# Who's Afraid of George Kingsley Zipf?

Charles Yang
Department of Linguistics & Computer Science
University of Pennsylvania
charles.yang@ling.upenn.edu

June 2010

## Abstract

We explore the implications of Zipf's law for the understanding of linguistic productivity. Focusing on language acquisition, we show that the item/usage based approach has not been supported by adequate statistical evidence. By contrast, the quantitative properties of a productive grammar can be precisely formulated, and are consistent with even very young children's language. Moreover, drawing from research in computational linguistics, the statistical properties of natural language strongly suggest that the theory of grammar be composed of general principles with overarching range of applications rather than a collection of item and construction specific expressions.

# 1 Introduction

Einstein was a very late talker. "The soup is too hot", as legend has it, were his first words, at the very ripe age of three. Apparently the boy genius hadn't had anything interesting enough to say.

The credulity of such tales aside—similar stories with other famous subjects abound—they do contain a kernel of truth: a child doesn't have to say something, *anything*, just because he can. And this poses a challenge for the study of child language when children's linguistic production is often the only, and certainly the most accessible, data on hand. Language use is the composite of linguistic, cognitive and perceptual factors many of which, in the child's case, are still in development and maturation. It is therefore difficult to draw inferences about the learner's linguistic knowledge from his linguistic behavior. Indeed, the moral holds for linguistic study in general: an individual's grammatical capacity may not be fully reflected in his or her speech. Since a goal of linguistic theory is to identify the possible and impossible structures of language, restricting oneself to naturalistic data is doubly limited: some expressions may not have been said but are nevertheless well formed, others—syntactic islands, for instance—will never be said for they are unsayable.

This much has been well known since Chomsky (1965) drew the competence/performance distinction. The pioneering work on child language that soon followed, include those who did not follow the generative approach, also recognized the gap between what the child knows and what the child says (Bloom 1970, Bowerman 1973, Brown & Fraser 1963, Brown & Bellugi 1964, McNeil 1966, Schlesinger 1971, Slobin 1971). Two examples from that period of time suffice to illustrate the necessity to go beyond the surface. Shipley, Gleitman & Smith (1969) show that children in the so-called telegraphic stage of language development nevertheless understand fully formed English sentences better than telegraphic patterns that resemble their own speech. Roger Brown, in his landmark study (1973) and synthesis of other work available at the time, provides distributional and quantitative evidence against the Pivot Grammar hypothesis (Braine 1963), under which early child syntax supposedly consists of templates centering around pivot words.[1] Brown advocates the thesis, later dubbed the Continuity Hypothesis, that child language be interpreted in terms of adult-like grammatical devices, which has continued to feature prominently in language acquisition (Wexler & Culicover 1980, Pinker 1984, Crain 1991, Yang 2002).

This tradition has been challenged by the *item* or *usage*-based approach to language most clearly represented by Tomasello (1992, 2000a, 2000b, 2003), which reflects a current trend (Bybee 2001, Pierrehumbert 2001, Goldberg 2003, Culicover & Jackendoff 2005, Hay & Baayen 2005, etc.) that emphasizes the storage of specific linguistic forms and constructions at the expense of general combinatorial linguistic principles and overarching points of language variation (Chomsky 1965, 1981). Child language, especially in the early stages, is claimed to consist of specific item-based schemas, rather than productive linguistic system as previously conceived. Consider, for instance, three case studies in Tomasello (2000a, p213-214) which have been cited as evidence for the item-based view at numerous places.

- The Verb Island Hypothesis (Tomasello 1992). In a longitudinal study of early child language, it is noted that "of the 162 verbs and predicate terms used, almost half were used in one and only one construction type, and over two-thirds were used in either one or two construction types ...". There is "great uneveness in how different verbs, even those that were very close in meaning, were used — both in terms of the number and types of constructions types used." Hence, "the 2-year-old child's

[1] A position that bears more than a passing resemblance to a strand of contemporary thinking to which we return momentarily.

syntactic competence is comprised totally of verb-specific constructions with open nominal slots", rather than abstract and productive syntactic rules under which presumably a broader range of combinations is expected.

- Limited morphological inflection. According to a study of child Italian (Pizutto & Caselli 1994), 47% of all verbs used by 3 young children (1;6 to 3;0) were used in 1 person-number agreement form, and an additional 40% were used with 2 or 3 forms, where six forms are possible (3 person $\times$ 2 number). Only 13% of all verbs appeared in 4 or more forms. Again, the low level of usage diversity is taken to show the limitedness of generalization characteristic of item-based learning.

- Unbalanced determiner usage. Citing Pine & Lieven (1997) and other similar studies, it is found that when children began to use the determiners *a* and *the* with nouns, "there was almost no overlap in the sets of nouns used with the two determiners, suggesting that the children at this age did not have any kind of abstract category of Determiners that included both of these lexical items". This finding is held to contradict the earliest study (Valian 1986) which maintains that child determiner use is productive and accurate like adults by the age of 2;0.

So far as we can tell, however, these evidence in support for item-based learning has been presented, and accepted, on the basis of intuitive inspections rather than formal empirical tests. For instance, among the numerous examples from child language, no statistical test was given in the major treatment (Tomasello 1992) where the Verb Island Hypothesis and related ideas about item-based learning are put forward. Specifically, no test has been given to show that the observations above are statistically *inconsistent* with the expectation of a fully productive grammar, the position that item-based learning opposes. Nor, for that matter, are these observations shown to be *consistent* with item-based learning, which, as we shall see, has not been clearly enough articulated to facilitate quantitative evaluation. In this paper, we provide statistical analysis to fill these gaps. We demonstrate that children's language use actually shows the opposite of the item-based view; the productivity of children's grammar is in fact confirmed. More broadly, we aim to direct researchers to certain statistical properties of natural language and the challenges they pose for the theory of language and language learning. Our point of departure is a name that has been, and will continue to, torment every student of language: *George Kingsley Zipf*.

## 2 Zipfian Presence

### 2.1 Zipfian Words

Under the so-called *Zipf's law* (Zipf 1949), the empirical distributions of words follow a curious pattern: relatively few words are used frequently—*very* frequently—while most words occur rarely, with many occurring only once in even large samples of texts. More precisely, the frequency of a word tends to be approximately inversely proportional to its rank in frequency. Let $f$ be the frequency of the word with the rank of $r$ in a set of $N$ words, then:

$$f = \frac{C}{r} \text{ where } C \text{ is some constant} \tag{1}$$

In the Brown corpus (Kucera & Francis 1967), for instance, the word with rank 1 is "the", which has the frequency of about 70,000, and the word with rank 2 is "of", with the frequency of about 36,000: almost exactly as Zipf's law entails (i.e., $70000 \times 1 \approx 36000 \times 2$). The Zipfian characterization of word frequency

can be visualized by plotting the log of word frequency against the log of word rank. By taking the log on both sides of the equation above ($\log f = \log C - \log r$), a perfect Zipfian fit would be a straight line with the slope -1. Indeed, Zipf's law has been observed in vocabulary studies across languages and genres, and the log-log slope fit is consistently in the close neighborhood of -1.0 (Baroni 2008).

