al.'s (2004) study constitutes suggestive evidence for the ECSS model.[13] Future research needs to identify the appropriate empirical domains where rule frequency can be controlled for and to further investigate the time course of rule selection under the ECSS model.

In sum, we believe that the ECSS model 4 is a reasonable approach to real-time morphological processing. But the Elsewhere Condition, when used to implement productive rules, may come at a considerable cost. If a rule is productive, and it happens to have too many exceptions, then significant complexity may incur as the exceptions must be searched and rejected before the computation of the rule-following items. The next section gives a precise model of how many is "too many".

4. The tolerance principle

Suppose the learner has observed a generalization in a set of N words $\mathcal{N}$:

R = **IF** X **THEN** Y

X is the structural description of $\mathcal{N}$, some properties, whatever they are, that hold for all words in $\mathcal{N}$, e.g., all end in /ing/, all are verbs, or all are feminine nouns. Y is the structural change, whatever it is, that results from matching X, e.g, "-t & Rime →/u/", "add -d", add a prefix followed by vowel harmony. However, the learner notes that $\mathcal{N}$ contains some exceptions, namely, a subset $\mathcal{M} \subseteq \mathcal{N}$ with the cardinality of M, the members of which do not follow R. In other words, $\mathcal{M}$ are the exceptions to the rule. If the learner wishes to make R productive, then the words in $\mathcal{M}$ must be explicitly committed to memory, as depicted in 4. Alternatively, if $\mathcal{M}$ is too large, then the learner may decide R is not productive after all and proceed to store all words in $\mathcal{N}$ as exceptions. Therefore, the learner has two ways of storing and processing words in $\mathcal{N}$ with respect to R:

(6) a. R is a productive ($w_i \in \mathcal{M}$)

```
IF $w_1$ THEN ...
ELSE IF $w_2$ THEN ...
...
```

```
        ELSE IF $w_M$ THEN ...
        ELSE APPLY R
b.  R is a morpholexical (i.e., unproductive) rule ($w_i \in \mathcal{N}$)
        IF $w_1$ THEN ...
        ELSE IF $w_2$ THEN ...
        ...
        ELSE IF $w_N$ THEN ...
```

Let $T(N, M)$ be the expected time of using (6a) where M exceptions co-exist with (N–M) rule-following items. On the one hand, if R is morpholexical, then all N items must be listed explictly as if all are exceptions, in which the expected time complexity would be $T(N, N)$ (i.e., $M = N$, all words are exceptions).

We conjecture:

(7) **Tolerance Principle (TP)**

a. If $T(N, N) < T(N, M)$ then R is morpholexical ($\mathcal{N}$ will all be explicitly stored, as if they have nothing in common).

b. Otherwise, $\mathcal{N}$ is computed by a productive rule R plus a list of M exceptions.

Note that the Tolerance Principle is an optimization procedure for online processing of words. It takes into account for both token and type frequencies. Token frequency optimization is reflected in the ECSS model that ranks clauses by frequency. The Tolerance Principle itself, however, is directly expressed as a function of type frequencies ($T(N, N)$ and $T(N, M)$; see Aronoff (1976), Bybee (1995), Bauer (2001) for similar views on productivity.

The Tolerance Principle suggests that if there are too many exceptions (e.g., M is large relative to N), then the cumulative cost of finding R for words in $\mathcal{N} - \mathcal{M}$ becomes prohibitive because the ECSS model must reject the words in $\mathcal{M}$ first. At this point, the learner may decide against making R into a productive rule. Under reasonable assumptions about word frequency distributions, the threshold value M_c at which a rule becomes or ceases to be productive can be calculated as a function of N. This is achieved by solving for M in $T(N, N) = T(N, M)$. Some of the mathematical details, and a number of algorithms that can rank a list of rules by frequency, are presented in the Appendix. We simply give the analytic result below:

(8) **Theorem:** $M_c \approx N/\ln N$

Figure 2 shows the special case of $N = 100$ and where $M_c = 22$.

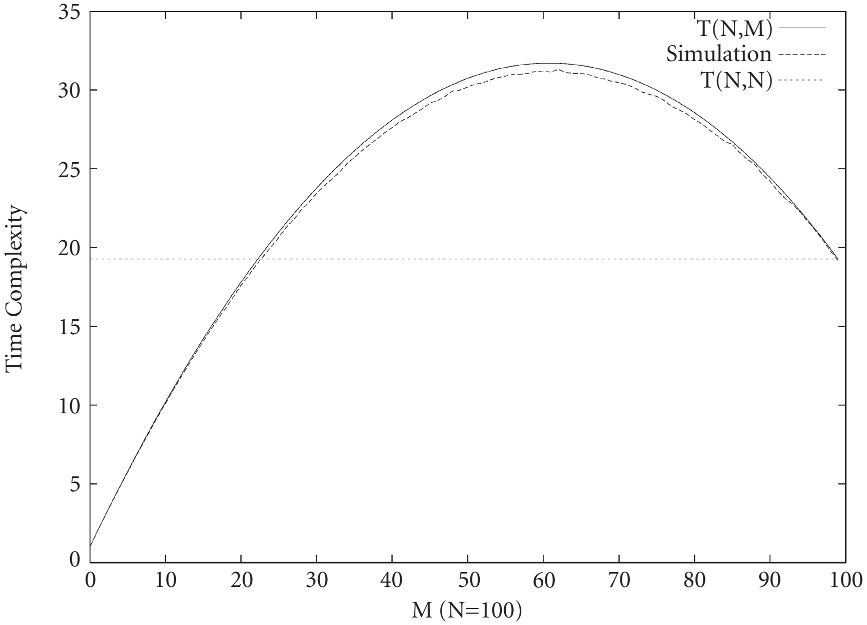


Figure 2. The cost of storing M exceptions and the rest as a productive rule ($T(N, M)$), compared with storing all words as irregulars ($T(N, N)$), where $N = 100$. The analytic approximation of M_c in Theorem 8 is compared compared with simulations average over 10,000 trials.

