A Statistical Test for Grammar

Charles Yang
Department of Linguistics & Computer Science
Institute for Research in Cognitive Science
University of Pennsylvania
charles.yang@ling.upenn.edu

Abstract

We propose a statistical test for measuring grammatical productivity. We show that very young children's knowledge is consistent with a systematic grammar that independently combines linguistic units. To a testable extent, the usage-based approach to language and language learning, which emphasizes the role of lexically specific memorization, is inconsistent with the child language data. We also discuss the connection of this research with developments in computational and theoretical linguistics.

1 Introduction

Eistein was a famously late talker. The great physicist's first words, at the ripe age of three, were to proclaim "The soup is too hot." Apparently he hadn't had anything interesting to say.

The moral of the story is that one's linguistic behavior may not be sufficiently revealing of one's linguistic knowledge. The problem is especially acute in the study of child language since children's linguistic production is often the only, and certainly the most accessible, data on hand. Much of the traditional research in language acquisition recognizes this challenge (Shipley et al. 1969, Slobin 1971, Bowerman 1973, Brown 1973) and has in general advocated the position that child language be interpreted in terms of adult-like grammatical devices.

This tradition has been challenged by the usage-based approach to language (Tomasello 1992, 2000a) which, while reviving some earlier theories of child grammar (Braine 1964), also reflects a current trend in linguistic theorizing that emphasizes the storage of specific linguistic forms and constructions at the expense of general combinatorial linguistic principles and overarching points of language variation (Goldberg 2003, Sag 2010, etc.). Child language, especially in the early stages, is claimed to consist of specific usage-based schemas, rather than productive linguistic system as previously conceived. The main evidence for this approach comes from the lack of combinatorial diversity–the hallmark of a productive grammar–in child language data (Tomasello 2000a). For instance, verbs in young children's language tend to appear in very few frames rather than across many; this "uneveness" has been attributed to the verb-specific predicate structures rather than general/categorical rules. Similar observations have been made in the acquisition of inflectional morphology, where many stems are used only in relatively few morphosyntactic contexts (e.g., person, number). Another concrete example comes from the syntactic use of the determiners "a" and "the", which can be interchangeably used with singular nouns.[1] An *overlap* metric has been defined as the ratio of nouns appearing with both "a" and "the" out of those appearing with either. Pine & Lieven (1997) find that overlap values are generally low in child language, in fact considerably below chance level. This finding is taken to support the view that the child's determiner use is bound with specific nouns rather than reflecting a productive grammar defined over the abstract categories of determiners and nouns (Valian 1986).

[1] Although "a" is typically described as combining with countable nouns, instances such as "a water", "a sun" and "a floor" are frequently attested in both child and adult speech from CHILDES.

Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, pages 30–38,
Portland, Oregon, June 2011.

The computational linguistics literature has seen the influence of usage-based approach: computational models have been proposed to proceed from an initial stage of lexicalized constructions toward a more general grammatical system (Felman 2004, Steels 2004, cf. Wintner 2009). However, as far as we can tell, the evidence for an unproductive stage of grammar as discussed above was established on the basis of intuition rather than rigorous assessments. We are not aware of a statistical test against which the predictions of usage-based learning can be verified. Nor are we of any demonstration that the child language data described above is *inconsistent* with the expectation of a fully productive grammar, the position rejected in usage-based learning. It is also worth noting that while the proponents of the grammar based approach have often produced tests for the *quality* of the grammar–e.g., the errors in child language are statistically significantly low–they have likewise failed to provide tests for the *existence* of the grammar. As has been pointed out in the usage-based learning literature, low error rates could be the result of rote memorization of adult linguistic forms.

In this paper, we provide statistical analysis of grammar to fill these gaps. The test is designed to show whether a corpus of linguistic expressions can be accounted for as the output of a productive grammar that freely combines linguistic units. We demonstrate through case studies based on CHILDES (MacWhinney 2000) that children's language shows the opposite of the usage-based view, and it is the productivity hypothesis that is confirmed. We also aim to show that the child data is inconsistent with the memory-and-retrieval approach in usage-based learning (Tomasello 2000b). Furthermore, through empirical data and numerical simulations, we show that our statistical test (correctly) over-predicts productivity for linguistic combinations that are subject to lexical exceptions (e.g., irregular tense inflection). We conclude by drawing connections between this work and developments in computational and theoretical linguistics.

