

Who's Afraid of George Kingsley Zipf?

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

June 2009

Abstract

We explore the statistical distributions of natural language known as Zipf's law and develop a new approach to assess the properties of the underlying grammar given a sample of linguistic production. Focusing on language acquisition, we show that the item or usage based approach to language and language learning fails to provide adequate statistical tests of linguistic productivity, and that even very young children's grammar is abstract, systematic, and fully generative.

1 Introduction

Einstein was a very late talker. As legend has it, the first thing the young Einstein ever uttered, at the age of three, was “The soup is too hot”. Apparently the boy genius had nothing interesting to say before that.

The credibility 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 the process of development and maturation; and it is difficult to draw inferences about the learner's linguistic knowledge from his linguistic behavior. This much has been well appreciated ever since Chomsky [1] drew the competence/performance distinction. The pioneering work on child language that soon followed, most notably Roger Brown's landmark study [2], recognized such potential gaps between what the child knows and what the child says and proposed to interpret child language in terms of adult-like grammatical devices. (See, in particular, Brown's

*For helpful comments, I would like to thank Virginia Valian, Julie Anne Legate, Sam Gutmann, Bob Frank, Mark Liberman, Qiuye Zhao, Ruochuan Liu and the audience at the 2009 Schultink Lecture, the University of Groningen, where these materials were first presented. Special thanks to Erwin Chan and Costas Lignos for their help with Table 3 and Figure 4.

critique of the Pivot Grammar hypothesis [3], bears more than a passing resemblance with some contemporary theorizing of child language reviewed here.)

This tradition is now challenged by the *item* or *usage*-based approach to language [4–7]. This change reflects a current trend in the study of language, as can be seen in these pages [8–12]), which emphasizes the storage of specific linguistic forms and constructions at the expense of general combinatorial linguistic principles and overarching points of language variation [1, 13]. 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 Box 1 [5], which have been cited as evidence for the item-based view at numerous places.

Box 1: Production Evidence for Item Based Approach to Language Learning

- The Verb Island Hypothesis [4]. 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 ...”. Hence, “the 2-year-old child’s syntactic competence is comprised totally of verb-specific constructions with open nominal slots”, rather than abstract and productive syntactic rules.
- Limited morphological inflection. A study of child Italian [14] finds that 47% of all verbs used by 3 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 × 2 number). Only 13% of all verbs appeared in 4 or more forms.
- Unbalanced determiner usage. [15] notes 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 observation is held to contradict the earliest study [16] which maintains that child determiner use is adult-like by the age of 2;0.

So far as we can tell, however, these purported evidence for item-based learning is given entirely on the basis of informal inspections rather than rigorous empirical tests. For all examples and observations from child language, not a single statistical test can be found in Tomasello’s work [4] where the Verb Island Hypothesis and related ideas about item-based learning are first put forward. In general, theoretical conclusions are only convincing when the null or alternative hypothesis can be statistically rejected; that is, the observation in Box 1 be shown to be inconsistent with the expectation from a fully productive grammar. In this note, we provide statistical analysis of what such alternative hypothesis would be. We demonstrate that children’s language use shows exactly the *opposite* of the item-based view, and the hypothesis of early productivity is

in fact supported. More broadly, we direct cognitive scientists to certain statistical properties of natural language that are widely known but not widely appreciated, and to discuss the challenges their properties pose for the theory of language and language learning. Our point of departure is a name that ought to strike fear in every living soul: *George Kingsley Zipf*.

2 Zipfian Presence

Under the so-called *Zipf's law* [17], the distributions of words follow a curious pattern: relatively few words are used frequently — *very* frequently — while most words are rare, 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 w with the rank of r in a set of N , then:

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

In the Brown corpus [18], 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. 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 ($\log f = \text{constant} - \log r$), 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 [19]. The top line in Figure 1 plots word rank and frequency on a log-log scale based on the Brown corpus: the Zipfian fit is excellent.

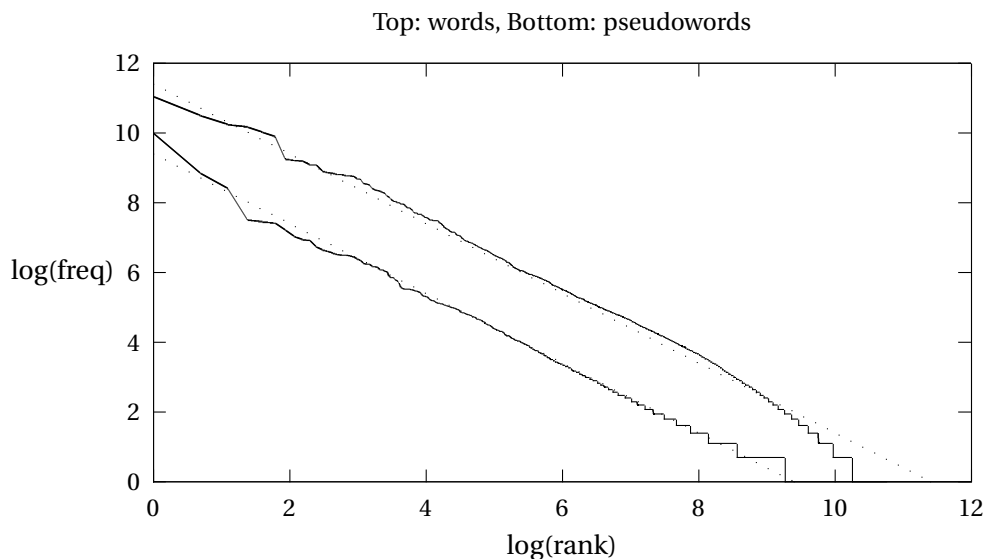


Figure 1. Zipfian distribution of words and pseudowords in the Brown corpus [18]. The lower line is plotted by taking “words” to be any sequence of letters between *e*'s [23]. The two straight dotted lines are linear functions with the slope -1, which illustrate the goodness of the Zipfian fit.

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 other physical and social systems. It is now clear that the observation of Zipfian distribution alone is of no inherent interest or significance, as certain random generating processes can produce outcomes that follow Zipf's law [20–22]. As noted in [23], 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 particularly concerned with the actual frequencies of words but their probability of occurrence. Zipf's law allow us to estimate the probability p of the word n , whose rank is r among N words in a linguistic sample:

$$p = \frac{\frac{C}{r}}{\sum_{i=1}^N \frac{C}{i}} = \frac{1}{r H_N} \text{ where } H_N \text{ is the } N\text{th Harmonic Number } \sum_{i=1}^N \frac{1}{i} \quad (1)$$

Zipf's law as applied to the distribution of words has been well known and studied. Yet relatively little attention has been given to the combinatorics of words 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.

3 The Unbearable Lightness of Productivity

Claims of item-based learning are established on the assumption that linguistic productivity entails usage diversity in linguistic production. Take the notion of “overlap” in the case of determiner use in early child language (Box 1), follows the logic of the Verb Island hypothesis [4]. 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”. Thus, determiner-noun overlap is defined as the percentage of nouns that appears with both determiners out of those that appear with either. The low overlaps in children's determiner use [15] are taken to support the item-based view of child language. However, using a similar metric, Valian and colleagues [24] find that the overlap measures for young children and their mothers are not significantly different, and they are both very low. Indeed, when applied to the Brown corpus (see Box 3 for methods), we obtain an overlap 25.2%, which is actually lower than those of some children reported in [15]. 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 “ $DP \rightarrow D N$ ”, where “ $D \rightarrow a|the$ ” and “ $N \rightarrow cat|book|desk|...$ ”. We use this rule for its simplicity and for the readily available data for empirical tests but one can easily substitute the rule for “ $VP \rightarrow V DP$ ”, “ $VP \rightarrow V$ 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 distinct 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 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: nouns in the Brown corpus, for instance, shows a log-log slope of -0.97 (see Box 3 for methods). This means that in any given sample, relatively few nouns occur often but most can be expected to occur only once or less than once, which necessarily have zero overlap with 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 linguistic pragmatics and conventions. 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. As noted earlier, about 75% of nouns in the Brown corpus occur with either “the” or “a” but not both. Even the remaining 25% which do occur with both show strong biases: only a further 25% (297) are used with “a” and “the” equally frequently. Overall, for nouns that appear with both determiners as least once, the frequency ratio between the more over the less favored determiner is 2.86:1. These general patterns hold for child and adult speech data as well. In the six child-adult pairs (thus 12 samples) we examined from the CHILDES database (Box 3), 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, these consequences of Zipf’s law ensure that the average determiner-noun overlap must be relatively low unless the sample size S is very large. Box 2 gives the theoretical analysis of overlap informally sketched above, and Figure 2 gives an illustration.

Box 2. Calculating Expected Overlap in Determiner Noun Usage

Let $O(N, S)$ be the overlap value of N nouns in a sample S pairs of determiner-noun pairs. Consider a noun n which has the rank of r out of N . Following equation (1), it has a probability of $p = 1/(rH_N)$ of being drawn at any single trial in S . Let the expected overlap of n be $O(r, N, S)$.

$$O(N, S) = \frac{1}{N} \sum_{r=1}^N O(r, N, S) \tag{2}$$

$$O(r, N, S) = 1 - \sum_{i=1}^D [(pd_i + (1-p))^S - (1-p)^S] - (1-p)^S \tag{3}$$

The probability of determiners $d_i (i = 1, 2)$ in (3) also follows Zipf's law (1), i.e., $d_1 = 2/3$ and $d_2 = 1/3$.^a

^aAlthough the empirical frequencies of 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 ratios in the calculation of (2) and (3). This is because most of average overlap value comes from the relatively few and high frequent nouns; see Figure 2.

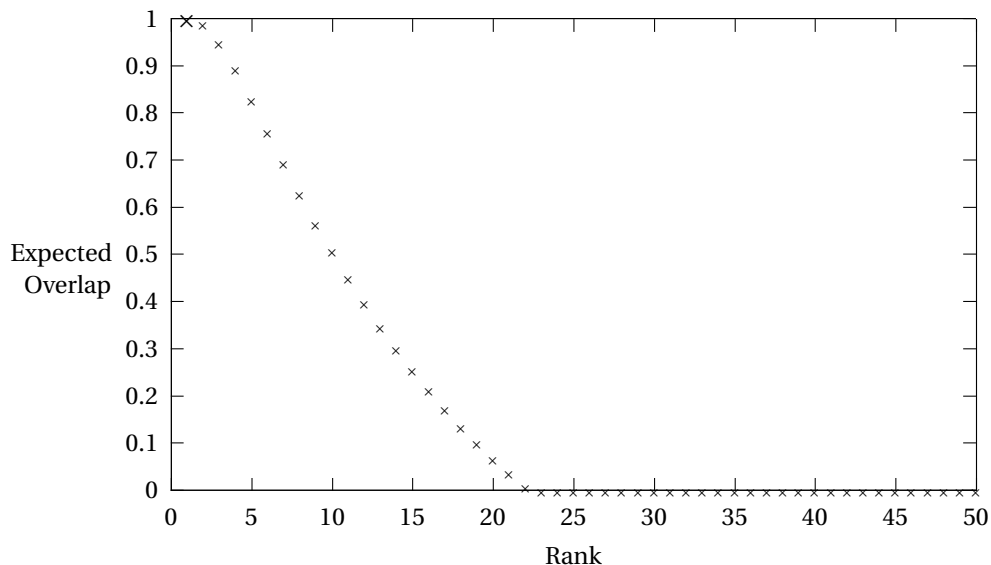


Figure 2. Expected overlap values for nouns ordered by rank, for $N = 50$ nouns in a sample size of $S = 100$. Word frequencies are assumed to follow the Zipfian distribution. As can be seen, few of nouns have high probabilities of occurring with both determiners, but most are (far) below chance. The average overlap is 20.6%.

We now study determiner-noun overlap in child language and compare the results with the

theoretical expectations. The empirical methods for data analysis are given in Box 3, and the results are summarized in Table 1.

Box 3. Empirical Studies of Overlap in Language Production

- a. We consider the data from Adam, Eve, Sarah, Naomi, Nina, and Peter [25]. These are the all and only children in the CHILDES database 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.
- b. 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 [26] (available <http://gposttl.sourceforge.net/>): 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%, which is sufficient for our purposes. For comparison, we also consider the Brown corpus [18], which has been manually tagged.
- c. With POS tagged datasets, we extracted adjacent determiner-noun pairs such as *D* is either “a” or “the”, and *N* has been tagged as a singular noun. Words that are marked as unknown as discarded. As is standard in child language research, repetitions counts 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.
- d. For an additional test, we have pooled together the first 100, 300, and 500 determiner-noun tokens of the six children and created three hypothetical children from the very earliest age of language acquisition, which would presumably be the least productive knowledge of determiner usage.
- e. For each learner, the theoretical expectation of overlap is calculated based on equations in Box 2, that is, only with the sample size *S* and the number of distinct nouns *N* in determiner-noun pairs.

Child	Sample Size (S)	a & <i>the</i> Noun types	a or <i>the</i> Noun types (N)	Overlap (expected)	Overlap (empirical)	S/\bar{N}
Naomi (1;1-5;1)	884	60	349	21.8	19.8	2.53
Eve (1;6-2;3)	831	61	283	25.4	21.6	2.94
Sarah (2;3-5;1)	2453	187	640	28.8	29.2	3.83
Adam (2;3-4;10)	3729	252	780	33.7	32.3	4.78
Peter (1;4-2;10)	2873	194	480	42.2	40.4	5.99
Nina (1;11-3;11)	4542	308	660	45.1	46.7	6.88
First 100	600	53	243	22.4	21.8	2.47
First 300	1800	141	483	29.1	29.1	3.73
First 500	3000	219	640	33.9	34.2	4.68
Brown corpus	20650	1175	4664	26.5	25.2	4.43

Table 1. Empirical and expected determiner-noun overlaps in child speech. The Brown corpus is included 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 Box 2.

The theoretical expectations and the empirical measures of overlap agree extremely well (column 5 and 6 in Table 1). Neither paired t-test nor paired Wilcoxon test show significant difference between the two sets of values. Perhaps a more revealing test is linear regression: a perfect agreement between the two sets of value would have the slope of 1.0, and the actual slope is 1.08 (adjusted $R^2 = 0.9716$). In other words, 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 results show (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. Given N nouns in a sample of S , greater overlap value will be obtained when more nouns are expected to occur more than once, or $Sp > 1$. That is, words whose occurrence probabilities that are greater than $1/S$ can contribute to the overlap value; Zipf’s law allows us to express this probability cutoff line in terms with ranks, following equation (1). The approximation below uses a well-known result from Euler’s summation formula.

$$S \frac{1}{r H_N} = 1$$

$$r = \frac{S}{H_N} \approx \frac{S}{\ln N} \quad (4)$$

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 is confirmed in strongest terms: S/N is a near perfect predictor for the empirical values of overlap (last two columns of Table 1): $r = 0.986$, $p < 0.00001$.

We now briefly explore the question whether the determiner usage data by children can be accounted for by the item based approach to language learning. Our effort is hampered by the lack of concrete models for the item-based learning approach, a point that Tomasello concedes [4, p274]. Analytical results like those in Box 2 cannot be similarly obtained. 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 specific and itemized combinations. Similar approaches such as construction grammar [10], usage [27] and exemplar based models [28] make similar commitment to the role of verbatim memory. To this end, we consider a type of learning model that memorizes determiner-noun pairs in the input, and these pairs are then sampled jointly, following the commitment of item-based learning, rather than independently (which would be the productive rule-based view).

Child	Sample Size (S)	Overlap (BIG learner)	Overlap (small learner)	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. Is the full productivity data in child language consistent with item-based learning? Two variants of learners are considered. One type, the BIG learner, is designed to mimic the long term commitment to memory; it stores a large set of determiner-noun pairs, which consists of a sample of 1.1 million child directed utterances from the CHILDES database (methods as described in Box 3). The other variant – the small learner – only memorizes the adult utterances present in each child’s transcript. For both learners, we draw an independent and random sample from these stored D-N pairs with respect to their joint empirical frequencies; this is contrasted with the rule-based model in which D and N are drawn independently. The sample size matches those in each child’s production (Table 1, column 2). The overlap values are then calculated as the percentage of nouns that appear with both “a” and “the” over those that appear with either. The results are given in Table 2, averaged over 1000 trials per child.

Both sets of overlap values from item-based learning (column 3 and 4) are significantly different 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 do not follow the predictions of the item-based learning approach. 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. Intuition is no substitute for theoretical analysis or statistical validation.

4 An Itemized Look at Verbs

The formal analysis in section 3 can be generalized to the study of child verb syntax and morphology (Box 1). Unfortunately, the acquisition data in support of the Verb Island Hypothesis [4] and the item-based nature of early morphology [14] is not available in the public domain.

But there is no escape from the Zipfian grasp: the combinatorics of verbs and their morphological and syntactic associates are similarly lopsided in their usage distribution as in the case of determiners. Consider first the kind of verbal syntax distributions 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 [4, p242], each unique lexical item in the object position counts as a unique construction for the verb. Figure 3 shows the construction frequencies of the top 15 transitive verbs in 1.1 million child directed utterances.

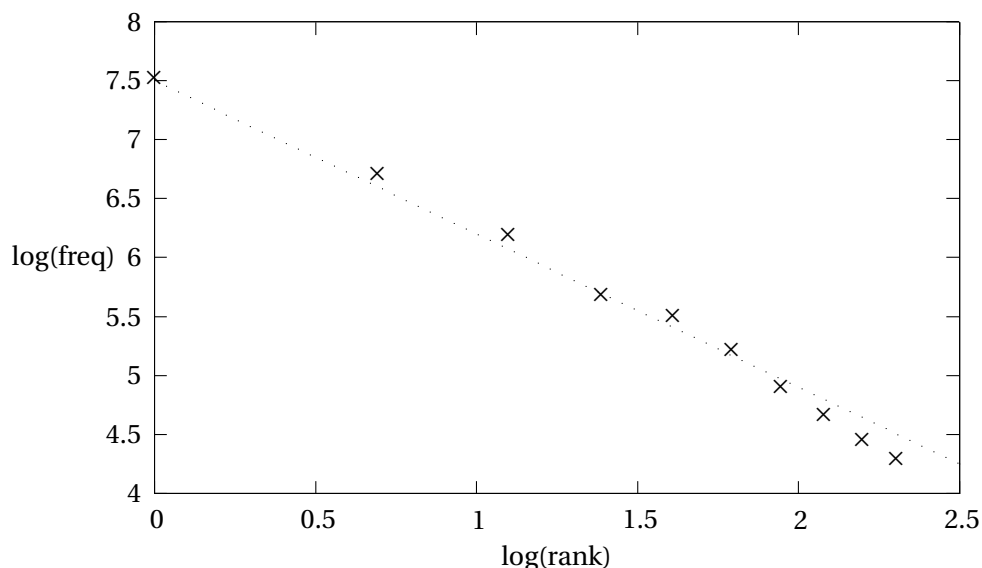


Figure 3. Rank and frequency of verb-object constructions based on 1.1 million child-directed utterances. Processing methods are as described in Box 3 except here we focus on adjacent verb-nominal pairs in part-of-speech tagged texts. The verbs are the top 15 most frequent transitive verbs: *put, tell, see, want, let, give, take, show, got, ask, make eat, like, bring* and *hear*. For each verb, we counted 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: the

frequency tallies are 1904, 838, 501, 301, 252, 189, 137, 109, 88, and 75. The verb-construction frequency is also strongly Zipfian: a verb appears in few constructions frequently and in most constructions infrequently. The observation of Verb Islands, that verbs tend to combine with one or few elements out of larger range, is in fact characteristic of a fully productive verbal syntax system; when the sample size is only modest, as is the case in [4] and indeed most child production studies, to expect anything other than low verb usage diversity is mathematically naive.

The statistical properties of morphology have been investigated by [29] 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 grammar books—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 morphological approaches that rely on exemplars and memorization [11, 30].

Our focus here is to provide a brief assessment of the statistical distribution of morphological forms in child and adult languages. Recall that Italian morphology acquisition study [14, Box 1] that 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 public domain [25].

Subject	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 morphological data is analyzed with a state-of-the-art natural language processing toolkit Freeling (<http://garraf.epsevg.upc.es/freeling/>) which specializes in Romance languages. 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. 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 (in multiple inflections).

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 (4): 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 modest discrepancy in the frequency of one-form verbs.

5 Zipfian Lessons

Outstanding Questions

- Children's full command of determiners does not mean that all aspects of child grammar are equally adult-like. In fact, one of the most revealing findings in language acquisition builds on children's systematic deviation from the target language which nevertheless reflects biologically possible grammatical systems [31, 32]. Moreover, even for components of the grammar acquired early on, as in the case of determiners, one still needs to provide a mechanistic account of *how* the child arrives at the target like state of linguistic knowledge [33].
- Demonstration of linguistic productivity in no way denies the role of memorization: as Sapir famously remarked, all grammars leak [34]. The recognition of exceptions has been explicit throughout the history of generative grammar [35, 36]. Memorization and generalization must work in tandem to ensure successful language acquisition. The central challenge is to develop a learning model which recognizes productive processes and their exceptions as such and proceeds to internalize them as different kinds of linguistic knowledge [37–39].
- It is tempting, though ultimately premature, to construct learning models that exploits the Zipfian distribution of language, e.g., one which treats the most frequent linguistic expressions as the basis of generalization. It is worth noting that linguistic exceptions often reside in the high frequency region — the case of English irregular past tense being a prominent example [8, 9, 32] — and children almost never extended the exceptional forms for productive use [40].
- Switching gears to computational linguistics, it is hoped that the very linguistic assumptions and learning mechanisms that the child brings to the task of language acquisition and successfully overcome the Zipfian hurdle will prove equally useful for natural language processing applications.

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

The *psychologist* and the *linguist*, as seen above, have just been deprived of a convenient means of assessing children's linguistic knowledge. For any type of linguistic expression that involve open class items — and that means *every* type of linguistic expression — even modest mea-

asures 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. Additional methods for probing linguistic knowledge must be sought. But this ought to be old news since Chomsky [1] and Brown [2].

As every *computer scientist* knows, Zipf’s law comes to haunt us as 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 [41]. Figure 4 plots the log-log plot of the top 200 syntactic syntactic rules of modern English from the Penn Treebank [42], which again show excellent Zipfian fit.

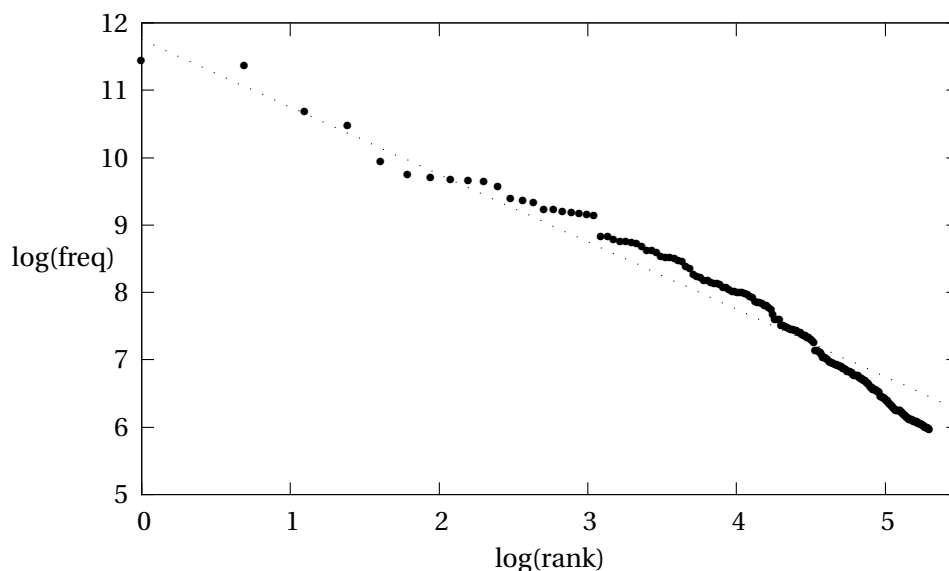


Figure 4. The frequency distribution of the top 200 syntactic rules in the Penn Treebank. Since the corpus has been manually annotated with syntactic structures, it is straightforward to extract rules and tally their frequencies. The most frequent rule is “PP→P NP”, followed by “S→NP VP”. The perfect Zipfian line of slope -1 is given as a reference point.

In a line of research similar in spirit to language acquisition, statistical induction of grammar [43] shows that focusing on lexically specific constructions pays very little dividend: much of the statistical language model’s generalizing ability (implicitly) resides in broad ranging rules. Item-based learning, with its heavy focus on specific and lexically defined constructions, seems ill-equipped for wide linguistic coverage.

But most significant victim of George Kingsley Zipf must be the *child* learner herself. The task faced by children acquiring language is no different from that of the computational linguist, for the input data are also Zipfian in character. The sparse data problem strikes just as hard, and thus the role of memory in language learning should not be overestimated. In linguistics and cognitive science, of course, the learner’s challenge bears another name: the argument from the

poverty of stimulus [44,45]. 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 from the get go still seems the best preparation a child can hope for.

References

- [1] Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- [2] Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.
- [3] Braine, M. (1963). The ontogeny of English phrase structure: The first phase. *Language*, 39, 3-13.
- [4] Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge, MA: Harvard University Press.
- [5] Tomasello, M. (2000a). Do young children have adult syntactic competence. *Cognition*, 74, 209-253.
- [6] Tomasello, M. (2000b). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4, 4: 156-164.
- [7] Tomasello, M. (2003). *Constructing a language*. Cambridge, MA: Harvard University Press.
- [8] Pinker, S. & Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, 6, 456-463.
- [9] McClelland, J. & Patterson, K. (2002). Rules or connections in past-tense inflection: What does the evidence rule out? *Trends in Cognitive Sciences*, 6, 465-472.
- [10] Goldberg, E. (2003). Constructions. *Trends in Cognitive Science*, 7, 219-224.
- [11] Hay, J. & Baayen, H. (2005). Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences*, 9, 342-348.
- [12] Culicover, P. & Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends in Cognitive Sciences*, 10, 413-418.
- [13] Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.
- [14] 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.

- [15] Pine, J. & Lieven, E. (1997). Slot and frame patterns in the development of the determiner category. *Applied Psycholinguistics*, 18, 123-138.
- [16] Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, 22, 562-579.
- [17] Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley.
- [18] Kučera, H & Francis, N. (1967). *Computational analysis of present-day English*. Providence, RI: Brown University Press.
- [19] Baroni, M. (2008). Distributions in text. In Lüdelign, A. & Kytö, M. (Eds.) *Corpus linguistics: An international handbook*. Berlin: Mouton de Gruyter.
- [20] Mandelbrot, B. (1954). Structure formelle des textes et communication: Deux études. *Words*, 10, 1-27.
- [21] Li, W. (1992). Random texts exhibit Zipf's law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38 (6), 1842-1845.
- [22] 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.
- [23] Chomsky, N. (1958). Review of *Langage des machines et langage humain* by Par Vitold Belevitch. *Language*, 34 (1), 99-105.
- [24] 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.
- [25] MacWhinney, B. (2000). *The CHILDES Project*. Lawrence Erlbaum.
- [26] 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.
- [27] Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.
- [28] Pierrehumbert, J. (2001). Exemplar dynamics. In Bybee, J. & Hopper, P. (Eds.) *Frequency and emergence of linguistic structure*. Amsterdam: Johns Benjamins. 137-158.
- [29] Chan, E. (2008). Structures and distributions in morphology learning. Ph.D. Dissertation. Department of Computer and Information Science. University of Pennsylvania. Philadelphia, PA.
- [30] Seidenberg, M. & Gonnerman, L. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*. 4, 353-361.

- [31] Hyams, N. (1986). *Language Acquisition and the Theory of Parameters*. Dordrecht: Reidel.
- [32] Yang, C. (2002). *Knowledge and Learning in Natural Language*. New York: Oxford University Press.
- [33] Yang, C. (2004). Universal Grammar, Statistics, or Both. *Trends in Cognitive Sciences*, 8, 451-456.
- [34] Sapir, E. (1928). *Language: An introduction to the study of speech*. New York: Harcourt Brace.
- [35] Chomsky, N. (1962). Explanatory models in linguistics. In Nagel, E., Suppes, P. & Tarski, A. (Eds.) *Logic, Methodology and Philosophy of Science*. Stanford, CA: Stanford University Press.
- [36] Chomsky, N. & Halle, M. (1968). *The Sound Patterns of English*. New York: Harper & Row.
- [37] Fodor, J. D. & Crain, S. (1987). Simplicity and generality of rules in language acquisition. In MacWhinney, B. (Ed.) *Mechanisms of Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum. 35-64.
- [38] Pinker, S. (1989). *Language Learnability and Cognition*. Cambridge, MA: MIT Press.
- [39] Yang, C. (2009). Three factors in language variation. *Lingua*. Forthcoming.
- [40] Xu, F., & Pinker, S. (1995). Weird Past Tense Forms. *Journal of Child Language*, 22:531-556.
- [41] Jelinek, F. (1998). *Statistical methods for speech recognition*. Cambridge, MA: MIT Press.
- [42] Marcus, M., Marcinkiewicz, M. & Santorini, B. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19, 313-330.
- [43] Bikel, D. (2004) Intricacies of Collins' parsing model. *Computational Linguistics*, 30, 479-511.
- [44] Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.
- [45] Legate, J. A. & Yang, C. (2002). Empirical reassessments of poverty stimulus arguments. *Linguistic Review*, 19, 151-162.