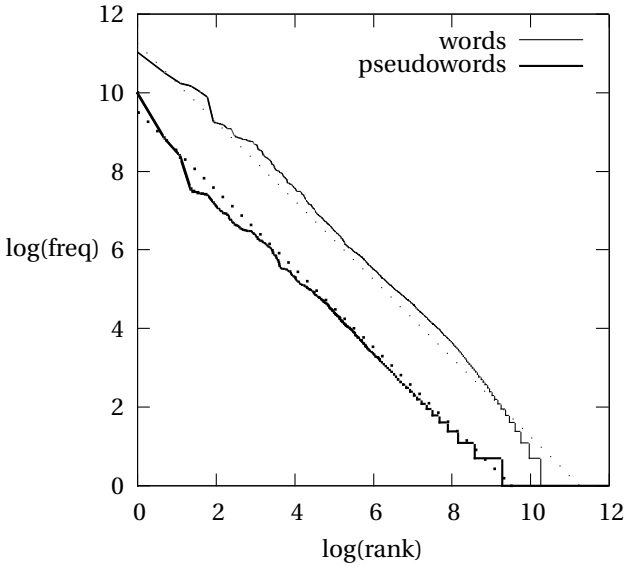


Figure 1. Zipfian distribution of words (top) and pseudowords (bottom) in the Brown corpus. The lower line is plotted by taking "words" to be any sequence of letters between *e*'s (Chomsky 1958). The two straight dotted lines are linear functions with the slope -1, which illustrate the goodness of the Zipfian fit.

There has been a good deal of controversy over the interpretation of Zipf's law, which shows up not only in the context of words but also many other physical and social systems (Bak, Tang, Wiesenfeld 1987, Gabaix 1999, Axtell 2001, among many others). It is now clear that the observation of Zipfian distribution alone is of no inherent interest or significance, as certain random letter generating processes can produce outcomes that follow Zipf's law (Mandelbrot 1954, Miller 1957, Li 1992, Niyogi & Berwick1995). As noted in Chomsky (1958), if we redefine "words" as alphabets between any two occurrences of some letter, say, "e", rather than space as in the case of written text, the resulting distribution may fit Zipf's law even better. This is illustrated by the lower line in Figure 1, which follows the Zipfian straight line at least as well as real words.

It is often the case that we are not concerned with the actual frequencies of words but their probability of occurrence; Zipf's law makes this estimation simple and accurate. Given (1), the probability $p_r$ of the word $n_r$ with the rank $r$ among $N$ words can be expressed as follows:

$$p_r = \left(\frac{C}{r}\right) \Big/ \left(\sum_{i=1}^{N} \frac{C}{i}\right) = \frac{1}{rH_N} \text{ where } H_N \text{ is the } N\text{th Harmonic Number } \sum_{i=1}^{N} \frac{1}{i} \tag{2}$$

The application of Zipf's law to words has been very well studied. Yet relatively little attention has been given to the combinatorics of linguistic units under a grammar and more important, how one might draw inference about the grammar given the distribution of word combinatorics. We turn to these questions immediately.

## 2.2 Zipfian Combinatorics

The "long tail" of Zipf's law, which is occupied by low frequency words, becomes even more pronounced when we consider combinatorial linguistic units. Take, for instance, $n$-grams, the simplest linguistic combination that consists of $n$ consecutive words in a text.[2] Since there are a lot more bigrams and trigrams than words, there are consequently a lot more low frequency bigrams and trigrams in a linguistic sample, as Figure 2 illustrates from the Brown corpus (for related studies, see Teahan 1997, Ha et al. 2002):

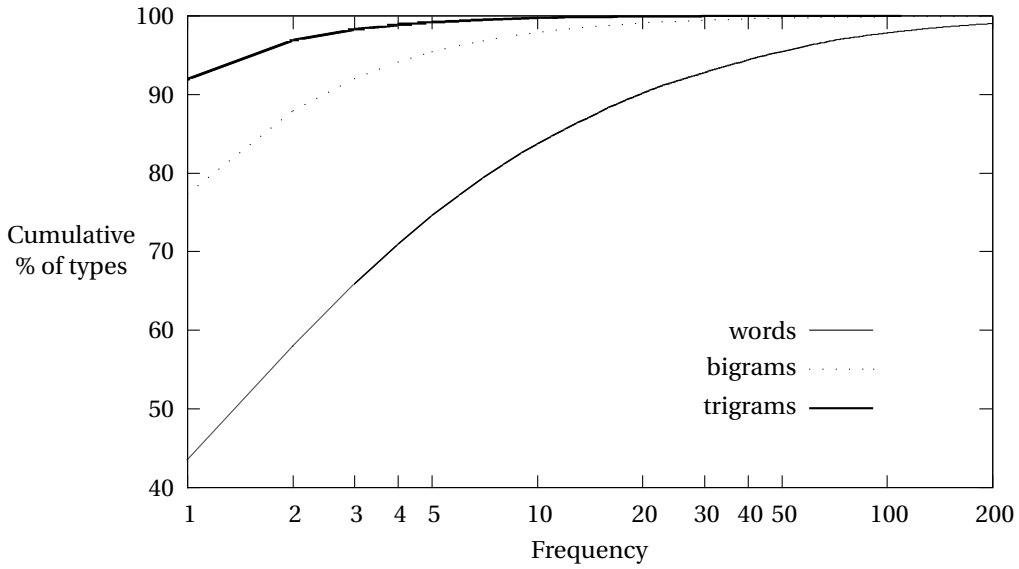


Figure 2. The vast majority of $n$-grams are rare events. The x-axis denotes the frequency of the gram, and the y-axis denotes the cumulative % of the gram that appear at that frequency or lower.

For instance, there are about 43% of words that occur only once, about 58% of words that occur 1-2 times, 68% of words that occur 1-3 times, etc. The % of units that occur multiple times decreases rapidly, especially for bigrams and trigrams: approximately 91% of distinct trigram types in the Brown corpus occur only once, and 96% occur once or twice.

The range of linguistic forms is so vast that no sample is large enough to capture all of its varieties even when we make a certain number of abstractions. Figure 3 plots the rank and frequency distributions of syntactic rules of modern English from the Penn Treebank (Marcus et al. 1993). Since the corpus has been manually annotated with syntactic structures, it is straightforward to extract rules and tally their frequencies.[3] The most frequent rule is "PP→P NP", followed by "S→NP VP": again, the Zipf-like pattern can be seen by the close approximation by a straight line on the log-log scale.

[2]For example, given the sentence "the cat chases the mouse", the bigrams ($n = 2$) are "the cat chases the mouse" are "the cat", "cat chases", "chases the", and "the mouse", and the trigrams ($n = 3$) are "the cat chases", "cat chases the", "chases the mouse". When $n = 1$, we are just dealing with words.

[3]Certain rules have been collapsed together as the Treebank frequently annotates rules involving distinct functional heads as separate rules.

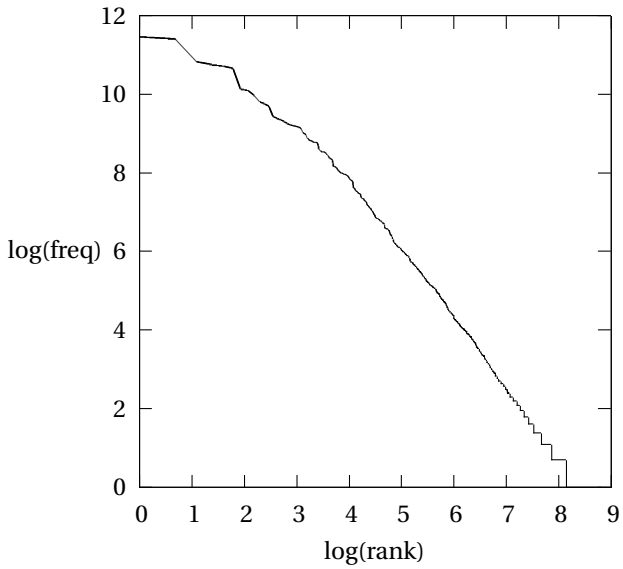


Figure 3. The frequency distribution of the syntactic rules in the Penn Treebank.

