Who's Afraid of George Kingsley Zipf?

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

April 11, 2009
Version 2.0

Abstract

We study the statistical distributions of natural language and develop a novel approach to assess the properties of the underlying grammar given a sample of linguistic production. 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 one version of the story 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 — there are similar stories with other famous subjects — 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 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 (1965) drew the competence-performance distinction. The

*For helpful comments, I would like to thank Virginia Valian, Julie Anne Legate, Mark Liberman, and the audience at the 2009 Schultink Lecture, the University of Groningen, where these materials were first presented. Special thanks to Erwin Chan for his help with the morphological data.

pioneering work on child language that soon followed, most notably Brown (1973),[1] recognized the potential gap between what the child knows and what the child says and laid out the influential tradition that interprets child language in terms of adult-like grammatical devices.

That tradition has been challenged by what is referred to as the *item*-based approach to language (Tomasello 1992, 2000a, 2000b, 2003). According to this view, child language, particularly in the early stages, is organized around specific item-based schemas, rather than productive linguistic system as previously conceived. Consider, for instance, three case studies (Tomasello 2000a; in Box 1) which have been cited as evidence for the item-based view in numerous places.

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

- The Verb Island Hypothesis (Tomasello 1992). In his child's early speech, 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. In a study of child Italian, Pizutto & Caselli (1994) find 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 $\times$ 2 number). Only 13% of all verbs appeared in 4 or more forms.
- Determiner usage. Pine & Lieven (1997) find 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 Valian's (1986) study, which maintains that child determiner use is adult-like by the age of 2;0.

So far as we can tell, these conclusions about children's grammar based children's language are made solely on the basis of intuition or informal inspections, rather than rigorous statistical analysis. For the numerous examples from his child's speech, not a single statistical test can be found in Tomasello (1992) where the Verb Island Hypothesis and related

[1]See, in particular, Brown's critique of the Pivot Grammar hypothesis (Braine 1963), which bears more than a passing resemblance with some contemporary theorizing of child language, some of which is reviewed here.

ideas about item-based learning are first put forward. It would seem more prudent to make assertions about child language only 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 this 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. Our broader aim is to direct language researcher to certain statistical properties of natural language that are widely known but not widely appreciated, and to discuss the implications of these properties 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* (1949), the distributions of words in natural language follow a curious pattern: 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:

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

In the Brown corpus (Kučera & 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. The Zipfian characterization of word frequency can be visualized by plotting the log of word frequency against the log of word rank. 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). 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.

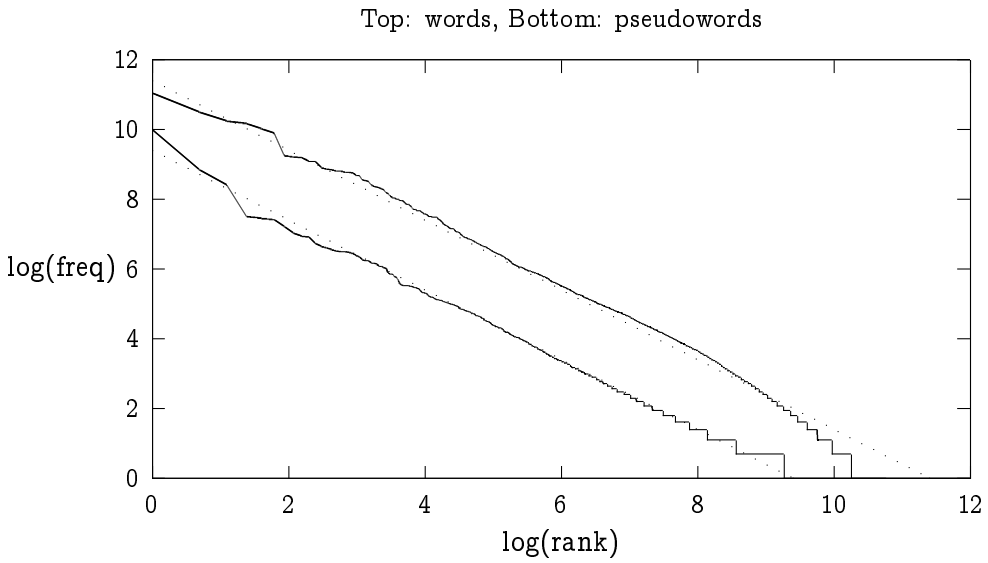

Figure 1. Zipfian distribution of words and pseudowords in the Brown corpus. The lower line is plotted by taking "words" to be any sequence of letters between *e*'s; see text. The two straight lines are linear functions with the slope -1, i.e., perfect Zipfian fit.