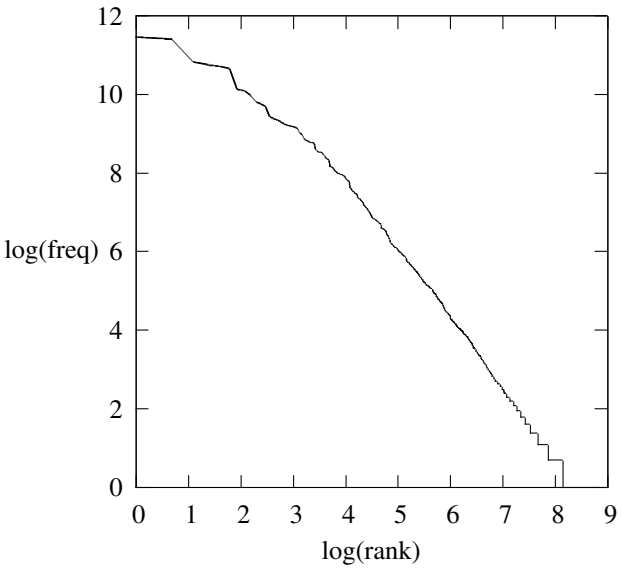

Figure 1: The power law frequency distribution of Treebank rules.

2 Quantifying Productivity

2.1 Zipfian Combinatorics

Zipf's law has long been known to be an omnipresent feature of natural language (Zipf 1949, Mendelbrot 1954). Specifically, the probability p_r of the word n_r with the rank r among N word types in a corpus can be expressed as follows:

$$p_r = \left(\frac{C}{r}\right) \Big/ \left(\sum_{i=1}^{N} \frac{C}{i}\right) = \frac{1}{rH_N}, \quad H_N = \sum_{i=1}^{N} \frac{1}{i} \tag{1}$$

Empirical tests show that Zipf's law provides an excellent fit of word frequency distributions across languages and genres (Baroni 2008).

It has been noted that the linguistic combinations such as n-grams show Zipf-like power law distributions as well (Teahna 1997, Ha et al. 2002), which contributes to the familiar sparse data problem in computational linguistics. These observations generalize the combination of morphemes (Chan 2008) and grammatical rules. Figure 1 plots the ranks and frequencies syntactic rules (on log-log scale) from the Penn Treebank (Marcus et al. 1993); certain rules headed by specific functional words have been merged.

Claims of usage-based learning build on the

premise that linguistic productivity entails diversity of usage: the "unevenness" in usage distribution such as the low overlap in D(eterminer)-N(oun) combinations is taken to be evidence against a systematic grammar. Paradoxically, however, Valian et al. (2008) find that the D-N overlap values in mothers' speech to children do not differ significantly from those in children's speech. In fact, when applied to the Brown corpus, 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): the language of the Brown corpus, which draws from various genres of professional print materials, must be regarded as less productive and more usage-based than that of a toddler—which seems absurd.

Consider the alternative to the usage based view, 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|...". Other rules can be similarly formulated: e.g., "VP→ V DP", "$V_{inflection}$ → V_{stem} + Person + Number + Tense". 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 in the study of D-N usage diversity. First, nouns 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 speech transcripts of six children (see section 3.1 for details for data analysis), the average value of log-log slope is -0.98. Thus, 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 in the DP rule is syntactically interchangeable, N's may 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. 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 both 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. These general patterns hold for child and adult speech data as well. In the six children's transcripts (section 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. As a result, even when a noun appears multiple times in a sample, there is still a significant chance that it has been paired with a single determiner in all instances.

We now formalize the overlap measure under the assumption of a rule and Zipfian frequencies of grammatical combinations.