The long tail of linguistic combinations must be taken into account when we assess the structural properties of the grammar. Claims of item-based learning build on the premise that linguistic productivity entails diversity of usage: the "unevenness" in usage distribution is taken to be evidence against a systematic grammar. The underlying intuition, therefore, appears to be that linguistic combinations might follow something close to a uniform distribution. Take the notion of overlap in the case of determiner use in early child language (Pine & Lieven 1997). If the child has fully productive use of the syntactic category determiner, then one might expect her to use determiners with any noun for which they are appropriate. Since the determiners "the" and "a" have (virtually) identical syntactic distributions, a linguistically productive child that uses "a" with a noun is expected to automatically transfer the use of that noun to "the". Quantitatively, determiner-noun overlap is defined as the percentage of nouns that appears with both determiners out of those that appear with either. The low overlap values in children's determiner use (Pine & Lieven 1997, among others) are taken to support the item-based view of child language.

However, using a similar metric, Valian and colleagues (Valian et al. 2009) find that the overlap measures for young children and their mothers are not significantly different, and they are both very low. In fact, when applied to the Brown corpus (see section 3.2 for methods), we find that "a/the" overlap for singular nouns is only 25.2%: almost three quarters that could have appeared with both determiners only appeared with one exclusively. The overlap value of 25.2% is actually lower than those of some children reported in Pine & Lieven (1997). It would follow that the language of the Brown corpus, which draws from various genres of professional print materials, is less productive and more item-based than that of a toddler—which seems absurd.

The reason for these seemingly paradoxical findings lies in the Zipfian distribution of syntactic categories and the generative capacity of natural language grammar. Consider a fully productive rule that combines a determiner and a singular noun, or "DP→ D N", where "D→ a|the" and "N→ cat|book|desk|...". We use this rule for its simplicity and for the readily available data for empirical tests but one can easily substitute the rule for "VP→ V DP", "VP→ V in Construction$_x$", "$V_{inflection} \rightarrow V_{stem}$ + Person + Number + Tense". All such cases can be analyzed with the methods provided here.

Suppose a linguistic sample contains $S$ determiner-noun pairs, which consist of $D$ and $N$ unique determiners and nouns. (In the present case $D = 2$ for "a" and "the".) The full productivity of the DP rule, by definition, means that the two categories combine independently. Two observations, one obvious and the other novel, can be made about the distributions of the two categories and their combinations. First, nouns (and open class words in general) will follow zipf's law. For instance, the singular nouns that appear in the form of "DP→ D N" in the Brown corpus show a log-log slope of -0.97. In the CHILDES (MacWhinney 2000) speech transcripts of six children (see section 3.2 for details), the average value of log-log slope is -0.98. This means that in a linguistic sample, relatively few nouns occur often but many will occur only once—which of course cannot overlap with more than one determiners.

Second, while the combination of $D$ and $N$ is syntactically interchangeable, $N$'s tend to favor one of the two determiners, a consequence of pragmatics and indeed non-linguistic factors. For instance, we say "the bathroom" more often than "a bathroom" but "a bath" more often than "the bath", even though all four DPs are perfectly grammatical. The reason for such asymmetries is not a matter of linguistic interest: "the bathroom" is more frequent than "a bathroom" only because bodily functions are a more constant theme of life than real estate matters.

We can place these combinatorial asymmetries in a quantitative context. As noted earlier, about 75% of distinct nouns in the Brown corpus occur with exclusively "the" or "a" but not both. Even the remaining 25% which do occur with tend to have favorites: only a further 25% (i.e. 12.5% of all nouns) are used with "a" and "the" equally frequently, and the remaining 75% are unbalanced. Overall, for nouns that appear with both determiners as least once (i.e. 25% of all nouns), the frequency ratio between the more over the less favored determiner is 2.86:1. (Of course, some nouns favor "the" while others favor "a", as the "bathroom" and "bath" examples above illustrate.) These general patterns hold for child and adult speech data as well. In the six children's transcripts (section 3.2), the average percentage of balanced nouns among those that appear with both "the" and "a" is 22.8%, and the more favored vs. less favored determiner has an average frequency ratio of 2.54:1. Even though these ratios deviate from the perfect 2:1 ratio under the strict version of Zipf's law—the more favored is even more dominant over the less—they clearly point out the considerable asymmetry in category combination usage. As a result, even when a noun appears several times in a sample, there is still a significant chance that it has been paired with a single determiner in all instances.

Together, Zipfian distributions of atomic linguistic units (words; Figure 1) and their combinations ($n$-grams Figure 2, phrases Figure 3) ensure that the determiner-noun overlap *must* be relatively low unless the sample size $S$ is very large. In section 4, we examine, and discover similar patterns, for the usage patterns of verbal syntax and morphology. For the moment, we develop a precise mathematical treatment and contrast it with the item-based learning approach in the context of language acquisition.

# 3 Quantifying Productivity

## 3.1 Theoretical analysis

Consider a sample $(N, D, S)$, which consists of $N$ unique nouns, $D$ unique determiners, and $S$ determiner-noun pairs. Here $D = 2$ for "the" and "a" though we consider the general case here. The nouns that have appeared with more than one (i.e. two) determiners will have an overlap value of 1; otherwise, they have the overlap value of 0. The overlap value for the entire sample will be the number of 1's divided by $N$.

Our analysis calculates the *expected value* of the overlap value for the sample $(N, D, S)$ under the

productive rule "DP→D N"; let it be $O(N,D,S)$. This requires the calculation of the expected overlap value for each of the $N$ nouns over all possible compositions of the sample. Consider the noun $n_r$ with the rank $r$ out of $N$. Following equation (2), it has the Zipfian probability $p_r = 1/(rH_N)$ of being drawn at any single trial in $S$. Let the expected overlap value of $n_r$ be $O(r,N,D,S)$. The overlap for the sample can be stated as:

$$O(D,N,S) = \frac{1}{N}\sum_{r=1}^{N} O(r,N,D,S) \tag{3}$$

Consider now the calculation $O(r,N,D,S)$. Since $n_r$ has the overlap value of 1 if and only if it has been used with more than one determiner in the sample, we have:

$$\begin{aligned} O(r,N,D,S) &= 1 - \Pr\{n_r \text{ is not sampled during } S \text{ trials}\} \\ &\quad - \sum_{i=1}^{D} \Pr\{n_r \text{ is sampled but with the } i\text{th determiner exclusively}\} \\ &= 1-(1-p_r)^S \\ &\quad - \sum_{i=1}^{D}\Big[(d_i p_r + 1 - p_r)^S - (1-p_r)^S\Big] \end{aligned}$$

The last term above requires a brief comment. Under the hypothesis that the language learner has a productive rule "DP→D N", the combination of determiner and noun is independent. Therefore, the probability of noun $n_r$ combining with the $i$th determiner is the product of their probabilities, or $d_i p_r$. The multinomial expression

$$(p_1 + p_2 + ... + p_{r-1} + d_i p_r + p_{r+1} + ... + p_N)^S$$

gives the probabilities of all the compositions of the sample, with $n_r$ combining with the $i$th determiner 0, 1, 2, ... $S$ times, which is simply $(d_i p_r + 1 - p_r)^S$ since $(p_1 + p_2 + p_{r-1} + p_r + p_{r+1} + ... + p_N) = 1$. However, this value includes the probability of $n_r$ combining with the $i$th determiner zero times—again $(1-p_r)^S$—which must be subtracted. Thus, the probability with which $n_r$ combines with the $i$th determiner exclusively in the sample $S$ is $[(d_i p_r + 1 - p_r)^S - (1-p_r)^S]$. Summing these values over all determiners and collecting terms, we have:

$$O(r,N,D,S) = 1 + (D-1)(1-p_r)^S - \sum_{i=1}^{D}\Big[(d_i p_r + 1 - p_r)^S\Big] \tag{4}$$

The formulations in (3)—(4) allow us to calculate the expected value of overlap using only the sample size $S$, the number of unique noun $N$ and the number of unique determiners $D$, under the assumption that nouns and determiners both follow Zipf's law as discussed in section 2.[4] Figure 4 gives an illustration, with $N = 100$, $D = 2$ and $S = 200$.