In Figure 2, the cost of storing all 100 words as exceptions is represented by the straightline $T(N, N)$. This constitutes the baseline against which $T(N, M)$ is compared. When M is small, it is clearly cheaper to have a rule. However, once M crosses the critical threshold M_c (22 for $N = 100$), it becomes more expensive to have a productive rule now that 78 words have to wait for the 22 time units before the selection of the productive rule. At this point, the economical solution actually is to store *everything* as exceptions. As the number of exceptions approaching 100, the cost of T(N, M) becomes closer and closer to $T(N, N)$, and they eventually meet again when M=N.

Perhaps a surprising consequence of these calculations is that it doesn't take a lot of exceptions to derail a potentially productive rule: $M_c = N/\ln N$ is in fact sub-linear, whose effects become more dramatic when N is large. (In natural languages, however, N is the size of a subset of the vocabulary and is unlikely to be very large). In other words, the postulation of a productive rule requires it to *actually* apply to a great majority of words that it *could* apply to. For example, a rule like "add -d", which is in principle applicable to all English verbs, can be productive only if the number of irregulars–those that do not add "-d"–is relatively small compared to the English verb vocabulary. English has about 150 irregular verbs, the theorem 8 implies that there need to at least 1000 verbs algother to ensure the productivity of "add -d" is safe. This requirement is easily met in the English language.

We turn to some empirical consequences of the Tolerance Principle in Section 5.

5. Tolerance principle in action

How do children use the Tolerance Principle in language acquisition? It goes like this. The learner first identifies a generalization, or a rule R based on the input data, perhaps using something like the Sussman-Yip learning model. She then notes that in the vocabulary she has acquired so far, R should work for N words but actually works for $(N - M)$ words, i.e., there are M exceptions. The learner proceeds to use the Tolerance Principle to calculate whether M has crossed the threshold value under N. If so, R is not productive and all N words are memorized explicitly. Otherwise, R is productive: M exceptions are memorized, and the $(N - M)$ rule-following items are handled by R automatically. Note that the productivity of R depends on N and M, which in turn depends on the composition of the vocabulary. As the vocabulary changes, the productivity of rules may change as well. We address these issues presently.

5.1 Tolerance and learning

The (low) tolerance of exceptions leads one to make some immediate predictions on child language learning in subtle connections with word frequency distributions. By the time a productive rule is recognized as such, it must be the case that the learner has accumulated a good deal more rule-following items than exceptions.

Consider now the case of past tense learning in English. If we take the first instance of over-regularization as evidence for the successful acquisition of the regular rule, then at that stage there must be considerably more regular verbs than irregulars in the child's vocabulary as predicted by the Tolerance Principle. Although it is very difficult to get precise estimates of the child's vocabulary at any stage of acquisition, Marcus et al. (1992) made a useful attempt. Their estimates of the regular vocabulary size were based on the verbs attested in children's speech. As they noted, this invariably underestimated the vocabulary size, as children do not use always say all the words they know. In addition, regular verbs are more affected by under-sampling than irregular verbs, the latter of which tend to have higher token frequencies, at least in English. Nevertheless, Marcus et al.'s (1992) counts lend general support to our model. Based on attested verb use, the three children in Brown's (1973) study had 63% (Adam), 55% (Eve), and 56% (Sarah) of regular verbs respectively at the first instance of over-regularization. By using the so-called Jackknife estimate from biostatistics, the percentages of regular verbs are 62% (Adam) and 64% (Sarah), where Eve's files did not yield sufficient data for extrapolation. These numbers are lower than what the theorem $M_c \approx N/\ln N$ predicts, which we tentatively attribute to the under-sampling of regular verbs. Significantly, though, there are considerably more regulars than irregulars when the "add -d" rule becomes productive, which is consistent with the general directions of our theoretical proposal. result.

A related prediction concerns the status of the U-shape curve in language learning. For some children, past tense acquisition starts with a period of correct inflections (of both regulars and irregulars).[14] This is followed by over-regularization as the regular rule emerges (see above). Given the specific word frequency distributions in English, this pattern is natural under the Tolerance Principle. Because irregular verbs occupy the majority of token frequencies in English, they are likely to be the earliest verbs a child acquires. Consequently, there will be relatively few regular verbs. Even so, just a few regular verbs could in fact lead to the emergence of the "add -d" rule. Recall the Sussman-Yip model of morphological learning, under which the structural description of a rule are given by the diversity of words it applies to: in implementation, the "add -d" rule can usually be learned on the basis of 20–30 regular verbs (Sussman & Yip 1997; Molnar 2001). The interesting point is that, even after "add -d" rule is induced, the child may not be able to use it productively. At the early stage of word learning, the significant proportion of irregulars in her verb vocabulary may push M above the critical threshold. The child, then, has no choice but to memorize regularly inflected verbs as if they were unrelated. It is only after more regular verbs are acquired – M drops below $M_c = N/\ln N$ – did the child realize that the "add -d" rule is productive after all: overregularization would follow as a result. Note that we do not predict the U-shape curve as a universal phenomenon of language or morphological learning: its existence or absence depends on the specific values of M and N for each child, whose vocabularies may vary considerably.

Note that *token* frequency of words directly affects the composition of early vocabulary, but it is the *type* frequency of words within the early vocabulary that directly influences the productivity of rules, as formulated by the Tolerance Principle. The emergence of productive rules may well differ in morphological domains where word type distributions may differ. Consider the English noun pluralization, where the regular rule is "add -s", and there are a few dozen irregulars scattered around. However, unlike the case of English irregular verbs, the irregular nouns are not heavily concentrated in the uppermost echelon of noun frequency. This means that among the first plural nouns that children learn, only very few are going to be irregulars. The Tolerance Principle then predicts that the "add -s" rule become productive very quickly–faster than the "add -d" rule, for instance. As far as we know, this prediction is empirically true (Brown 1973), nor do we observe any longitudinal evidence for a U-shape learning curve (Falco & Yang 2005)

5.2 Tolerance and productivity

The Tolerance Principle makes sharp predictions about morphological productivity. In general, there are three kinds of evidence that directly bear on the validity of our proposal.