There has been a good deal of controversy over the interpretation of Zipf's law in the context of language, cognition, and other physical systems. It is now clear that the observation of a Zipfian distribution alone does not reveal anything interesting or specific about language but only statistical properties of certain random generating processes; see Mandelbrot (1954), Li (1992) and Niyogi & Berwick (1995). As noted by Chomsky long ago (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 particularly concerned with the actual frequencies of words but their probability of occurrence. Zipf's law gives us the probability p of the word n, whose rank is r among N words:

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

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 case of determiner use in early child language (Box 1). The *overlap* metric (Pine & Lieven 1997) follows the logic of the Verb Island hypothesis (Tomasello 1992). 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, the linguistically productive child that uses "a" with a noun is expected to automatically transfer the use of that noun to "the". Thus, the determiner overlap is defined as the percentage of nouns that appears with both determiners out of those that appear with either. That several children show overlap measures significantly below chance is taken to be evidence for item-based characterization of child language and against a fully productive grammar. Using a similar but somewhat different metric, Valian, Solt & Stewart (2008) replicate low measures of determiner overlap but they also find no difference in the speech of young children compared with their mothers. Indeed, when applied to the Brown corpus (see Box 3 for methods), we obtain an overlap value of 25.2%, which is actually lower than those reported by Pine & Martindale's (1996) and Pine & Lieven (1997), the lowest of which are in the region of 30%. It thus follows that the language of the Brown corpus, which draws various genres of professional print materials, would be less productive and more item-based than a toddler, a conclusion that 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 available data for empirical tests but one can easily substitute the rule for "VP$\rightarrow$ V DP", "VP$\rightarrow$ V in Construction$_{\chi}$", "$V_{inflection} \rightarrow V_{stem}$ + Person + Number + Tense". All such cases can be analyzed with the methods provided here.

Suppose we have taken a sample in which D determiners and N nouns have combined to form S pairs. The full productivity of the DP rule, by definition, means that D and N can combine independently. Several observations can be made about the distributions of D and N. First, nouns (and open class words in general) exhibit excellent fit of Zipf's law: the Brown corpus, for instance, shows a log-log slope of -0.97 (see Box 3 for methods). Second, while the combination of D ("the" and "a") and N are 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 DPs are perfectly grammatical. Such skewed distributions of D-N distribution is exactly what Zipf's law entails. As noted above, 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 (1175 in all) show strong biases: only a further 25% (297) are

used with "a" and "the" equally frequently. Overall, the frequency ratio between the more favored over less favored determiner is 2.86:1. These general pattern hold for child and child directed data as well. In the six child-adult pairs (thus 12 individuals) 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 interpretation of Zipf's law, they unambiguously point out the considerable asymmetry in category combination usage.

We now turn to the theoretical analysis of D-N overlap (Box 2).

Box 2. Calculating Expected Overlap in Determiner Noun Usage

Let $O(N, S)$ be the percentage of N nouns in a sample S pairs of D-N pairs. Consider a noun n whose rank is r out of N. Following (1), it has a probability of $p = 1/(rH_N)$ of being drawn at any single trial in S; its expected occurrence in S is thus simply Sp. The expected probability of n being used with more than 1 determiners is 1 - the expected probability of n being used with *exactly* 1 determiner in all of the Sp trials. Obviously, if a noun is expected to be sampled once or less, it will have an overlap of zero. Let the expected overlap of n be $O(r, N, S)$.

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

$$O(r, N, S) = \begin{cases} 1 - \sum_{i=1}^{D} d_i^{(Sp)} & \text{if } Sp > 1 \text{ where } d_i = \frac{1}{iH_D}, p = \frac{1}{rH_N} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The probability of determiners $d_i (i = 1, 2)$ in (3) also follows Zipf's law (1).[a]

[a] Although 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 rations in the calculation of (2) and (3). This is because most of overlap comes from the relatively few and high frequent nouns, which can almost surely be expected to occur with both determiners.

The frequency distribution of both D and N ensures that the expected overlap be relatively low unless the sample size S is very large. First, the majority of nouns will have zero overlap in S because they will have been sampled only once or less thanks to their Zipfian frequencies. Second, even if a noun is sampled more than once, there is still a significant chance that it is combined with a unique determiner, thanks to the Zipfian distribution of the determiners. Figure 2 plots the expected probability of overlap for 50 nouns (ordered by rank) when combined with two determiners in a sample of 100 D-N pairs: that is, $D = 2$,

$N = 50$, and $S = 100$. A few of nouns almost surely occur with both determiners but the number of those that do so drops off sharply. The $O(50, 100) \approx 20.6\%$.