[4]For the present case involving only two determiners "the" and "a", $d_1 = 2/3$ and $d_2 = 1/3$. As noted in section 2.2, the empirical probabilities of the more vs. less frequent determiners deviate somewhat from the strict Zipfian ratio of 2:1, numerical results show that the 2:1 ratio is a very accurate surrogate for a wide range of actual rations in the calculation of (3)—(4). This is because most of average overlap value comes from the relatively few and high frequent nouns, as Figure 4 makes clear.

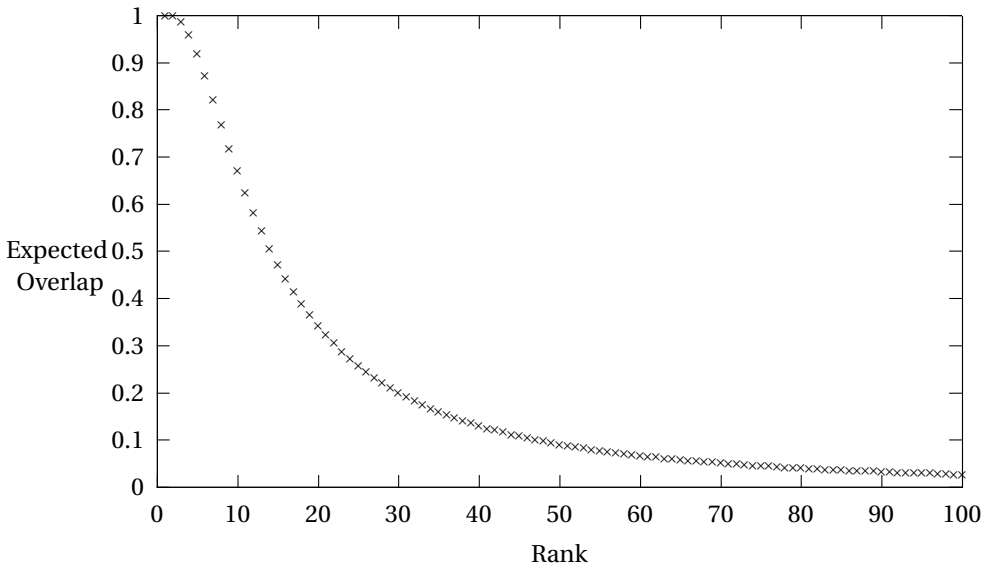


Figure 4. Expected overlap values for nouns ordered by rank, for $N = 100$ nouns in a sample size of $S = 200$ with $D = 2$ determiners. Word frequencies are assumed to follow the Zipfian distribution. As can be seen, few of nouns have high probabilities of occurring with both determiners, but most are (far) below chance. The average overlap is 21.1%.

Under Zipfian distribution of categories and their productive combinations, low overlap values are a mathematical necessity. As we shall see, the theoretical formulation here nearly perfectly match the distributional patterns in child language, to which we turn presently.

### 3.2 Determiners and productivity

**Methods**. To study the determiner system in child language, we consider the data from six children Adam, Eve, Sarah, Naomi, Nina, and Peter. These are the all and only children in the CHILDES database (MacWhinney 2000) with substantial longitudinal data that starts at the very beginning of syntactic development (i.e, one or two word stage) so that the item-based stage, if exists, could be observed. For comparison, we also consider the overlap measure of the Brown corpus (Kucera & Francis 1967), for which productivity is not in doubt.

We first removed the extraneous annotations from the child text and then applied an open source implementation of a rule-based part-of-speech tagger (Brill 1995):[5] words are now associated with their part-of-speech (e.g., preposition, singular noun, past tense verb, etc.). For languages such as English, which has relatively salient cues for part-of-speech (e.g., rigid word order, low degree of morphological syncretism), such taggers can achieve high accuracy at over 97%. This already low error rate causes even less concern for our study, since the determiners "a" and "the" are not ambiguous and are always correctly tagged, which reliably contributes to the tagging of the words that follow them. The Brown Corpus is available with manually assigned part-of-speech tags so no computational tagging is necessary.

With tagged datasets, we extracted adjacent determiner-noun pairs for which $D$ is either "a" or "the", and $N$ has been tagged as a singular noun. Words that are marked as unknown, largely unintelligible

[5]Available at http://gposttl.sourceforge.net/.

transcriptions with special marks, are discarded. As is standard in child language research, consecutive repetitions count only once toward the tally. For instance, when the child says "I made a queen. I made a queen. I made a queen", "a queen" is counted once for the sample.

In an additional test suggested by Virginia Valian (personal communication), we pooled together the first 100, 300, and 500 determiner-noun tokens of the six children and created three hypothetical children from the very earliest stages of language acquisition, which would presumably be the least productive knowledge of determiner usage.

For each child, the theoretical expectation of overlap is calculated based on equations in (3)—(4), that is, only with the sample size $S$ and the number of unique nouns $N$ in determiner-noun pairs while $D = 2$. These expectations are then compared against the empirical overlap values computed from the determiner-noun samples extracted with the methods above. The results are summarized in Table 1.

| Subject | Sample Size ($S$) | *a* or *the* Noun types ($N$) | Overlap (expected) | Overlap (empirical) | $\frac{S}{\overline{\mathrm{N}}}$ |
|---|---|---|---|---|---|
| Naomi (1;1-5;1) | 884 | 349 | 21.8 | 19.8 | 2.53 |
| Eve (1;6-2;3) | 831 | 283 | 25.4 | 21.6 | 2.94 |
| Sarah (2;3-5;1) | 2453 | 640 | 28.8 | 29.2 | 3.83 |
| Adam (2;3-4;10) | 3729 | 780 | 33.7 | 32.3 | 4.78 |
| Peter (1;4-2;10) | 2873 | 480 | 42.2 | 40.4 | 5.99 |
| Nina (1;11-3;11) | 4542 | 660 | 45.1 | 46.7 | 6.88 |
| First 100 | 600 | 243 | 22.4 | 21.8 | 2.47 |
| First 300 | 1800 | 483 | 29.1 | 29.1 | 3.73 |
| First 500 | 3000 | 640 | 33.9 | 34.2 | 4.68 |
| Brown corpus | 20650 | 4664 | 26.5 | 25.2 | 4.43 |

Table 1. Empirical and expected determiner-noun overlaps in child speech. The Brown corpus is included in the last row for comparison. Results include the data from six individual children and the first 100, 300, 500 determiner-noun pairs from all children pooled together, which reflect the earliest stages of language acquisition. The expected values in column 5 are calculated using only the sample size $S$ and the number of nouns $N$ (column 2 and 4 respectively), following the analytic results in section 3.1.

The theoretical expectations and the empirical measures of overlap agree extremely well (column 5 and 6 in Table 1). Neither paired t- nor Wilcoxon test reveal significant difference between the two sets of values. Perhaps a more revealing test is linear regression (Figure 5): a perfect agreement between the two sets of value would have the slope of 1.0, and the actual slope is 1.08 (adjusted $R^2 = 0.9716$).