First, one may design artificial language learning experiments where the distributions of regulars and (unpredictable) irregulars are carefully controlled. We can then observe how children generalize when faced with inconsistent patterns. Kam & Newport (2005) is a recent example though the learning data there consists of grammatical constructions rather than morphological patterns. They discovered that children could tolerate a certain level of exceptions and impose systematicity on the learning data, while adults are far more hesitant (and appear to be probability matching as in the parameter setting model of Yang 2002). It would be interesting to see whether the level of inconsistency tolerated by children in such and similar studies can be predicted by our theoretical calculations. Furthermore, we speculate that the Tolerance Principle may be correlated with the "critical period" in language acquisition, and is thus active in young children and inert in adults.

Second, there is potentially a wealth of evidence to test the model in historical change. Over time, words have shifted morphological classes. For example, *cleave-clove-cloven* is now *cleave-cleaved-cleaved*, and for many speakers, *strive-strove-striven* has become *strive-strived-strived*; this kind of change is called "analogical leveling" where an irregular word got regularized. There are also cases of "analogical extension" where a regular got irregularized: the past tense of *wear* was *werede*, which would have been regular had it survived, but in fact it took on the *bear-bore* and *swear-swore* class and is now *wear-wore*. If something like the current proposal is correct, then the morphological rule that had just picked up a new word must have been productive at that particular period of time in history. Moreover, productivity under the present model is a categorical one, a position held by morphologists such as Bauer (2001) and Anderson (1992): either *M* is below the critical threshold or not. This would lead to a reformulation of the notion of "analogy" in historical morphology. We can examine the relevant historical corpus to find the corresponding values of *N* and *M* at the time when a given word shifted classes. See Yang (2005) for a preliminary investigation along this direction.

Finally, one may carry out corpus analysis of synchronic morphological systems and find the relevant sets of $\mathcal{N}$ and $\mathcal{M}$. The model would make predictions about the regularities within $\mathcal{N}$, which can be then checked against the Wug test and other benchmarks for productivity (though see Schütze 2005 for methodological issues associated with such tasks). The relative ease of this approach will constitute the bulk of research in our producitivity study and will take us to revisit some of the traditional discussions in morphological productivity such as the English affix problem (Aronoff 1976). For the moment, we turn our attention to some mysteries and intricacies in the German noun system.

5.3 Regular "irregular" nouns

I once thought that the German irregular nouns were individually memorized for plurals: not directly as in the WAR model, but still sorted to their respective rules by repeated exposure, just like English irregular verbs in the ROW model. This is what I wrote a few years ago:

> (A) quick glance at German shows that the four irregular classes of plural show no systematic similarity whatever. The horrors of German are real: one must sort each irregular noun into its proper class, as in the traditional rule-based view. (Yang 2000)

The problem is, the horrors of German are *too* real. There are a lot of irregulars to slot under their respective irregular classes. According to the CELEX database, the default class ("add -s") consists of only 7% of nouns (by type): it means that 93% of the nouns would have to be memorized by brute force.[15] It didn't feel quite right back then, and now I think this statement is completely mistaken. At least one of the four "irregular" classes of German nouns, the feminine nouns to be specific, is actually pluralized by a productive rule: add -(e)n. No memorization for feminine nouns is necessary, save for a few dozens of feminine nouns that do not add -(e)n and thus must be treated separately as exceptions.

To begin with, every German textbook, while quick to point out the exceptional cases, do not deny that the reliable generalizations exists in the pluralization of nouns. And these generalization usually concern the correlation between gender (and sometimes phonological properties) and plural inflection. It has been noted that masculine and neuter nouns whose final syllables contain a schwa usually add nothing, polysyllabic feminine nouns add -n, feminine nouns never add -er, and so on.

In the empirical literature, there have been some problematic findings for the simplistic regular vs. irregular division of the German nouns. These findings, by contrast, suggest that some "irregular" nouns actually form plurals highly productive ways–at least as productive as the "add -d" rule in English past tense. First, acquisition studies. German children overregularize their nouns about 10% of the time (Veit 1986; cited in Clahsen et al. 1992) and it has been known since Park (1978) that the -n suffix, an irregular one according to the WAR model, is in fact the most frequently overused suffix in children's natural production. Moreover, Clahsen et al. (1996) find in an elicitation study of low frequency nouns that that German children use wrong plural forms 18.5% of time overall, and 7.6% if overregularizations involving -s are excluded. These error rates are far greater than those with which English children misuse irregular past tense form: recall that over-irregularization errors make up only 0.2% of English children's past tense (Xu & Pinker 1995). On this basis alone, we must conclude that at least some of German "irregular" rules are as *regular* as the English regular past tense rule.

A highly predictable class seems to be that of feminine nouns, most of which, according to descriptions of German grammar, add -(e)n in plurals. There are also feminine nouns that do not add -(e)n: this is German after all! Thus, it is interesting to assess, on purely mathematical basis a la the Tolerance Principle, whether the following rule is productive in German plural formation:[16]

(9) **IF** feminine **THEN** add -(e)n

Here we have a straightforward application of the theorem 8. On the one hand, we count the total number of feminine nouns, namely, *N*: those are the candidates for adding -e(n) in plurals if rule 9 is productive. On the other, we count the number of the feminine nouns that do not add -(e)n, namely, *M*.[17] These would be the exceptions to rule 9, if that rule is productive. First, let's find the exceptions (see Footnote 17). Some feminine nouns do Umlaut: Axt (ax), Bank (bench), Braut (bride), Brust (breast), Faust (fist), Hand (hand), Mutter (mother)… Wand (wall), Wurst (sausage). A few, all monosyllabic, simply add -e: Tann (wood), Trupp (troop), etc. Some feminine nouns add -s: those are loan words: Kamera (camera), Restaurant (restaurant), etc. And at least one adds -ø: die Mark (Deutsche Mark, the currency). Overall, we have found *M* = 80 exceptions. The Tolerance Principle tells us that there needs to be at least 500 feminine nouns altogether–or 420 that *do* add -(e)n–to support a productive rule 9. This is easily met: there are in fact at least 9,500 feminine nouns in German based on the CELEX corpus (3600 if excluding compounds). The Tolerance Principle, then, predicts a productive rule of feminine noun inflection, while the some eighty exceptions must be explicitly memorized as exceptions.

Our conclusion that German feminine nouns follow a productive rule is not a novel one. A number of German morphology scholars (Wiese 1996; Dressler 1999; Wunderlich 1999) have regarded the WAR model to be too coarse-grained to capture the considerable regularities within the "irregulars". Indeed, a rule like 9 has been proposed to account for the feminine-(e)n correlation. The significant aspect of the present work is that the existence of a productive rule is predictable on purely mathematical basis–one doesn't even have to know German to do that.