2.2 Theoretical analysis

Consider a sample (N, D, S), which consists of N unique nouns, D unique determiners, and S determiner-noun pairs. The nouns that have appeared with more than one (i.e. two, in the case of "a" and "the") 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 (1), it has the 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:

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

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:

$$\begin{aligned} O(r,N,D,S) &= 1 - \Pr\{n_r \text{ not sampled during } S \text{ trials}\} \\ &\quad - \sum_{i=1}^{D} \Pr\{n_r \text{ sampled } i\text{th exclusively}\} \\ &= 1-(1-p_r)^S \\ &\quad - \sum_{i=1}^{D}\Big[(d_i p_r + 1 - p_r)^S - (1-p_r)^S\Big] \end{aligned} \tag{3}$$

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 ith determiner is the product of their probabilities, or $d_i p_r$. The multinomial expression

$$(p_1 + p_2 + ... + p_{r-1} + d_i p_r + p_{r+1} + ... + p_N)^S \tag{4}$$

gives the probabilities of all the compositions of the sample, with n_r combining with the ith 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 ith determiner zero times—again $(1-p_r)^S$—which must be subtracted. Thus, the probability with which n_r combines with the ith 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:

$$O(r,N,D,S) = 1+(D-1)(1-p_r)^S - \sum_{i=1}^{D}\Big[(d_i p_r + 1 - p_r)^S\Big] \tag{5}$$

The formulations in (2)—(5) 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.[2] We now turn to the empirical evaluations of the overlap test (2).

3 Testing Grammar Productivity

3.1 Testing grammar in child language

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 with substantial longitudinal data that starts at the very beginning of syntactic development (i.e, one or two word stage) so that the usage-based stage, if exists, could be observed. For comparison, we also consider the overlap measure of the Brown corpus (Kucera & Francis 1967), for which the writers' productivity is not in doubt.

We applied a variant of the Brill tagger (1995) (http://gposttl.sourceforge.net/) to prepare the child data before extracting adjacent pairs of determiners followed by singular nouns. While no tagger works perfectly, the determiners "a" and "the" are not ambiguous which reliably contribute the tagging of the following word. The Brown Corpus is already manually tagged and the D-N pairs are extracted directly. In an additional test, we pooled together the first 100, 300, and 500 D-N pairs from the six children and created three hypothetical children in the very earliest, and presumably least productive, stage of learning.

For each child, the theoretical expectation of overlap is calculated based on equations in (2)—(5), 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; i.e., the percentage of nouns appearing with both "a" and "the". The results are summarized in Table 1.

The theoretical expectations and the empirical measures of overlap agree extremely well (column 5 and 6 in Table 1). Neither paired t- nor paired Wilcoxon test reveal significant difference between the two sets of values. A linear regression produces empirical $= 1.08 \times$ theoretical, $R^2 = 0.9716$: a perfect fit between theory and data would have the slope of 1.0. Thus we may conclude that the determiner usage data from child language is consistent

[2] For the present case involving only two determiners "the" and "a", $d_1 = 2/3$ and $d_2 = 1/3$. As noted in section 2.1, 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 (2)—(5). This is because most of average overlap value comes from the relatively few and high frequent nouns.

Subject	Sample Size (S)	*a* or *the* Noun types (N)	Overlap% (expected)	Overlap% (empirical)	$\frac{S}{\overline{\text{N}}}$
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 and the Brown corpus (last row).

with the productive rule "DP→ D N".

The results in Table 1 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. Given N unique nouns in a sample of S, greater overlap value can be obtained if more nouns occur more than once. Zipf's law (1) 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 $H_N = \sum_{i=1}^{N} 1/i$ can be approximated by $\ln N$.

$$S\frac{1}{rH_N} = 1$$
$$r = \frac{S}{H_N} \approx \frac{S}{\ln N} \quad (6)$$

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

3.2 Testing usage-based learning

We turn to the question whether children's determiner usage data can be accounted for equally well by the usage based approach. In the limiting case, the usage-based child learner could store the input data in its entirety and simply retrieve these memorized determiner-noun pairs in production.

Our effort is hampered by the lack of concrete predictions about child language from the usage-based literature. Explicit models in usage-based learning and similar approaches (e.g., Chang et al. 2005, Freudenthal et al. 2007, etc.) generally involve programming efforts for which no analytical results such as (2)–(5) are possible. Nevertheless, a plausible approach can be construed based on a central tenet of usage-based learning, that the child does not form grammatical generalizations but rather memorizes and retrieves specific and item-based combinations. For instance, Tomasello (2000b) suggests "(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. These pairs will then be sampled; 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 in which the learner memorizes all past linguistic ex-

Child	sample	% (global)	% (local)	% (emp.)
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 usage-based learning and empirical results.

perience. To implement this, we extracted all D-N pairs from about 1.1 million child directed English utterances in CHILDES. The second model is a *local memory* learner, which is construed to capture the linguistic experience of a particular child. The local memory learner only memorizes the determiner-noun pairs from the adult utterances in that particular child's CHILDES transcripts. In both models, the memory consists of a list of jointly memorized D-N pairs, which are augmented with their frequencies in the input.

For each child with a sample size of S (see Table 1, column 2), and for each variant of the memory model, we use Monte Carlo simulation to randomly draw S pairs from the memorized lists. The probability with which a pair is drawn is proportional to its frequency. We then calculate the D-N overlap value, i.e, the the percentage of nouns that appear with both "a" and "the", for each sample. The results are averaged over 1000 draws and presented in Table 2.

Both sets of overlap values from the two variants of usage-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 usage-based learning approach. This conclusion is tentative, of course, as we reiterate the need for the usage-based approach to provide testable quantitative predictions about child language. At the minimum, child language does not appear to stem from frequency sensitive retrieval of jointly stored determiner-noun constructions (Tomasello 2000b).

Similar considerations apply to other linguistic examples. For instance, it is often noted (Lieven, Pine & Baldwin 1997) that child language is dominated by a small number of high frequency frozen frames (e.g, "give me (a) X").[3] True, but that appears no more than the reflection of the power law distribution of linguistic units. In the Harvard corpus of child English (Brown 1973), the frequencies of "give me", "give him" and "give her" are 93:15:12, or 7.75:1.23:1, and the frequencies of "me", "him" and "her" are 2870:466:364, or the virtually identical 7.88:1.28:1.

3.3 Testing for Unproductivity

Any statistical test worth its salt should be able to distinguish occurrences from non-occurrences of the pattern which it is designed to detect. If the productivity test predicts *higher* overlap values than empirically attested–assuming that these classes and their combinations follow Zipfian distribution–then there would be reason to suspect that the linguistic types in question do not combine completely independently, and that some kind of lexically specific processes are at work.

We test the utility of the productivity test on inflectional morphology. In English, the -ing suffix can attach to all verb stems, only some of which can take the -ed suffix–the rest are irregulars. Chan (2008) shows that in morphological systems across languages, stems, affixes, and their combinations tend to show Zipf-like distributions. Therefore, if we apply the productivity test to -ing and -ed inflected forms (i.e, assuming that -ing and -ed were fully interchangeable), then the predicted overlap value should be higher than the empirical value. Table 3 gives the results based on the verbal morphology data from the Brown corpus and the six children studied in section 3.1. Clearly there are very significant discrepancies between the empirical and predicted overlap values.

It can be reasonably objected that English irregular paste tense forms are highly frequent, which may contribute to the large discrepancies observed in Table 3. To address this concern, we created an artificial morphological system in which 100 stems

[3]Thanks to an anonymous reviewer for bringing up this example.

Subject	sample	# stems	% emp.	% pred.
Adam	6774	263	31.3	75.6
Eve	1028	120	20.0	61.7
Sarah	3442	230	28.7	76.8
Naomi	1797	192	32.3	61.9
Peter	2112	139	25.9	78.8
Nina	2830	191	34.0	77.2
Brown	62807	3044	45.5	75.6

Table 3: Empirical vs. predicted overlap values for -ing and -ed inflections.

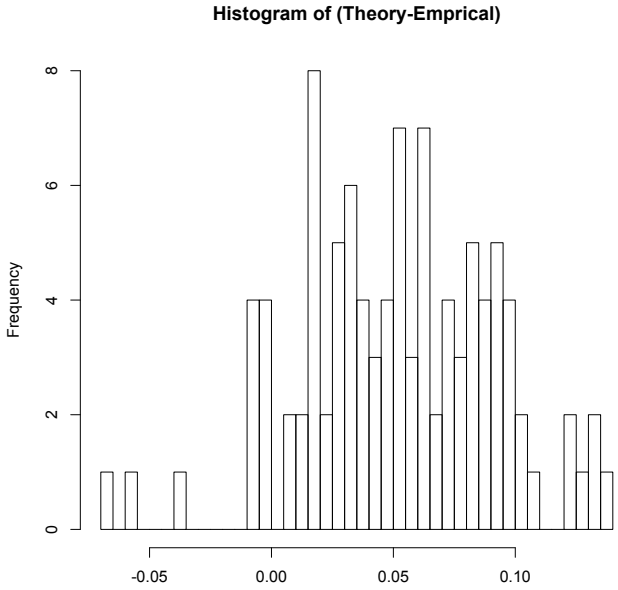

Figure 2: Overlap test applied to linguistic combinations with lexical exceptions.

may take two affixes A and B: A can attach to all stems but B can only attach to 90 while the other 10, randomly chosen from the 100, are exceptions. Again, we assume that frequencies of the stems and their combinations with affixes follow Zipfian distribution. We random combine stems with affixes 1000 times obtaining a sample size of 1000, and count the percentage of stems that are combined with both A and B. We then compare this value against the calculation from (2) which assumes A and B are fully interchangeable (where in this case they are not). The histogram of the difference between the theoretical and empirical values from 100 such simulations are given in Figure 3. The overlap test correctly overpredicts ($p < 10^{-15}$).

4 Discussion

For the study of child language acquisition, our results show that the usage-based approach to language learning is not supported by the child data once the statistical properties of linguistic units and their combinations are taken into account. A grammar based approach is supported (section 3.1) These results do not resolve the innateness debate in language acquisition: they only point to the very early availability of an abstract and productive grammar.

The simulation results on the inadequacy of the memory-and-retrieval approach to child language (section 3.2) show the limitations of lexically specific approach to language learning. These results are congruent with the work in statistical parsing that also demonstrates the diminishing returns of lexicalization (Gildea 2001, Klein & Manning 2003, Bikel 2004). They are also consistent with previous statistical studies (Buttery & Korhonen 2005) that child directed language data appear to be even more limited in syntactic usage diversity. The "uneveness" in verb islands (Tomasello 1992) is to be expected especially when the language sample is small as in the case of most child language acquisition studies. It thus seems necessary for the child learner to derive syntactic rules with overarching generality in a relatively short period of time (and with a few million utterances).

Finally, we envision the overlap test to be one of many tests for the statistical properties of grammar. Similar tests may be constructed to include a wider linguistic context (e.g., three or more words instead of two, but the sparse data problem becomes far more severe). The ability to detect lexicalized processes (section 3.3) may prove useful in the automatic induction of grammars. Such tests would be a welcome addition to the quantitative analysis tools in the behavioral study of language, which tend to establish mismatches between observations and null hypotheses; the favored hypotheses are those that cannot be rejected (though cannot be confirmed either). The present work shows that it is possible to test for statistical matches between observations and well formulated hypotheses.

References

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.

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.

Buttery, P. & Korhonen, A. (2005). Large-scale analysis of verb subcategorization differences between child directed speech and adult speech. Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarland University.

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

Feldman, J. (2004). Computational cognitive linguistics. In COLING 2004.

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.

Gildea, D. (2001) Corpus variation and parser performance. In 2001 Conference on Empirical Methods in Natural Lan- guage Processing (EMNLP).

Goldberg, A. (2003). Constructions. *Trends in Cognitive Science*, 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.

Klein, D. & Manning, C. (2003). Accurate unlexicalized parsing. In ACL 2003. 423-430.

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

Lieven, E., Pine, J. & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, 24, 187-219.

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

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

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

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

Sag, I. (2010). English filler-gap constructions. *Language*, 486–545.

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.

Steels, L. (2004). Constructivist development of grounded construction grammars. In ACL 2004.

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

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.

Wintner, S. (2009). What science underlies natural language engineering. *Computational Linguistics*, 641–644.

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