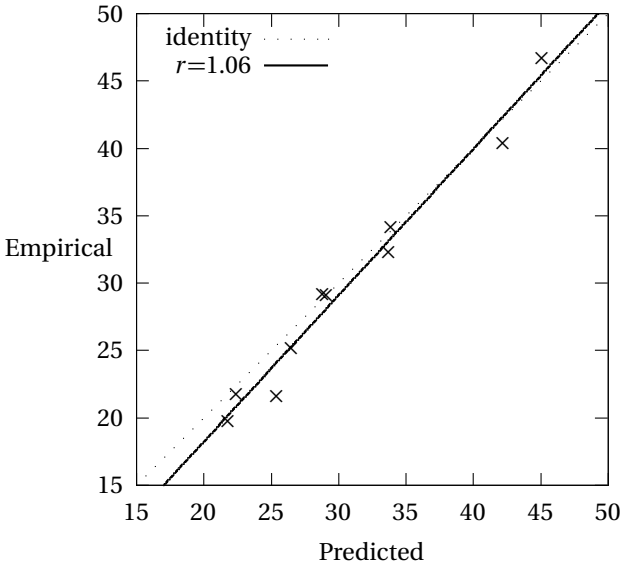


Figure 5. The solid line represents the linear regression fit of the expected vs. empirical values of overlap in Table 1 column 5 and 6 ($r$=1.08, adjusted $R^2 = 0.9716$). The dotted line indicates a perfect fit (i.e., the identity function $y = x$).

Therefore, we could that the determiner usage data from child language is consistent with the productive rule "DP→ D N".

The empirical studies also reveal considerable individual variation in the overlap values, and it is instructive to understand why. As the Brown corpus result shows (Table 1 last row), sample size $S$, the number of nouns $N$, or the language user's age alone is not predictive of the overlap value. The variation can be roughly analyzed as follows; see Valian et al. (2009) for a related proposal. Given $N$ unique nouns in a sample of $S$, greater overlap value can be obtained if more nouns occur more than once. That is, words whose probabilities are greater than $1/S$ can increase the overlap value. Zipf's law (2) allows us to express this cutoff line in terms with ranks, as the probability of the noun $n_r$ with rank $r$ has the probability of $1/(rH_N)$. The derivation below uses the fact that the $N$th Harmonic Number $\sum_{i=1}^{N} 1/i$ can be approximated by $\ln N$.

$$S\frac{1}{rH_N} = 1$$
$$r = \frac{S}{H_N} \approx \frac{S}{\ln N} \tag{5}$$

That is, only nouns whose ranks are lower than $S/(\ln N)$ can be expected to be non-zero overlaps. The total overlap is thus a monotonically increasing function of $S/(N \ln N)$ which, given the slow growth of $\ln N$, is approximately $S/N$, a term that must be positively correlated with overlap measures. This result is confirmed in strongest terms: $S/N$ is a near perfect predictor for the empirical values of overlap (last two columns of Table 1): $r = 0.986$, $p < 0.00001$.

### 3.3 Evaluating item-based learning

The study in section 3.2 shows that the evidence for item-based learning has been taken at the face value. The alternative hypothesis of productivity, when formulated precisely, matches the acquisition data nearly perfectly.

We then turn to the question whether the determiner usage data by children can be accounted for equally well by the item based approach. In the limiting case, the item-based child learner could store the input data in its entirety and simply retrieve these memorized determiner-noun pairs in production. Since the input data, which comes from adults, is presumably productive, children's repetition of it may show the same degree of productivity.

Unfortunately, our effort to investigate this possibility is hampered by the lack of concrete models for the item-based learning approach, a point that its advocates concede. Tomasello, for instance, acknowledges that "[t]he more emic approaches of Cognitive Linguistics and development psychology ... may appear, and indeed may be, less rigorously specifiable than generative approaches, a disadvantage to some theorists, perhaps" (1992, p274). Explicit models in item-based learning and similar approaches (e.g., Chang et al. 2005, Freudenthal et al. 2007, 2009, among others) generally involve programming efforts for which no analytical results such as those in section 3.1 are possible. Moreover, some of the most extensively studied models (e.g., Freudenthal et al. 2007 and subsequent work) are explicitly not intended as learning models but as means of data exploration (Pine 2009). Thus, it is not clear what kind of predictions these models make.

A plausible approach can be construed based on a central tenet of item-based learning, that the child does not form grammatical generalizations but rather memorizes and retrieves specific and itemized combinations (Tomasello 2001, 2003). Similar approaches such as construction grammar (Goldberg 2003), usage and exemplar based models (Bybee 2001, Pierrehumbert 2001) make similar commitment to the role of verbatim memory. In a clear statement, Tomasello (2000c, p77) suggests that "(w)hen young children have something they want to say, they sometimes have a set expression readily available and so they simply retrieve that expression from their stored linguistic experience." Following this line of reasoning, we consider a learning model that memorizes *jointly* formed, as opposed to productively composed, determiner-noun pairs from the input; presumably these "stored linguistic experience" as such nouns (and determiners) constitute a large part of adult-child linguistic communication in everyday life. These pairs will then be sampled directly; for each sample, the overlap values can be calculated and compared against the empirical values in Table 1.

We consider two variants of the memory model. The first can be called a *global memory* learner. Using the computational methods outlined in section 3.2, we extracted all determiner-noun combinations from an approximately 1.1 million random sample of English adult utterances (about 6.5 million words) from the CHILDES database. The global memory learner is a hypothetical child who memorizes all such combinations in the input along with their frequencies,[6] which is an attempt to replicate the long-term effect of learning and storage. The second model is a *local memory* learner, which is construed to capture the linguistic experience of the particular child (i.e, the six children plus the three composite learners from the earliest stages of acquisition). The local memory learner only memorizes the determiner-noun pairs from the adult utterances in that particular child's CHILDES transcripts.

For each child, then, there are two sets of data: the determiner-noun pairs along with their frequen-

[6]We put aside the important question how a learner can selectively retain determiner-noun pairs as relevant items for memorization, rather than other adjacent category/word combinations such as auxiliary-determiner ("is a") or pronoun-verb ("He put") which are highly frequent but do not appear systematically in children's speech.

cies from that child's input (local memory learner) and the determiner-noun pairs along with their frequencies in the entire 1.1 million utterances of adult speech (global memory learner). For each child with a sample size of $S$ (see Table 1, column 2), and for each variant of the memory model, we use the Monte Carlo simulation to randomly draw $S$ pairs from the two sets of data that correspond to the local and global memory learning models. The probability with which a pair is drawn is proportional to its frequency in the two sets of data. Thus, a more frequently-used pairs in the input will have a higher chance of being drawn, which reflects frequency effects in learnings often emphasized in the item/usage-based approach (e.g., Tomasello 2001, 2003, Matthews et al. 2005, Bybee & Hopper 2001, among others). Each sample, then, consists of a list of determiner-noun pairs with varying occurrence counts. We calculate the value of overlap from this list, that is, the percentage of nouns that appear with both "a" and "the" over the total number of nouns. The results are averaged over 1000 draws. These results are given in Table 2.

| Child | Sample Size ($S$) | Overlap (global memory) | Overlap (local memory) | Overlap (empirical) |
|---|---|---|---|---|
| Eve | 831 | 16.0 | 17.8 | 21.6 |
| Naomi | 884 | 16.6 | 18.9 | 19.8 |
| Sarah | 2453 | 24.5 | 27.0 | 29.2 |
| Peter | 2873 | 25.6 | 28.8 | 40.4 |
| Adam | 3729 | 27.5 | 28.5 | 32.3 |
| Nina | 4542 | 28.6 | 41.1 | 46.7 |
| First 100 | 600 | 13.7 | 17.2 | 21.8 |
| First 300 | 1800 | 22.1 | 25.6 | 29.1 |
| First 500 | 3000 | 25.9 | 30.2 | 34.2 |

Table 2. The comparison of determiner-noun overlap between two variants of item-based learning and empirical results.