The most direct evidence bearing on the feminine noun rule comes from a study of plural formation in both aphasic and unimpaired German speakers (Penke & Krause 2002). In an elicitation study, aphasic patients showed no frequency effect in error rate for feminine nouns that add -(e)n in plurals. Moreover, unimpaired subjects show no frequency effect for -(e)n feminine nouns in reaction time during a lexical decision task. The lack of frequency effect is generally a hallmark for productive rule, as has been demonstrated for the "add -s" noun class in German (Clahsen et al. 1997; Clahsen 1999).

The present discussion of feminine nouns leads to a more articulated picture of the German noun plural system. We will not discuss the noun cases that add -e, -er, and -ø: some of these classes may turn out to be productive, and

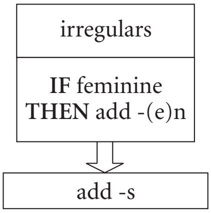


Figure 3. The inflection of German noun plurals: only the "**IF** feminine **THEN** add -(e)n" class is depicted.

they await future investigations along the lines sketched here. German noun plurals appear to be handled by a nested structure of rules, as illustrated in Figure 3.

The inflection of nouns goes through a cascade of rules. A feminine noun is first fed through the feminine rule as in 9. If it's one of the exceptions to the rule, then some special form will be used for plurals. Otherwise, -(e)n will be added. The "add -s" rule is the waste basket that takes over when the more specific rules–productive or otherwise–fail to apply. We note that this structure is roughly equivalent to the inheritance tree model proposed by Wunderlich (1999) in a different theoretical framework.

6. Conclusions

This paper can be viewed as an extrapolation from a set of known facts of morphological theory, processing, and acquisition. The point is not whether $M_c = N/\ln N$ is ultimately right or wrong. Rather, it is that this kind of work can be done: productivity can be studied rigorously in a formal framework that is empirically motivated.

The logic of the Tolerance Principle extends to syntactic learning as well, as illustrated in Figure 4.

As in morphology, where exceptions are individually listed and accessed prior to the rule-following items, a phrasal construction can be preceded by a set of idiomatic expressions, a core parameter value can be preceded by a set of constructions that use the opposite value (such as the range of facts regarding subject use in Englsh shown in Figure 4). For instance, suppose that an English child, guided by UG and perhaps general learning principles, has acquired a

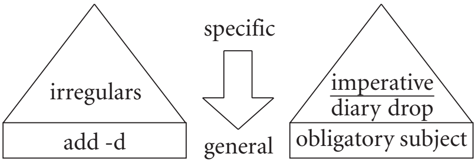


Figure 4. Elsewhere Condition in morphology and syntax.

set of verbs with some common characteristics (e.g., motion verbs). She notes that some of these verbs allow post-verbal subject (as in the case of locative or expletive inversions), while others do not. Now we have on hand the very situation as in the Tolerance Principle for morphology: N now corresponds to the total number of verbs in this class, and M is the number of verbs that do not trigger inversion. She can then use the Tolerance Principle to see whether inversion is a productive process by checking whether M falls below the threshold of $N/\ln N$. If so, then all these verbs trigger inversion save those M that do not: the child has learned a productive process despite the inconsistent data in the input. On the offer is of course a thought experiment, but these are research topics that one may pursue under the methodologies and results of the present work.

I understand that not everyone is interested in grammar learning under exceptions, be it morphological or syntactic. And that is just as well. Assuming that the problem of productivity and exceptions can be handled by a principled theory of learning, as we have tried to develop in the present paper, linguists can stay focused on the core problem of systematic regularities in language; that is, the part of the grammar that doesn't leak.

Appendix

Consider $T(N, N)$: the expected time to find the appropriate rule for a word in a list ranked by frequency of access.[18] Assume, as in standard in corpus-based linguistic modeling, that the frequencies of words in $\mathcal{N}$ follow a Zipfian distribution with the exponent of 1. That is,

$$\forall i, W_i \in \mathcal{N}, \quad p_i = \frac{1}{iH_n},$$

where

$$H_N = \sum_{i=1}^{N} \frac{1}{i}, \text{ the Nth Harmony Number}$$

We make the Zipfian assumption because, for whatever (uninteresting) reason, it appears to give a good fit for actual word frequencies, and it makes the calculation easy. In an optimal listing of $\mathcal{N}$, the time it takes to access the ith clause is its rank $r_{i,\mathcal{N}}$, which is i. Thus, the optimal time complexity of a list of N items is

$$T(N, N) = N/H_N$$

The naive algorithm for realizing N/H_N is computationally expensive, for it requires the learner to keep track of the frequencies of the exceptions and then rank them accordingly. Fortunately, there is a class of randomized, or "on-line", algorithms designed expressly for optimizing frequency-sensitive computation. One simple instantiation is the `MOVE-TO-FRONT` algorithm: whenever an exception is encountered, the corresponding clause is promoted to the beginning of the list. Another is the `MOVE-UP` algorithm (Rivest 1976), which swaps a clause just encountered with the one preceding it. Both algorithms are computationally trivial and both are known to be only worse than the optimal algorithm by a (small) linear factor (Sleator & Tarjan 1985a). Hence, it is reasonable to assume that a human learner is capable to constructing a near optimal list of irregulars.[19]

Consider now $T(N, M)$. Recall $\mathcal{M}$ is the set of exceptions. According to the ECSS model 4, we consider two cases for each word w_i:

$$T(N, M) = \sum_{i=1}^{N} p_i C_i,$$

where

$$C_i = \begin{cases} E[K_{i,\mathcal{M}}], & \text{if } w_i \in \mathcal{M}, \text{ and } K_{i,\mathcal{M}} \text{ is its rank in } \mathcal{M} \\ M, & \text{if } w_i \notin \mathcal{M} \end{cases}$$

Assume $\mathcal{M}$ is a random subset of $\mathcal{N}$, and $w \in \mathcal{N}$ has a uniform probability of $1/N$ being an irregular. (We turn to the justification of this assumption after presenting the theoretical results below.) It is easy to show that the probability of a word being chosen as an exception is $x = M/N$. We have:

$$\begin{aligned} T(N,M) &= \sum_{i=1}^{N} p_i [xE[K_{i,\mathcal{M}}] + (1-x)M] \\ &= x \sum_{n=1}^{N} p_i E[K_{i,\mathcal{M}}] + (1-x)M \end{aligned}$$

For our purposes, $T(N,M)$ can be further simplified as:

$$T(N,M) = xT(M,M) + (1-x)M, \text{ where } x = M/N$$

This is made on the assumption that a random and independent sample from a Zipfian distribution is still largely a Zipfian distribution (with fewer items). Simulation results suggest that this is sound.