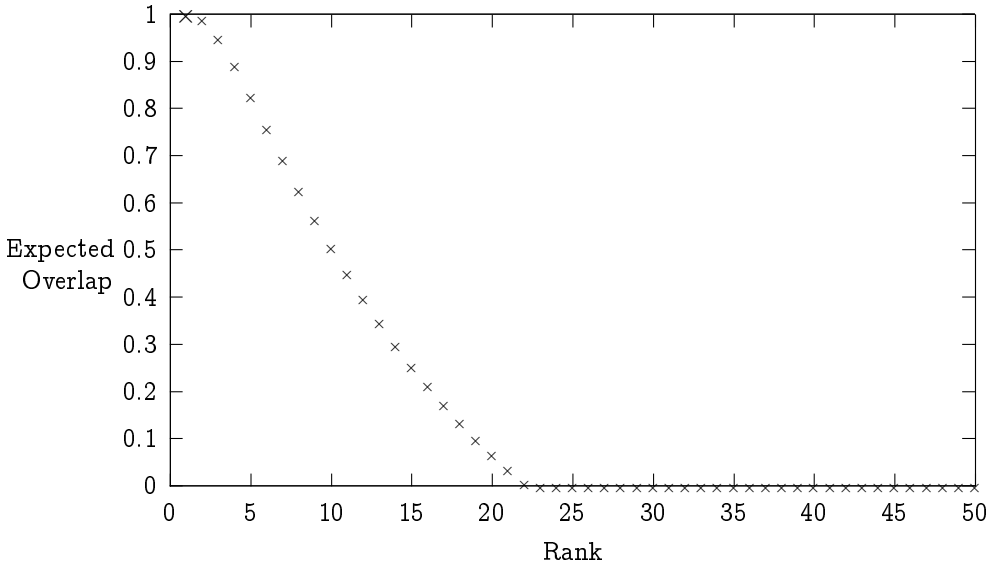

Figure 2. Expected overlap probability of nouns ordered by rank $O(r, 50, 100), r = 1...50$.

Box 3. Empirical Studies of Overlap in Language Production

a. The Brown corpus (Kučera & Francis 1967) has been previously tagged with part-of-speech (POS). These POSs are used to extract D-N pairs following the procedure in step c. below.

b. For language acquisition datasets, we consider the data of Adam, Eve, Sarah (Brown 1973), Naomi (Sachs 1983), Nina (Suppes 1974) and Peter (Bloom, Lightbrown, & Hood 1975). These are 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. 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) supplemented with statistical information (available http://gposttl.sourceforge.net/). For languages such as English (and many other Indo-European languages), 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.

c. With POS tagged datasets, we extracted adjacent D-N 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 S.

d. For an additional test, we have pooled together the first 100, 300, and 500 D-N 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.

Child	Sample Size (S)	*a* & *the* Noun types	*a* or *the* Noun types (N)	Overlap (expected)	Overlap (empirical)	$\frac{S}{N}$
Naomi (1;1-5;1)	884	60	349	19.0	19.8	2.53
Eve (1;6-2;3)	831	61	283	22.7	21.6	2.94
Sarah (2;3-5;1)	2453	187	640	26.4	29.2	3.83
Adam (2;3-4;10)	3729	252	780	32.0	32.3	4.78
Peter (1;4-2;10)	2873	194	480	43.0	40.4	5.99
Nina (1;11-3;11)	4542	308	660	47.2	46.7	6.88
First 100	600	53	243	19.6	21.8	2.47
First 300	1800	141	483	26.7	29.1	3.73
First 500	3000	219	640	32.3	34.2	4.68
Brown corpus	20650	1175	4664	23.8	25.2	4.43

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

The theoretical expectations and the empirical measures of overlap agree extremely well (column 5 and 6 in Table 1). Paired t-test and Wilcoxon test show no significant difference between the two sets of values. Perhaps a more revealing test is linear regression: a perfect agreement between theoretical and empirical overlap values would have the slope of 1.0, and the actual slope is 1.08. In other words, the determiner usage data from child language is consistent with the hypothesis that the underlying grammar in early child language is the fully productive "DP$\rightarrow$ D N".

The empirical test also reveals considerable variation in the overlap values across individuals. As the Brown corpus shows, sample size S, the number of nouns N, and the language user's age are not predictive of the overlap value. The variation can be formally analyzed. Consider again Box 2. Given N nouns in a sample of S, the greater value of overlap will be obtained if more nouns have the expected value of occurrence greater than 1, or $Sp > 1$. We thus solve

$$S\frac{1}{rH_N} = 1$$
$$r = \frac{S}{H_N} \approx \frac{S}{\ln N} \tag{4}$$

That is, nouns whose ranks are lower than $S/(\ln N)$ can be expected to be non-zero overlaps. The total overlap $O(N, S)$ is thus a monotonically increasing function of $S/(N \ln N)$, which, given the slow growth of $\ln N$, is approximately S/N–which 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): $r = 0.9608$, $p < 0.0001$.

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 himself concedes (1992, p274). Analytical results (Box 2) cannot be similarly obtained. A plausible approach can be established around a central tenet of item-based learning, that the child does not form grammatical generalizations but rather memorizes specific and itemized combinations. Similar frameworks such as construction grammar (Goldberg 2003), usage (Bybee 2001) and exemplar based models (Pierrehumbert 2001) make similar commitment to the role of verbatim memory. To this end, we consider an item-based learner that simply memorizes all the input tokens in the input data, which consists of a sample of 1.1 million adult utterances from the CHILDES database. Using the same methods described in Box 3, we all D-N pairs (approximately 140,000) in the input, which consists of the item based learner's memory.[2] We then 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. Matching for sample size, we obtain several itemized learning children (denoted with the subscript I) and compare the overlap value thus obtained with the actual values in Table 1. The results are given in Table 2, averaged over 1000 trials per learner.

Child	Sample Size (S)	*a* & *the* Noun types	*a* or *the* Noun types (N)	Overlap (itemized)	Overlap (empirical)
Eve_I	831	70	438	16.0	21.6
Naomi_I	884	76	456	16.6	19.8
Sarah_I	2453	203	832	24.5	29.2
Peter_I	2873	232	906	25.6	40.4
Adam_I	3729	285	1035	27.5	32.3
Nina_I	4542	328	1147	28.6	46.7
First 100_I	600	49	356	13.7	21.8
First 300_I	1800	155	701	22.1	29.1
First 500_I	3000	239	926	25.9	34.2

Table 2. Overlap comparison between item-based learning and empirical measures.

The overlap values thus obtained are significantly different from the empirical values under a paired t-test ($p <= 0.001$), suggesting that children's use of determiners do not follow the prediction 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. Intuitions are no substitute for theoretical models or empirical evaluations.

[2]We put aside the important question how such a learner can selectively retain D-N pairs as relevant items for memorization, rather than, say, categorical combinations such as Aux-D ("is a") or Pronoun-Verb ("I see") which are highly frequent.

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 (Tomasello 1992) and the item-based nature of early morphology (Pizutto & Caselli 1994) are not available in the public domain for examination.

But there is no escape from Zipf's 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 transitive verbs and their immediately adjacent nominal objects, including pronouns and noun phrases, which can be readily extracted from minimally processed corpora. Using the 1.1 million child directed utterances on which Table 2 is based, we extracted the top 15 most frequently transitive verbs: *put, tell, see, want, let, give, take, show, got, ask, make eat, like, bring* and *hear*. For each verb, we counted, and then ranked, the top 10 frequencies with which it appears with a unique lexical item: this corresponds to the definition of "sentence frame" in Tomasello's original verb Island Hypothesis study (1992, p242). For each of the 10 ranks, the frequencies of all 15 verbs are then tallied. Figure 3 then gives the log-log plot of rank and total frequency: it can be observed that verb-object combinations follow a Zipfian distribution virtually perfectly. The observation of Verb Islands, that verbs tend to combine with one or few elements, is in fact characteristic of a fully productive verbal syntax system; when the sample size is only modest, as is the case in Tomasello's (1992) study and indeed most child production studies, to expect anything other than low verb usage diversity is mathematically naive.

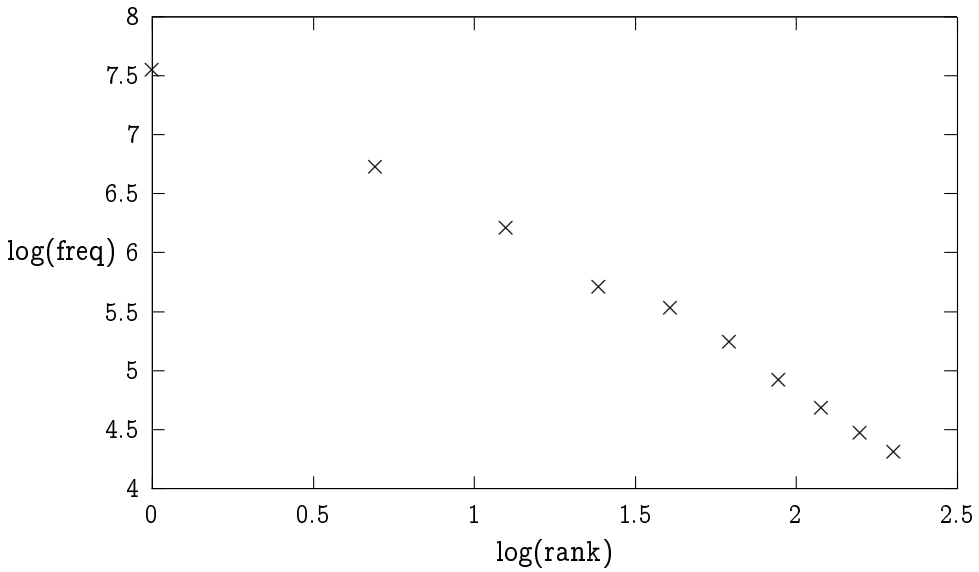

Figure 3. Rank and frequency of verb and nominal object combinations. Raw frequency

tallies are 1904, 838, 501, 301, 252, 189, 137, 109, 88, and 75, based on 1.1 million child directed utterances.

The statistical properties of morphology have been investigated by Chan (2008) in an independent context. It is shown that Zipf-like distributions are strongly attested not only at the word level, but also at the level of morphological paradigm. There are relatively few stems that appear in a great number of inflections. However, they never approach anywhere near the maximum number of possible inflections, which may be dozens and dozens as in the case of moderately richly inflected languages. 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, not even for a single stem. Furthermore, the inflections themselves do not have uniform frequencies either: few are used very frequently but most are used rarely.

The interested reader should turn to Chan (2008) for details for that important work, and perhaps more significantly, their implications on the theory of morphology and models of language acquisition. Our focus here is to provide a brief assessment of the statistical distribution of morphological forms in child and adult languages. Recall that Pizutto & Caselli's results cited by Tomasello as the evidence for usage-based learning that 47% of verbs appear in exactly one of 6 possible agreement forms – i.e., (1, 2, 3 person) $\times$ (singular, plural) – while 40% of verbs appear with 2 or 3 out of 6 possible agreement forms, and only 13% of verbs appear in 4 or more. 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 CHILDES (MacWhinney 2000).[3].

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 cell represents the percentage of verb stems that are used in 1, 2, 3, 4, 5, and 6 inflectional forms.

A formal treatment of the agreement distributions similar to the overlap study requires multinomial analysis that we do not pursue here. Nevertheless, the logic of the problem

[3]The morphological data is analyzed through the open source natural language processing toolkit freeling (http://garraf.epsevg.upc.es/freeling/), which puts special attention to Romance languages. Only tensed forms are counted; infinitives, which do not bear person/number agreement in these languages, are ignored. We thank Erwin Chan for his help in extracting

remains the same as in (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 token/type 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 Summary

Even when Einstein became a world renowned scientist and a talkative one no less, he would still be classified as an item-based learner according to Tomasello's measures. Indeed, the determiner and verbal syntax and morphology studies show that productivity measures based on diversity of usage cannot be taken as evidence for the item-based approach to language. Given the omnipresence of Zipf-like distributions, it is not surprising that these measures turn out consistently low across genres, categories, and speakers old and young. The alternative hypothesis of linguistic productivity not only cannot be rejected; it is in fact strongly supported.

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 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. Additional methods for probing linguistic knowledge must be sought. But this ought to be old news since Chomsky (1965) and Brown (1973).

As every *natural language engineer* knows, Zipf's law comes to haunt us 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. As a result, 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 1993). Again, Chan's (2008) works shows that the sparse data problem persists even for words and morphologies, which at a first glance appear most amenable to item-based learning and other approaches that put extensive requirement on memorization.

But most significant victim of George Kingsley Zipf must be the *child* learner himself. The task faced by children acquiring language is no different from that of the computational linguist, for the input data from adults are 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 (Chomsky 1975, Legate & Yang 2002). 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. This much is the statistical reality of natural language. It seems to us that a grammatical system with full generative potentials from the get go remains the best preparation that a child could hope for.

References

Bloom, L., Lightbrown, P., & Hood, L. (1975). Structure and variation in child language. *Monographs of the Society for Research in Child Development*, 40, (Serial No. 160).

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.

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

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

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.

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

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.

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

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.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. In Nelson, K. E. (Ed.) *Children's Language*. Vol 4. Hillsdale, NJ: Lawrence Erlbaum.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, 29, 103-114.

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

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.

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