Both sets of overlap values from the two variants of item-based learning (column 3 and 4) differ significantly from the empirical measures (column 5): $p < 0.005$ for both paired t-test and paired Wilcoxon test. This suggests that children's use of determiners does not follow the predictions of the item-based learning approach; it certainly does not seem to be the result of the child retrieving jointly stored determiner-noun pairs from the input in a frequency sensitive fashion. Naturally, our evaluation here is tentative since the proper test can be carried out only when the theoretical predictions of item-based learning are made clear. And that is exactly the point: the advocates of item-based learning not only rejected the alternative hypothesis without adequate statistical tests, but also accepted the favored hypothesis without adequate statistical tests.

## 4 An Itemized Look at Verbs

The formal analysis in section 3 can be generalized to child verb syntax and morphology, which are among the main supporting cases for item-based learning. Unfortunately, the acquisition data in support of the Verb Island Hypothesis (Tomasello 1992) and the item-based nature of early morphology (Pizutto & Caselli 1994) cited in section 1 has not been made available in the public domain.

But the Zipfian reality is inherent: the combinatorics of verbs and their morphological and syntactic associates are similarly lopsided in usage distribution as is with the determiners. We now turn to examine the statistical distributions of verbal morphology and syntax.

### 4.1 The sparsity of verbal morphology

The statistical properties of morphology have been investigated by Chan (2008) in an independent context, and again Zipf's law reigns supreme. Few stems appear in a great number of inflections, which, however, never approach anywhere near the maximum number of possible inflections. Moreover, most stems are used very sparsely, the majority of which occur in exactly one inflection. In other words, there are languages in which one could go through his entire life without ever hearing the full content of a paradigm table favored by linguists—not even for a single stem. Furthermore, the inflections themselves are also Zipfian: few are used very frequently but most are used sparsely. These findings pose interesting challenges to both the traditional Word and Paradigm approach to morphology as well as recent work that emphasizes stored exemplars.

The reader is directed to Chan (2008) for details for these important observations and how the learning model might cope with such sparsity of morphological data. Our focus here is to provide a brief assessment of the statistical distribution of morphological forms in child and adult languages. Recall the Italian morphology acquisition study (Pizutto & Caselli 1994) where most verbs appear only one or two of the six possible agreement forms. Table 3 summarizes the results from the corpus analysis of all of child and child-directed data in Italian, Spanish, and Catalan that are currently available in the CHILDES database. The morphological data is analyzed with a state of the art natural language processing toolkit Freeling, which specializes in Romance languages.[7] Only tensed forms are counted; infinitives, which do not bear person/number agreement in these languages, are ignored. Each cell represents the percentage of verb stems that are used in 1, 2, 3, 4, 5, and 6 inflectional forms.

| Subjects | 1 form | 2 forms | 3 forms | 4 forms | 5 forms | 6 forms | $S/N$ |
|---|---|---|---|---|---|---|---|
| Italian children | 81.8 | 7.7 | 4.0 | 2.5 | 1.7 | 0.3 | 1.533 |
| Italian adults | 63.9 | 11.0 | 7.3 | 5.5 | 3.6 | 2.3 | 2.544 |
| Spanish children | 80.1 | 5.8 | 3.9 | 3.2 | 3.0 | 1.9 | 2.233 |
| Spanish adults | 76.6 | 5.8 | 4.6 | 3.6 | 3.3 | 3.2 | 2.607 |
| Catalan children | 69.2 | 8.1 | 7.6 | 4.6 | 3.8 | 2.0 | 2.098 |
| Catalan adults | 72.5 | 7.0 | 3.9 | 4.6 | 4.9 | 3.3 | 2.342 |

Table 3. Verb agreement distributions in child and adult Italian, Spanish, and Catalan. The last column reports the ratio between the total number of inflected forms ($S$) over the total number of stems ($N$), which is the average number of opportunities for a stem to be used.

A formal treatment of the agreement distributions similar to the overlap study uses multinomial analysis which we do not pursue here. Nevertheless, the logic of the problem remains the same as in equation (5): the diversity of usage depends on the number of opportunities for a verb stem to appear multiple forms, or $S/N$. As can be seen in Table 3, children learning Spanish and Catalan show very similar agreement usage to adults—and the $S/N$ ratios are also very similar for these groups. Italian children use somewhat more stems in only one form than Italian adults (81.8% vs. 63.9%), but that follows from the $S/N$ ratio (2.544 vs. 1.533). That is, for each verb, the Italian adults have roughly 66% more opportunities to use it than the Italian children, which would account for the discrepancy in the frequency of one-form verbs.

[7] Available at http://garraf.epsevg.upc.es/freeling/.

### 4.2 All verbs are islands

We now study the distributional properties of verbal syntax that have been attributed to the Verb Island Hypothesis. We focus on constructions that involve a transitive verb and its nominal objects, including pronouns and noun phrases. Following the definition of "sentence frame" in Tomasello's original Verb Island study (1992, p242), each unique lexical item in the object position counts as a unique construction for the verb.

Figure 6 shows the construction frequencies of the top 15 transitive verbs in 1.1 million child directed utterances. Processing methods are as described in section 3.2 except here we extract adjacent verb-nominal pairs in part-of-speech tagged texts. For each verb, we count its top 10 most frequent constructions, which are defined as the verb followed a unique lexical item in the object position (e.g., "ask him" and "ask John" are different constructions.) For each of the 10 ranks, we tallied the construction frequencies for all 15 verbs.[8]

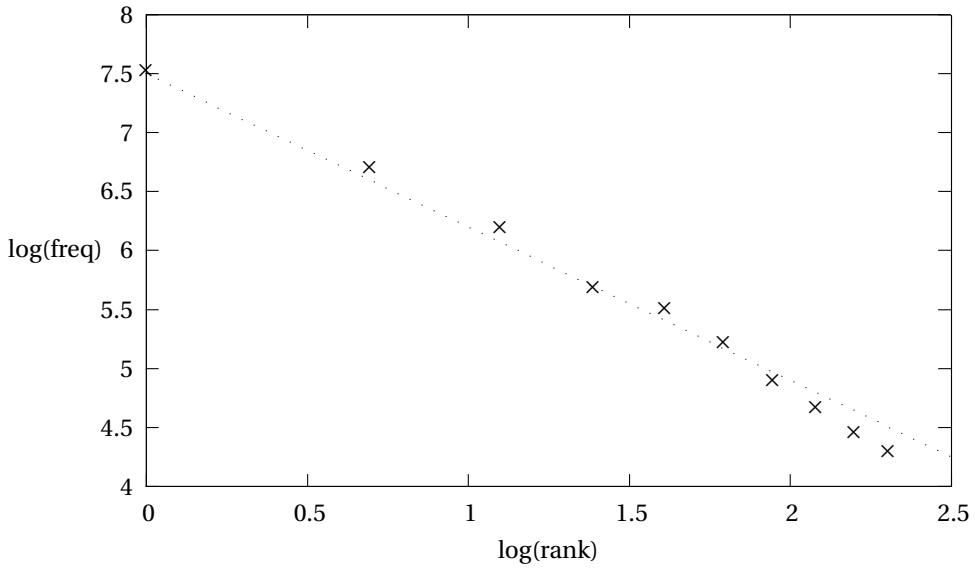


Figure 6. Rank and frequency of verb-object constructions based on 1.1 million child-directed utterances.