M_c can be obtained by solving for M in $T(N,N) = T(N,M)$. By approximating H_N with $\ln N$, it is not difficult to show, we have

$$x\frac{M}{\ln N} + (1-x)M = \frac{N}{\ln N}$$

It follows that

$$x^2 \frac{1}{\ln N + \ln x} + (1-x)x = \frac{1}{\ln N}$$

Let

$$f(x) = x^2 \frac{1}{\ln N + \ln x} + (1-x)x - \frac{1}{\ln N}$$

We now solve f(x) = 0.

Note that

$$\begin{aligned} f\left(\frac{1}{\ln N}\right) &= \frac{(1/\ln N)^2}{\ln N + \ln\ln N} + \left(1 - \frac{1}{\ln N}\right)\frac{1}{\ln N} - \frac{1}{\ln N} \\ &= -\left(\frac{1}{\ln N}\right)^2 + (\frac{1}{\ln N})^3 \frac{\ln N}{\ln N + \ln\ln N} \\ &\approx -\left(\frac{1}{\ln N}\right)^2 \end{aligned}$$

Moreover,

$$f'(x) = \frac{2x(\ln N + \ln x) - x^2(\frac{1}{x})}{(\ln N + \ln x)^2} - 2x + 1$$

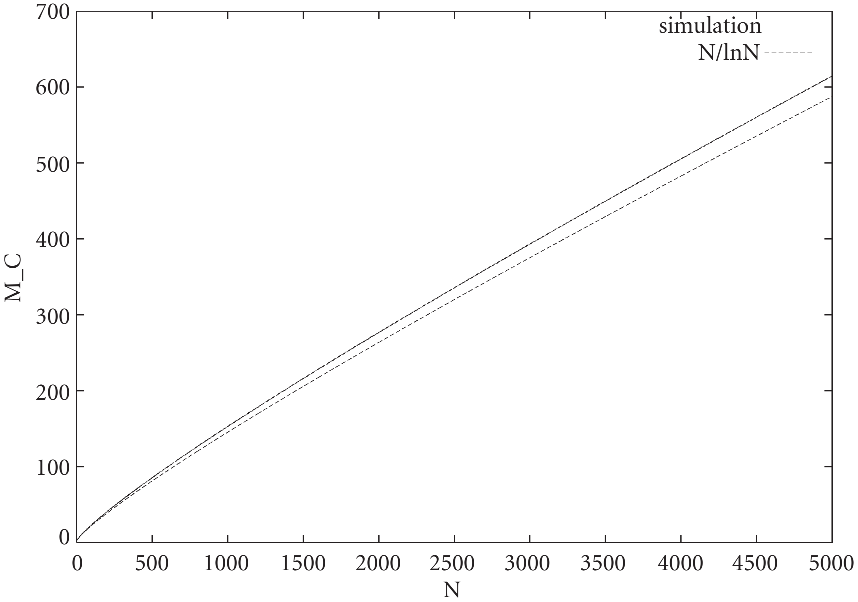
And

$$f'\left(\frac{1}{\ln N}\right) \approx 1 \text{ for } x = \frac{1}{\ln N}$$

Hence, $1/\ln N$ is a good approximation to the root of $f(x) = 0$

$$M_c \approx \frac{N}{\ln N}$$

The figure below shows the approximation of M_c for N = 1 to 5000, in comparison to the results obtained by numerical simulations: $N/\ln N$ is a very good approximation.



Remarks. In the calculation of $T(N, M)$, we made the assumption that words in $\mathcal{N}$ have uniform probabilities of being in the exception subset $\mathcal{M}$. This may strike one as unwarranted as it is often said that irregulars are among the most (token) frequent items for a given class (Bybee 1985; Pinker 1999). But there are a number of reasons for challenging this, which consequently preserves our assumption.

First, the tacit assumption under irregulars-are-frequent is that irregularity can only be sustained by high frequency, as held in the WAR model. A look at the English verb distributions certainly gives one that impression, and children do, in general, learn high-frequency irregulars faster and better. However, as noted in the text (Section 3.3), relatively low frequency verbs can be learned well if they piggyback on more frequent or systematic rules in the language. These verbs, then, will have good staying power as irregulars. Thus, high frequency is not the exclusive means of maintaining irregularity.

Second, the claim that irregulars are frequent does not always generalize to other cases. An obvious example is the English noun system where the irregulars are not as heavily concentrated in the high-frequency region. Another example is the German noun system discussed in the text. Under the WAR model, the -s class is computed by a rule while all other nouns–the vast majority in German–are irregulars and thus memorized by association. But here lies a contradiction. The irregulars must include many very low frequency nouns: if irregularity is maintained by high frequency, then these low frequency irregular nouns will be on a massive exodus to the default/regular -s class. But this has not happened in German. Again, irregularity can–and must, in this case–be maintained by means other than token frequency. One possibility is that these nouns are *not* irregular, but are handled by productive rules which applies to words regardless of their frequencies. We discuss this issue further in Section 5.3.

Finally, the current model is a procedure for the child learner to determine the productivity of rules. Children's acquisition of morphology, as reviewed in Section 2, shows that such decisions are almost always made correctly, and remarkably early. Hence, the determination of productivity via the Tolerance Principle is made on a relative small vocabulary, which contains mostly of high frequency words. Hence, even if the concentration of irregulars is mostly in the high frequency region of the *entire* lexicon, as in the case of English but not necessarily in all languages, it will become less pronounced in a young child's vocabulary, a sample of high frequency words that include both regulars and irregulars. See footnote 13 for additional discussions.

Notes

* Much of this paper derives from a 2002 manuscript (*A principle for word storage*) and has benefited from Morris Halle's suggestions and advice. Portions of this work were presented at University of Maryland, Massachusetts Institute of Technology, University of Pennsylvania University of Southern California, University of Illinois at Urbana-Champagne, and CUNY Graduate Center; I thank the audiences there, in particular Mark Liberman and Janet Dean Fodor, for questions and discussions. Bill Idsardi, Julie Anne Legate, Lisa Pearl and Sarah Van Wagenen made useful comments on a draft version of this paper though they'd be crazy to believe everything I say here. I thank Sam Gutmann for helping out with the mathematics. Finally, I am grateful to an anonymous reviewer whose suggestions improved this paper considerably.

[1] In the Optimality Theory literature, one may find similar notions such as morpheme-specific constraint rankings (Anttila 2000).

[2] There *could* be regularities among words that a morpholexical rule applies too: "ing→ang" applies to verbs that end in "ing" as in *sing* and *ring*: it just doesn't apply to all verbs that end in *ing*.

[3] Of these, "bring-brang" alone makes up almost half of over-irregularization errors. If dialectal variations are taken into account–*brang* is acceptable in some parts of the United States–the error rate would be even lower.

[4] In the actual implementation, words are represented as phonological feature bundles, and the change would be "add -t". We continue to use "add -d" for convenience of exposition.

[5] Logically, there needn't be any productive rules in language either. As far as I know, this situation does not arise in any language–such a language would not be very useful, as its generative capacities would be finite. However, it is perfectly possible for a child learner, at particular stage of language learning, to have no productive rules: consider a child that knows only one word.

[6] See Lewis et al. (2005) for a recent attempt that attributes a class of sentence processing phenomenon to an independently motivated theory of working memory (Anderson & Lebiere 1998).

[7] We do not deny additional contextual factors in lexical access, though the lexical search theory, which makes precise claims about specific psychological mechanisms, is a theory with strong predictive power. It has remained an important and fruitful research topic and produced findings that are difficult to capture under alternative models. For a cogent defense of this perspective, see Forster (1989) and Murray & Forster (2004).

[8] We leave open the possibility that the *M* exceptions are organized in some structural ways as well, and needn't be gone through exhaustively. See Footnote 17 for discussions with specific reference to an empirical case.

[9] This results in the application of some rule (other than *R*), if one follows the ROW model. On the other hand, the WAR model would say that a direct memory retrieval of the derived form follows the **THEN** statement. However, this detail needn't concern us in the present context.

10. Stem-cluster frequency is the sum of the frequencies of all forms of a word. For example, the stem-cluster frequency of the verb *arrive* is the sum of the frequencies of *arrive*, *arrives*, *arrived* and *arriving*. By contrast, the lexeme frequency refers to the frequency of a specific form of interest, e.g., *arrived*.

11. Jaeger et al. (1996) report that regulars are *much* faster than irregulars, but that result was obtained through the use of a flawed experimental design; see Seidenberg & Hoeffner (1998) for details.

12. Sereno & Jongman's framework has some similarities with the ROW model: both recognize, though in different theoretical terms, the frequency of *rules* in the morphological computation in addition to the frequency of words.

13. The reverse pattern appears to be true for low frequency verbs, where regulars are faster than irregulars. This is not the place for what might account for this curious phenomenon. It suffices for our purposes that the results on high frequency verbs may be all that the ECSS model needs. When children solve the productivity problem–and they solve it fairly early on, as noted in Section 2.1–they do so on the basis of a relatively small vocabulary, which in general consists of words in the high frequency region. As long as the rule search predictions hold for high frequency words, the ECSS model serves its purpose as the motivation for a principled decision procedure on productivity.

14. When they do inflect the verbs. Sometimes the verb may be left in the stem/infinitive form.

15. If I were right in Yang (2000), and if "add -s" is completely unrestricted, we have a situation that blatantly violates the Tolerance Principle. Simply, there are far too many exceptions for the "add -s" rule to be productive. However, German scholars have regarded the "add -s" class as a reservoir for "untypical" nouns (Wunderlich 1999), including proper names, onomatopoetics, truncated forms, loan words, and so on. In an informal Wug test conducted with the help of Marianne Pouplier and Tanja Kupisch, we found that when presented with nonce nouns that obeyed German phonotactic constraints, none of the German speakers we consulted formed plurals by adding -s, but used the -(e)n and -e suffixes instead, sometimes Umlauting as well. Our results suggest two conclusions. First, German speakers have a native sense of what counts as an "untypical" noun and reserve "-s" for its inflection: Marcus et al. (1995) were able to elicit "add -s" forms with foreign sounding nonce nouns. Second, it suggests that at least some "irregular" classes are in fact productive, as they apply to novel nouns. Unfortunately, we did not present the speakers with the gender of the nonce nouns to locate more precisely the connection between gender and pluralization.

16. There are non-feminine nouns that add -(e)n also, but those are obviously irrelevant to the rule 9 under consideration.

17. Note that this would be the most inclusive estimate of M, as we treat the set $\mathcal{M}$, i.e., feminine nouns that do not add -(e)n, to be completely unsystematic. In fact, it has been pointed that $\mathcal{M}$ may contain regularities within. Wunderlich (1999), for example, notes that umlaunt nouns regardless of gender, have certain phonological properties. Moreover, monosyllabic feminine nouns tend to add -e. These generalizations need to be put through tests to determine their productivity. If they are productive, the execution of 9 involves searching through a list of *rules*, each of which defines a set of nouns with certain properties (and may have to explicit list their own exceptions). The net effect of these productive sub-rules may shorten the rule search time for nouns that do follow 9. For instance, suppose "**IF** monosyllabic feminine **THEN** add -e" is productive. This sub-rule may speed up the rule selection process for a disyllabic feminine noun, because monosyllabic sub-rule needn't be searched through at all. In other words, if the sub-rules within $\mathcal{M}$ turn out to be productive, then we would in effect have a smaller number of M than we are currently estimating. These issues are quite complex and require further studies, and we have chosen the maximum value of M to carry out the productivity calculation of rule 9.

[18] I thank Sam Gutmann for his help in finding an analytical solution for $M_C \approx N/\ln N$.

[19] Similar algorithms with more sophisticated data structures may be used for a finer-grained model of morphological processing. For instance, it may be interesting to incorporate the Cohort effects (Marslen-Wilson 1987) in lexical lookup by the use of lexicographic tree search algorithms (Sleator & Tarjan 1985b). We are currently investigating this direction of research.

References

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

Anderson, S. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press.

Anderson, J. & Lebiere, C. 1998. *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Anderson, S. 1969. West Scandinavian vowel systems and the ordering of phonological rules. Ph.D. dissertation, MIT.

Anderson, S. 1974. *The organization of phonology*. New York: Academic Press.

Anderson, S. 1988. Morphology as a parsing problem. *Linguistics*, 26, 521–544.

Anttila, A. 2000. Morphologically conditioned phonological alternations. *Natural Language and Linguistic Theory*, 20, 1–42.

Aronoff, M. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Aronoff, M. 1976. *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

Baayen, H. 1989. A corpus-based approach to morphological productivity. Ph.D. thesis, Free University, Amsterdam.

Baayen, H. 2003. Probabilistic approaches to morphology. In Bod, R., Hay, J. & Jannedy, S. (Eds.) *Probabilistic Linguistics*. Cambridge, MA: MIT Press. 229–287.

Baayen, H. & Lieber, R. 1991. Productivity and English derivation: a corpus-based study. *Linguistics* 29:801–8433.

Balota, D. & Chumbley, J. 1984. Are lexical decisions a good measure of lexical access? The role of frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, 10:340–357

Barton, E., Berwick, R. & Ristad, E. 1987. *Computational complexity and natural langauge*. Cambridge, MA: MIT Press.

Bauer, L. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

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.

Berwick, R. & Weinberg, A. 1984. *Grammatical basis of linguistic performance*. Cambridge, MA: MIT Press.

Brent, M. & Cartwright, T. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61:93–125.

Brown, R. 1973. *A First Language*. Cambridge, Ma: Harvard University Press.

Burani, C., Salmaso, D., & Caramazza, A. 1984. Morphological structure and lexical access. *Visible Language*, 4:348–358.

Bybee, J. 1985. *Morphology: A study of form and meaning*. Amsterdam: Johns Benjamins.

Bybee, J. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes*, 10, 425–455.

Bybee, J. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Caramazza, A. 1997. How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, 14, 177–208.

Chomsky, N. 1955. *Logical Structure of Linguistic Theory*. Manuscript, Harvard and MIT. Published in 1975 by Plenum.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1980. *Rules and Representations*. New York: Columbia University Press.

Chomsky, N. & Halle, M. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Clahsen, H. 1999. Lexical entries and rules of language. *Brain and Behavioral Sciences*, 22:991–1013.

Clahsen, H. 2005. Dual-mechanism morphology. In Brown, K. (ed.) *Encyclopedia of Language and Linguistics*. New York: Oxford University Press. (To appear).

Clahsen, H., Eisenbeiss, S. & Sonnenstuhl, I. 1997. Morphological structure and the processing of inflected words. *Theoretical Linguistics*, 23, 201–249.

Clahsen, H., Hadler, M., & Weyerts, H. 2004. Speeded production of inflected words in children and adults. *Journal of Child Language*, 31:683–712.

Clashen, H., Marcus, G., Bartke, S., & Wiese, R. 1996. Compounding and inflection in German child language. *Yearbook of Morphology*, 115–142.

Clahsen, H. & Rothweiler, M. 1993. Inflectional Rules in Children's Grammars: Evidence from German Participles. In G. Booij & J.v. Marle (eds.) Yearbook of Morphology 1992. Dordrecht. 1–34.

Clahsen, H., Rothweiler, M., Woest, A. & Marcus, G. 1992. Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, 45, 225–255.

Clark, R. 2001. Information theory, complexity, and linguistic descriptions. In Bertolo, S. (Ed.) *Parametric Linguistics and Learnability*. Cambridge: Cambridge University Press. 126–171.

Crain, S. & Fodor, J. D. 1985. How can grammars help parsers. In Dowty, D., Kartuunen, L. & Zwicky, A. (Eds.) *Natural language parsing: Psychological, computational, and theoretical perspectives*. Cambridge: Cambridge University Press. 94–128.

Dabrowska, E. 2001. Learning a morphological system without a default: The Polish genitive. *Journal of Child Language*. 28, 545–574.

Dressler, W. 1999. Why collapse morphological concepts? *Behavioral and Brain Sciences*, 22, 1021.

Falco, J. & Yang, C. 2005. I want wash my baby feets. Ms. Yale University. New Haven, CT.

Forster, K. 1976. Accessing the mental lexicon. In Wales, R. & Walker, E. (Eds.) *New approaches to language mechanisms*. Amsterdam: North Holland. 257–287.

Forster, K. 1989. Basic issues in lexical processing. In Marslen-Wilson, W. (Ed.) *Lexical representation and process*. Cambridge, MA: MIT Press. 75–107.

Forster, K. 1992. Memory-addressing mechanisms and lexical access. In Frost, R. & Katz, K. (eds.) *Orthography, phonology, morphology, and meaning*. Amsterdam: Elsevier. 413–434.

Forster, K. & Chambers, S. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12, 627–635.

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

Guasti, M. 2002. *Language acquisition: The growth of grammar*. Cambridge, MA: MIT Press.

Halle, M. 1990. An approach to morphology. *Proceedings of the Northeast Linguistic Society*, 20:150–84.

Halle, M. 1997. The stress of English words 1968–1998. *Linguistic Inquiry*, 29, 539–568.

Halle, M., & Marantz, A. 1993. Distributed Morphology. In K. Hale & S. J. Keyser (eds.) *The View From Building 20*. Cambridge, MA: MIT Press, 111–176.

Hankamer, J. 1989. Lexical representation and processes. In Marslen-Wilson, W. (ed.) *Morphological Parsing and the Lexicon*. Cambridge, MA: MIT Press.

Howes, D. & Solomon, R. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental psychology*, 41:401–410.

Jaeger, J., Lockwood, A., Kemmerer, D., Van Valin, R., Murphy, B. & Khalak, H. 1996. A positron emission tomographic study of regular and irregular verb morphology in English. *Language* 72, 451–497.