The verb construction frequency thus also follow a Zipf-like pattern: even for large corpora, a verb appears in few constructions frequently and in most constructions infrequently if at all. The observation of Verb Islands, that verbs tend to combine with one or few elements out of a large range, is in fact characteristic of a fully productive verbal syntax system.

As far as we know, the quantitative predictions of the Verb Island Hypothesis have not been spelled out but we may estimate the necessary amount of language sample that would mask these island effects. The appeal to unevenness of verbal construction frequencies seems to reflect the expectation that under full productivity, most verbs ought to appear with most of the possible range of arguments. Substituting nouns and determiners for verbs and nominals, the formal analysis could be carried out for the verbal syntactic system. Instead of calculating the expected numbers of determiners that a noun appears with, one would calculate the expected number of objects a verb appears with.

[8]These verbs are: *put, tell, see, want, let, give, take, show, got, ask, make eat, like, bring* and *hear*. The frequency tallies of the top 10 most frequent constructions are 1904, 838, 501, 301, 252, 189, 137, 109, 88, and 75.

Here we carry out a Monte Carlo simulation to determine the range of sample size for which the observed unevenness will disappear or considerably diminish. Let us suppose that, fairly conservatively, the child knows only 100 verbs, which can combine with only 100 distinct objects, following the definition of the sentence frame in Tomasello (1992). There are thus 10,000 unique verb-object combinations. For a given sample size, we draw verbs and objects independently according to their Zipfian frequencies. We then calculate the average number of objects that a verb appears with. We increase the sample size until the average number of objects per verb is 50% of all possible objects.

For 100 verbs and 100 objects, the sample size needs to consist of approximately 28,000 verb-object pairs in order to meet the 50% requirement. Since our 1.1 million adult utterances, or 6.5 million words, only yielded about 19,000 verb-object pairs (with approximately 850 distinct verbs and objects respectively), it is reasonable to conclude that a sample of 9.6 million words, probably the first few years of speech altogether, would be required to meet the 50% criterion, for a very modest verb and object vocabulary size. Note that even when the sample size is this large, there is still a good deal of remaining unevenness: after all, the number of objects per verb is mostly due to the highly frequent verbs that have appeared with virtually all objects, and many verbs, even in a sample of almost 10 million words, will have appeared with far less than 50% of the objects.

The necessary sample size to eliminate the verb island effects goes up sharply when we increase the number of verbs and objects. Under a still conservative estimate of 1500 verbs and 1500 objects for a language user, to achieve the 50%-object criterion requires an estimated 4.8 billion words, which amount to 46 years of non-stop talking at a rate of 200 words per minute. In light of those considerations, when the sample size is only extremely modest, as is the case in Tomasello (1992) and indeed most child production studies (Tomasello & Stahl 2004), it is impossible to spot anything other than verb islands.

## 5 Conclusions

So who's afraid of George Kingsley Zipf? The answer must be, *everyone*.

For the *psychologist* of child language, our study demonstrates the necessity to take the Zipfian nature of language into account when assessing linguistic knowledge. For any type of linguistic expression that involve open class items—and that means *every* type of linguistic expression—modest measures of usage diversity requires extremely large samples. This may not be possible in principle for the study of young children's language, even those not nearly as reticent as baby Einstein. Moreover, we note that many psycholinguistic studies of language rely on the differentiation from a null hypothesis, some baseline measure which is often provided by intuitive expectations (e.g., "chance level performance"). Our study shows that all hypotheses can and should be formulated precisely, including the null. When possible, one should strive to test for quantitative *matches* between hypothesis and data, rather than mismatches between competing hypotheses. We have just presented a case where the null hypothesis, once rigorously formulated, not only cannot be rejected but is be confirmed. Finally, while the current study shows that children's production is consistent with a productive grammar, it should not distract us from the important question how the child learns that grammar in the first place.

The *computer scientist* has been well aware of Zipf's challenges, which come in the form of the *sparse data problem*. As statistical models of language grow more sophisticated, the number of parameters that must be empirically valued shoots up exponentially. Hence one rapidly runs out of available data to estimate these parameters—thanks to Zipf's law—even when the statistical models of language are very simple, and drastic simplifying assumptions are made about the independence of linguistic structures

(Jelinek 1998); the $n$-gram and rule distributions discussed in section 2.2 make these points very clearly.

For the *linguist*, the Zipfian nature of language raises important questions for the development of linguistic theories. First, Zipf's law hints at the inherent limitations in approaches that stress the storage of construction-specific rules or processes (e.g., Goldberg 2003, Culicover & Jackendoff 2005). For instance, the central tenets of Construction Grammar views constructions as "stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns" and "the totality of our knowledge of language is captured by a network of constructions" (Goldberg 2003, p219). Yet the Zipfian distribution of linguistic combinations, as illustrated in Figure 3 for the Wall Street Journal and Figure 4 for child directed speech, ensure that most "pairings of form and function" simply will never be heard, never mind stored, and those that do appear may do so with sufficiently low frequency such that no reliable storage and use is possible.

Second, and more generally, Zipf's law challenges the conventional wisdom in current syntactic theorizing that makes use of a highly detailed lexical component; there have suggestions that all matters of language variation are in the lexicon which in any case needs to be acquired for individual languages. Yet the effectiveness of lexicalization in grammar has not been fully investigated in large scale studies. However, useful inferences can be drawn from the research on statistical induction of grammar in computational linguistics (Charniak 1993, Collins 2003). These tasks typically take a large set of grammatical rules (e.g., probabilistic context free grammar) and find appropriate parameter values (e.g., expansion probabilities in a probabilistic context free grammar) on the basis of an annotated training data such as the Treebank where sentences have been manually parsed into phrase structures. The performance of the trained grammar is evaluated by measuring parsing accuracy on a new set of unanalyzed sentences, thereby obtaining some measure of generalization power of the grammar.

Obviously, inducing a grammar on a computer is hardly the same thing as constructing a theory of grammar by the linguist. Nevertheless, statistical grammar induction can be viewed as a tool that explores what type of grammatical information is in principle available in and attainable from the data, which in turn can guide the linguist in making theoretical decisions. Contemporary work on statistical grammar induction makes use of wide range of potentially useful linguistic information in the grammar formalism. For instance, an phrase "drink water" may be represented in multiple forms:

(a) $\mathrm{VP} \rightarrow \mathrm{V\ NP}$

(b) $\mathrm{VP} \rightarrow \mathrm{V}_{\mathrm{drink}}\ \mathrm{NP}$

(c) $\mathrm{VP} \rightarrow \mathrm{V}_{\mathrm{drink}}\ \mathrm{NP}_{\mathrm{water}}$

(a) is the most general type of context free grammar rule, whereas both (b) and (c) include additional lexical information: (b) provides a lexically specific expansion rule concerning the head verb "drink", and the bilexical rule in (c) encodes the item-specific pairing of "drink" and "water", which corresponds to the notion of sentence frame in Tomasello's Verb Island hypothesis (1992; see section 4.2).