Kam, C. & Newport, E. 2005. Regularizing unpredictable variations: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1:151–195.

Karttunen, L. & Beesley, K. 2003. *Finite state morphology*. Stanford, CA: CSLI Publications.

Kiparsky, P. 1973. "Elsewhere" in Phonology. In Anderson, S. & Kiparksy, P. (eds.) *A Festschrift for Morris Halle*. New York, NY: Holt, Rinehart & Winston. 93–106.

Kiparsky, P. 2000. Opacity and cyclicity. *Linguistic Review*, 17, 351–366.

Koskenniemi, K. 1983. Two-level morphology: A general computational model for word form recognition and production. Publication No. 11. Department of General Linguistics, University of Helsinki, Helsinki.

Levelt, W., Roelofs, A. & Meyer, A. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–37.

Lewis, R., Van Dyke, J., Vasishth, S, & Nakayama, M. 2005. Parsing as memory retrieval. *Journal of Memory and Language*. (To appear).

de Marcken, C. 1996. Unsupervised language acquisition. Ph.D. dissertation. MIT.

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.

Marslen-Wilson, W. 1987. Functional parallelism in spoken word recognition. *Cognition*, 25:71–102.

Miller, G. & Chomsky, N. 1963. Introduction to the Formal Analysis of Natural Languages. In *Handbook of Mathematical Psychology* II, Luce, D., Bush, R., & Galanter, E. (Eds.) New York: Wiley and Sons. 268–321.

Mitchell, T. 1982. Generalization as search. *Artificial Intelligence*, 18, 203–226.

Molnar, R. 2001. Generalize and Sift as a model of inflection acquisition. Master's thesis, MIT.

Mooney, R. & Califf, M. 1995. Induction of First-Order Decision Lists: Results on learning the past tense of English verbs. *Journal of Artificial Intelligence Research*, 3, 1–24.

Murray, W. & Forster, K. 2004. Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*, 111, 721–756.

Myers, S. 1987. Vowel shortening in English. *Natural Language and Linguistic Theory*, 5, 485–518.

Niemi, J., Laine, M. & Tuominen, J. 1994. Cognitive morphology in Finnish: Foundations of a new model. *Language and Cognitive Processes*, 9, 423–446.

Park, T.-Z. 1978. Plurals in child speech. *Journal of Child Language*, 5, 237–250.

Penke, M. & Krause, M. 2002. German noun plurals: A challenge to the dual-mechanism model. *Brain and Language*, 81:303–311.

Phillips, C. 1995. Syntax at Age 2: Crosslinguistic differences. MIT Working Papers in Lingustics. Cambridge, MA.

Pinker, S. 1999. *Words and Rules*. New York: Basic Books.

Pinker, S. & Ullman, M. 2002. The past and future of the past tense. *Trends in Cognitive Sciences*, 6:456–463.

Prasada, S. & Pinker, S. 1995. Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8, 1–56.

Prasada, S., Pinker, S., & Snyder, W. 1990. Some evidence that irregular forms are retrieved from memory but regular forms are rule generated. Poster presented at the Psychonomic Society Meeting, Nov. 1990. New Orleans.

Ristad, E. 1994. Complexity of morpheme acquisition. In Ristad, E. (Ed.) *Language Computation*. DIMACS.

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

Rubenstein, H., Garfield, L. & Milliken, J. 1970. Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, 9, 487–494.

Sapir, E. 1928. *Language: An introduction to the study of speech*. New York: Harcourt Brace.

Schütze, C. 2005. Thinking about what we are asking speakers to do. Manuscript, UCLA.

Seidenberg, M. 1992. Connectionism without tears. In Davis, S. (Ed.). *Connectionism: Theory and practice*. Oxford: Oxford University Press. 84–122.

Seidenberg, M. & Gonnerman, L. 2000. Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, 4:353–361.

Seidenberg, M. & Hoeffner, J. 1998. Evaluating behavioral and neuroimaging data on past tense processing. *Language*, 74, 104–122.

Sereno, J. & Jongman, A. 1997. Processing of English inflectional morphology. *Memory and Cognition*, 25, 425–437.

Sleator, D. & Tarjan, R. 1985a. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28:202–208.

Sleator, D. & Tarjan, R. 1985b. Self-adjusting binary search trees. *Journal of the ACM*, 32:652–686.

Sober, E. 1975. *Simplicity*. Oxford: Oxford University Press.

Snyder, W. 2001. On the nature of syntactic variation: Evidence from complex predicates and complex word-formation. *Language*, 77:324–342.

Sussman, G., & Yip, K. 1997. A Computational Model for the Acquisition and Use of Phonological Knowledge. MIT Artificial Intelligence Laboratory, Memo 1575

Taft, M. 1979. Lexical access via an orthographic code: The basic orthographic syllabic structure BOSS. *Journal of Verbal Learning and Verbal Behavior*, 18, 21–39.

Ullman, M. 1999. Acceptability ratings of regular and irregular past tense forms: Evidence for a dual-system model of language from word frequency and phonological neighborhood effects. *Language and Cognitive Processes*, 14, 47–67.

Veit, 1986. Das Verständnis von Plural- und Komparativformen bei (entwicklungs)-dysgrammatischen Kindern im Vorschulalter. In Kegel, G. et al. (Eds.) *Sprechwissenschaft und Psycholinguistik*. Opladen: Westdeutscher Verlag. 217–286.

Whaley, C. 1978. Ward-nonword classification time. *Journal of Verbal Learning and Verbal Behavior* 17, 143–154.

Wiese, R. 1996. *The phonology of German*. Cambridge: Cambridge University Press.

Wunderlich, D. 1999. German noun plural reconsidered. Manuscript. University of Düsseldorf. Germany.

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

Yang, C. 2000. Dig-dug, think-thunk. *London Review of Books*, 22.

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

Yang, C. 2004. Universal Grammar, statistics, or both. *Trends in Cognitive Sciences*, 8:451–456.

Yang, C. 2005. The origin of linguistic irregularity. In Wang, W. S.-Y. & Minett, J. (Eds.) *Language Acquisition, Change, and Evolution*. Hong Kong: City University of Hong Kong Press. 297–328.