By including or excluding the rules of the types above in the grammatical formalism, and evaluating parsing accuracy of the grammar thus trained, we can obtain some quantitative measure of how much each type of rules, from general to specific, contributes to the grammar's ability to generalize to novel data. Bikel (2004) provides the most comprehensive study of this nature. Bilexical rules (c), similar to the notion of sentence frames and constructions, turn out to provide virtually no gain over simpler models that only use rules of the type (a) and (b). Furthermore, lexicalized rules (b) offer only modest improvement over general categorical rules (a) alone, with which almost all of the grammar's generalization power lies. These findings are not surprising given the Zipfian nature of linguistic productivity:

item-specific combinations are useful to keep track of only if they recur in the data but that is highly unlikely for most combinations.

The most significant victim of George Kingsley Zipf, however, must be the *child* learner herself. The task faced by children acquiring language is no different from that of the psychologist, computer scientist, and linguist, for the input data are also Zipfian in character, except that the child does not have controlled experiments, fast computer chips, or annotated corpora by her side. The sparse data problem strikes just as hard, and the role of memory in language learning should not be overestimated. In linguistics and cognitive science, of course, the learner's Zipfian challenge bears another name: the argument from the poverty of stimulus (Chomsky 1975; see Legate & Yang 2002 for a quantitative treatment). To attain full linguistic competence, the child learner must overcome the Zipfian distribution and draw generalizations about language on the basis of few and narrow types of linguistic expressions. In the face of such statistical reality of language, a grammatical system with full generative potentials (Chomsky 1965, Brown 1973) from the get go still seems the best preparation a child can hope for.

## References


Axtell, R. L. (2001). Zipf Distribution of U.S. Firm Sizes. *Science*, 293, 1818-1820.

Bak, P, Tang, C, & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of the 1/f noise. *Physical Review Letters*. 59, 381-384.

Baroni, M. (2008). Distributions in text. In Lüdelign, A. & Kytö, M. (Eds.) *Corpus linguistics: An international hanbook*. Berlin: Mouton de Gruyter.

Bikel, D. (2004) Intricacies of Collins' parsing model. *Computational Linguistics*, 30, 479–511.

Bloom, L. (1970). *Language development: Form and function in emerging grammars*. Cambridge MA: MIT Press.

Bowerman, M. (1973). *Early syntactic development: A cross-linguistic study with special reference to Finnish*. Cambridge: Cambridge University Press.

Braine, M. (1963). The ontogeny of English phrase structure: The first phase. *Language*, 39, 3-13.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21 (4), 543–565.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Brown, R. & Fraser, Collin. (1963). The acquisition of syntax. In Cofer, Charles & Musgrave, Barbara (Eds.) *Verbal behavior and learning: Problems and processes*. New York: McGraw-Hill. 158–201.

Brown, R. & Bellugi, U. (1964). Three processes in the acquisition of syntax. *Harvard Educational Review*, 34, 133-151.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. & Hopper, P. (2001). P. *Frequency and emergence of linguistic structure*. Amsterdam: Johns Benjamins.

Chan, E. (2008). Structures and distributions in morphology learning. Ph.D. Dissertation. Department of Computer and Information Science. University of Pennsylvania. Philadelphia, PA.

Chang, F., Lieven, E., & Tomasello, M. (2006). Using child utterances to evaluate syntax acquisition algorithms. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver, Canada

Chomsky, N. (1958). Review of *Langage des machines et langage humain* by Par Vitold Belevitch. *Language*, 34 (1), 99-105.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrectht: Foris.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*. 14, 597-650.

Culicover, P. & Jackendoff, R. (2005). *Simpler syntax*. New York: Oxford University Press.

Freudenthal, D., Pine, J. M., Aguado-Orea, J. & Gobet, F. (2007). Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. *Cognitive Science*, 31, 311-341.

Freudenthal, D., Pine, J. M. & Gobet, F. (2009). Simulating the referential properties of Dutch, German and English root infinitives. *Language Learning and Development*, 5, 1-29.

Gabaix, X. (1999). Zipf 's Law for Cities: An Explanation. *The Quarterly Journal of Economics*. 114, 739-767.

Goldberg, E. (2003). Constructions. *Trends in Cognitive Science*, 7, 219–224.

Ha, Le Quan, Sicilia-Garcia, E. I., Ming, Ji. & Smith, F. J. (2002). Extension of Zipf's law to words and phrases. *Proceedings of the 19th International Conference on Computational Linguistics*. 315-320.

Hay, J. & Baayen, H. (2005). Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences*, 9, 342-348.

Jelinek, F. (1998). *Statistical methods for speech recognition*. Cambridge, MA: MIT Press.

Kučera, H & Francis, N. (1967). *Computational analysis of present-day English*. Providence, RI: Brown University Press.

Legate, J. A. & Yang, C. (2002). Empirical reassessments of poverty stimulus arguments. *Linguistic Review*, 19, 151-162.

Li, W. (1992). Random texts exhibit Zipf's law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38 (6), 1842-1845.

MacWhinney, B. (2000). *The CHILDES Project*. Lawrence Erlbaum.

Mandelbrot, B. (1954). Structure formelle des textes et communication: Deux études. *Words*, 10, 1–27.

Matthews, D., Lieven, E., Theakston, A. & Tomasello, M. (2005). The role of frequency in the acquisition of English word order. *Cognitive Development*, 20, 121-136.

McNeill, D. (1963). The creation of language by children. In Lyons, J. & Wales, Roger. (Eds.) *Psycholinguistic Papers*. Edinburgh: Edinburgh University Press. 99-132.

Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, 70, 2, 311-314.

Niyogi, P. & Berwick, R. (1995). A note on Zipf's law, natural language, and noncoding DNA regions. Artificial Intelligence Laboratory Memo No. 1530. Massachusetts Institute of Technology. Cambridge, MA.

Pierrehumbert, J. (2001). Exemplar dynamics. In Bybee, J. & Hopper, P. (Eds.) *Frequency and emergence of linguistic structure*. Amsterdam: Johns Benjamins. 137–158.

Pine, J. (2009). Simulating the developmental pattern of finiteness marking in English, Dutch, German, French and Spanish using MOSAIC. Paper presented at the Workshop on Input and Syntactic Acquisition. University of California, Irvine.

Pine, J. & Lieven, E. (1997). Slot and frame patterns in the development of the determiner category. *Applied Psycholinguistics*, 18, 123-138.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pizutto, E. & Caselli, C. (1994). The acquisition of Italian verb morphology in a cross-linguistic perspective. In Levy, Y. (Ed.) *Other children, other languages*. Hillsdale, NJ: Erlbaum.

Schlesinger, I. M. (1971). Production of utterances and language acquisition. In In Slobin, Dan (Ed.) *The Ontogenesis of grammar*. New York: Academic Press. 63-101.

Shipley, E., Smith, C. & Gleitman, L. (1969). A study in the acquisition of language: Free reponses to commands. *Language*, 45, 2: 322-342.

Slobin, Dan. (1971). Data for the symposium. In Slobin, Dan (Ed.) *The Ontogenesis of grammar*. New York: Academic Press. 3-14.

Teahan, W. J. (1997). Modeling English text. DPhil thesis. University of Waikato, New Zealand.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2000a). Do young children have adult syntactic competence. *Cognition*, 74, 209-253.

Tomasello, M. (2000b). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4, 4: 156-164.

Tomasello, M. (2000c). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11, 61-82.

Tomasello, M. (2003). *Constructing a language*. Cambridge, MA: Harvard University Press.

Tomasello, M. & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough. *Journal of Child Language*, 31, 101-121.

Marcus, M., Marcinkiewicz, M. & Santorini, B. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19, 313-330.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, 22, 562-579.

Valian, V., Solt, S. & Stewart, J. (2008). Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, 35, 1-36.

Wexler, K. & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

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

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley.