J. Child Lang. **37** (2010), 671–703. © Cambridge University Press 2010
doi:10.1017/S0305000910000024

An empirical generative framework for computational modeling of language acquisition*

HEIDI R. WATERFALL

Department of Psychology, Cornell University and Department of Psychology, University of Chicago

BEN SANDBANK

School of Computer Science, Tel Aviv University

LUCA ONNIS

Department of Second Language Studies, University of Hawaii

AND

SHIMON EDELMAN

Department of Psychology, Cornell University and Department of Brain and Cognitive Engineering, Korea University

(*Received* 22 *December* 2008 – *Revised* 10 *August* 2009 – *Accepted* 24 *December* 2009)

ABSTRACT

This paper reports progress in developing a computer model of language acquisition in the form of (1) a generative grammar that is (2) algorithmically learnable from realistic corpus data, (3) viable in its large-scale quantitative performance and (4) psychologically real. First, we describe new algorithmic methods for unsupervised learning of generative grammars from raw CHILDES data and give an account of the generative performance of the acquired grammars. Next, we summarize findings from recent longitudinal and experimental work that suggests how certain statistically prominent structural properties of child-directed speech may facilitate language acquisition. We then present a series of new analyses of CHILDES data indicating that the desired properties

[*] During the preparation of this paper, Shimon Edelman was partially supported by the WCU (World Class University) program through the National Research Foundation of Korea, funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0). Address for correspondence: Heidi R. Waterfall, Department of Psychology, 211 Uris Hall, Cornell University, Ithaca, NY 14853 USA. Tel: 607 229 2061. Fax: 607 255 8433. E-mail: heidi.waterfall@gmail.com

are indeed present in realistic child-directed speech corpora. Finally, we suggest how our computational results, behavioral findings, and corpus-based insights can be integrated into a next-generation model aimed at meeting the four requirements of our modeling framework.

INTRODUCTION AND OVERVIEW

In psycholinguistics, the main challenge is to discover the nature of grammar – the knowledge of language as it is represented in the brain – subject to the constraint that it be simpler than the corpus of linguistic experience from which it is induced and which it in turn explains (Chomsky, 1957). A concomitant challenge in developmental psycholinguistics and computational linguistics is to discover the algorithms – precisely and exhaustively specified computational procedures – through which grammar is constrained by the learner's innate biases and shaped by experience.

The conceptual framework for developmental computational psycholinguistics adopted in this paper construes language acquisition by infants as an interplay of (i) the information available in the regularities in the corpus that the learner encounters as a part of its socially situated experience and (ii) the probabilistic structure discovery algorithms charged with processing that information. In the modeling of language acquisition, work on algorithmic techniques for grammar induction has often advanced independently of the traditional psycholinguistic characterization of the role of corpus regularities in acquisition (e.g. Solan, Horn, Ruppin & Edelman, 2005). Consequently, one of our main goals in the present paper is to argue for a greater integration of computational and behavioral developmental psycholinguistics.

In this paper, we lay the foundation for a novel framework for modeling language acquisition, which calls for the creation of (1) a fully generative grammar that is (2) algorithmically learnable from realistic data, (3) that can be tested quantifiably and (4) that is psychologically real. In the first section of this paper, we motivate these four requirements. Second, we detail our progress to date in creating algorithms that acquire a generative grammar from corpora of child-directed speech. In particular, it outlines a new algorithm for grammar induction, and states its performance in terms of recall and precision. Third, we address the requirement of learnability from realistic data and summarize recent findings from observational studies and artificial language studies which examine variation sets, a highly informative cue to language structure. In the fourth section, we offer an extensive analysis of a large subset of the English CHILDES corpus in terms of variation set structure. Finally, we discuss potential future research in which our results could be integrated into a comprehensive computational

model of language acquisition that would meet all four requirements stated above.

The need for a generative grammar

Much of the effort in the computational modeling of language development focuses on understanding specific phenomena, such as word segmentation (e.g. Batchelder, 2002). More general and abstract work typically involves computational models capable of learning certain classes of formal languages generated by small artificial grammars (e.g. Christiansen & Chater, 2001; Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996). A considerably more ambitious goal, and the only one commensurate with the achievement inherent in language acquisition by human infants, is to develop an algorithm capable of learning a grammar that is GENERATIVE of the target language, given a realistic corpus of child-directed speech (normally supplemented by a plethora of cues stemming from the embodiment and social situatedness of language, which are outside the scope of the present paper).

Formally, a grammar is generative of a language if it is capable of producing all and only the acceptable sentences in it (Chomsky, 1957). The growing realization in psycholinguistics that acceptability judgments offered by subjects are better described as graded rather than all-or-none (e.g. Schütze, 1996) is spurring an ongoing revision of the classic notion of generativity. The emerging modified version requires that the grammar reproduce the natural probability distribution over sentences in the linguistic community (instead of a binary parameter that represents idealized 'grammaticality'). This approach (e.g. Goldsmith, 2007) is compatible with the standard practice in natural language processing (NLP, a branch of artificial intelligence), where one of the goals of learning is to acquire a probabilistic language model (e.g. Chater & Vitányi, 2007; Goodman, 2001; other NLP tasks for which effective learning methods have been developed, such as word sense disambiguation or anaphora resolution, are only indirectly related to generativity and are therefore of less relevance to the present project). A language model is a probability distribution over word sequences (i.e. partial and complete utterances). Given a partial utterance, a language model can be used to estimate the probabilities of all possible successor words that may follow it. One family of such models that is generative in the required sense is the Probabilistic Context Free Grammar (PCFG), discussed later in this paper.

The need for a realistic corpus of language data

In order to integrate computational and behavioral approaches to language acquisition, it is necessary to use a common corpus of recorded language

that is natural enough to support valid psycholinguistic research and extensive enough to afford automatic learning of grammar, the algorithms for which tend to be data-hungry. A large and growing collection of corpora that is getting better and better at meeting both requirements is freely available from the CHILDES repository (MacWhinney & Snow, 1985; MacWhinney, 2000).

The need for an empirical evaluation of the resulting grammar

As any other science, psycholinguistics is expected to garner empirical support for the conceptual structures – in the present case, the grammar, whether acquired algorithmically or constructed by hand – in terms of which it attempts to explain its primary data. Good scientific practice requires, therefore, that the following two questions be addressed: (1) How well does the constructed or learned grammar perform? (2) Are the structures posited by the grammar psychologically real?

Measuring the performance of a grammar

A grammar that has been induced from a corpus of language must be capable of accepting novel utterances not previously encountered in that corpus; moreover, any utterances that it generates must be acceptable to native speakers of the language in question. (Here we intentionally gloss over the distinction between acceptability and grammaticality.) This consideration suggests that the generative performance of a grammar could be measured by two figures: RECALL, defined as the proportion of unfamiliar sentences that a parser based on the grammar accepts, and PRECISION, defined as the proportion of novel sentences generated by the grammar that are deemed acceptable by native-speaker subjects, preferably in a blind, controlled test (Solan *et al.*, 2005). These definitions of recall and precision are related but not identical to those used in NLP (Klein & Manning, 2002). Specifically, most acquisition-related work in NLP is concerned with learning manually defined gold-standard tree structures and thus measures recall and precision in terms of the proportion of correct constituent trees relative to such a standard; in comparison, we focus on the acceptability of entire utterances. Furthermore, we insist on excluding training utterances from this evaluation, so as better to assess the innovative generative performance of the learned grammars. Thus, our recall and precision figures tend to be more conservative than those in the literature.

It is worth noting that neither high recall nor high precision suffices on its own: it is trivially easy to construct a grammar that has recall of 1 (the grammar should simply accept all possible sequences of words) or conversely precision of 1 (the grammar should simply consist of the test corpus); the

precision of the former grammar and the recall of the latter will be very poor. A perfectly performing grammar will have both precision and recall of 1 on a convincingly large, statistically representative test corpus. Furthermore, as mentioned above, an ideal probabilistic language model associated with the grammar that is under evaluation should approximate closely the probability distribution over utterances that prevails in the language community (a COMPLETE situated language model would be able to approximate the joint probability over utterances and behavioral/social contexts).

Assessing the psychological reality of a grammar

A final requirement of an ideal probabilistic language model is that it should be psychologically real. It is conceivable that a grammar for a particular language could do well on all the dimensions of performance mentioned above, and yet rely on structures (e.g. rewriting rules) that have no counterpart in the brains of the speakers of that language (cf. Chomsky, 1995). Thus, if the goal of a study is to develop a cognitively valid model of grammar rather than a descriptively adequate one, it is important to ask whether the structures constituting a grammar have a grounding in psychological (and, ultimately, neurobiological) reality.

The question of psychological reality applies not only to the structures posited by a grammar but also to the mode of their acquisition and subsequent use in processing language. We note that the grammars induced by the algorithms mentioned later in this paper have not been vetted for psychological reality, nor do we address issues of language processing. Insofar as psychological reality is concerned, the purpose of the present paper is to serve merely as an illustration of the feasibility of unsupervised learning of high-performance generative grammars from realistic, unannotated corpus data.

Effective learning of psychologically real grammars from naturalistic corpus data

We now briefly summarize the main tenets of the proposed framework for modeling language acquisition, which is still under construction: its goal is to specify a grammar that is (1) fully generative, (2) algorithmically learnable from realistic data, (3) quantifiably successful and (4) psychologically real. In the rest of this paper, we describe working algorithms capable of unsupervised grammar induction from raw CHILDES data, which constitute significant progress in achieving the first three of these objectives. We also describe a set of novel quantitative analyses of CHILDES data that suggest how the new algorithmic techniques can be

made to model language acquisition in children, thus laying the groundwork for meeting objective #4.

ALGORITHMIC INDUCTION OF GENERATIVE GRAMMARS FROM TRANSCRIBED CHILD-DIRECTED SPEECH DATA

In this section, we approach the problem of grammar induction from first principles (motivated by the insights of Zellig Harris, 1954).

The task

Grammar induction algorithms typically assume that natural sentences are generated from some well-defined probabilistic distribution (a common assumption in the field of machine learning; cf. Valiant, 1984). The grammar induction task is to infer the underlying sentence distribution from a (potentially large) sample of sentences, called the training corpus. (All grammar induction algorithms implemented to date treat sentences as independent of each other, ignoring supra-sentential discourse structure. As we show later in this paper, this assumption throws away valuable information that can in principle be used to boost learning – such as variation sets, which naturally occur in child-directed speech, described below.)

This inference would be impossible without assuming some restrictions on the class of possible distributions of those sentences. Several such classes have been studied in the past. The class that is important to us, which seems to capture a substantial portion of natural language phenomena, is PROBABILISTIC CONTEXT FREE GRAMMARS (PCFGs), a probabilistic extension of the classic context free grammars. Most of the early work in this field produced algorithms that were demonstrated to work only for very small corpora generated by simple artificial grammars (e.g. Stolcke & Omohundro, 1994; Wolff, 1988). More recently, proofs of convergence of the learning algorithm to the correct grammar given certain constraints on the training corpus were published for certain subclasses of PCFGs (e.g. Adriaans, 2001; Clark, 2006).

Given the present framework's focus on psychological reality, the approaches to grammar acquisition that are of most interest to us are those that work in a completely unsupervised fashion on completely unannotated corpora – that is, algorithms that start with no explicit knowledge of potential structures and no data beyond the raw text or transcribed speech. Most existing algorithms for grammar induction have not been designed or tested for operation that is realistic in that sense (e.g. the highly successful algorithm of Klein and Manning (2002) learns structures from data annotated for part of speech information). A most notable exception in this respect is the Unsupervised Data-Oriented Parsing (U-DOP) algorithm developed by

Bod (2009). The DOP approach uses the tree-substitution grammar formalism, representing the structure of a novel sentence in terms of probabilistically weighted structural analogies to trees gleaned from a training corpus. In the unsupervised version, these trees are obtained by simply listing all the possible binary tree descriptions of sentences in the training corpus. As reported by Bod (2009), the U-DOP algorithm performs well in the task of learning a grammar from CHILDES data annotated with part of speech information, as assessed by comparing the structures it induces to those from a hand-annotated gold-standard syntactic parse of the corpus (its performance on raw CHILDES data is somewhat lower). The resulting grammar has been shown capable of replicating a number of syntactic phenomena long considered to be central to language acquisition (Bod, 2009).

With the exception of U-DOP and the ADIOS algorithm (Solan *et al.*, 2005; see below), none of the previously published algorithms for grammar acquisition were shown to scale up well to raw natural language corpora. Furthermore, no published algorithm except ADIOS had its performance assessed on the generativity dimensions of entire-sentence unseen-corpus precision and recall for realistic corpora such as CHILDES. We describe the ADIOS algorithm briefly in the next section.

The ADIOS algorithm

ADIOS (for Automatic Distillation Of Structure) is a fully unsupervised algorithm for grammar inference from unannotated text data (Solan *et al.*, 2005). The ADIOS algorithm rests on two principles: (1) probabilistic inference of pattern significance and (2) recursive construction of complex patterns. ADIOS starts by representing a corpus of sentences as an initially highly redundant directed graph, in which the vertices are the lexicon entries and the paths correspond to corpus sentences.

The graph can be informally visualized as a tangle of sentences (i.e. paths) that are partially segregated into bundles (i.e. two or more sentences containing the same word or words). The bundle unravels when the sentences diverge – that is, contain different words. In a given corpus, there will be many bundles, with each sentence possibly participating in several. The algorithm iteratively searches for SIGNIFICANT bundles (i.e. collocations) using a simple context-sensitive probabilistic criterion defined in terms of local flow quantities in the graph (cf. Solan *et al.*, 2005). A distinctive feature of ADIOS is that it only admits equivalence classes (e.g. lexical categories, verb phrases, etc.) that appear inside statistically significant collocations.

The ADIOS graph is rewired every time a new pattern (collocation and/or equivalence class) is detected, so that a bundle of element sequences subsumed by it is represented by a single new vertex or node. Following

the rewiring, which is specific to the context in which the pattern was discovered, potentially far-apart symbols that used to straddle the newly abstracted pattern become close neighbors. Patterns thus become hierarchically structured in that their elements may be either terminals (i.e. fully specified strings) or non-terminals (i.e. partially specified strings that include some variables). The ability of new patterns and equivalence classes to incorporate those added previously leads to the emergence of recursively structured units that support generalization. Moreover, patterns may refer to themselves, which opens the door for true recursion.

The main goal of the ADIOS project was to test the ability of unsupervised grammar induction methods to learn from realistic large-scale corpora. To that end, the ADIOS algorithm has been tested on corpora generated by large artificial context-free grammars, as well as on natural language corpora of moderate size, achieving impressive scores on the precision (0·63) and recall measures (0·50) (on a portion of the English CHILDES; comparable performance was achieved on the Mandarin CHILDES corpora; cf. Brodsky, Waterfall & Edelman, 2007).

The ConText algorithm

ConText, a much simpler algorithm developed in response to ADIOS, operates directly on the distributional statistics of the corpus and characterizes words and phrases by the local linguistic contexts in which they appeared. Distributional statistics, a major cue in language acquisition, were also instrumental for the automatic acquisition of syntactic categories (Redington, Chater & Finch, 1998), the grouping of nouns into semantic categories (Pereira, Tishby & Lee, 1993), unsupervised parsing (Clark, 2001; Klein & Manning, 2002) and text classification (Baker & McCallum, 1998).

In ConText, the distributional statistics of a word or a sequence of words (w) are determined by the surrounding words (i.e. local context). The width of this local context, L, is a user-specified parameter, set in most of our experiments to be two words on either side of w. To calculate the distributional statistics of w, ConText constructs its left and right context vectors. For each L that appears in the corpus, there is a corresponding coordinate in the left (right) context vector indicating how many times it appears to the left (right) of w. The left and right context vectors are then concatenated to form a single vector representation of the context distribution of w. ConText constructs these context vectors for each word sequence that occurs more than K times in the corpus, K being another user-specified parameter.

Thus, like ADIOS, ConText aligns word sequences to perform DISTRIBUTIONAL CLUSTERING of progressively more complex structures. Like

ADIOS, this distributional clustering creates equivalence classes. The distance between two word sequences is defined as the distance between their corresponding vectors. Sequences that are closer than D (a user-defined parameter) are viewed as equivalent and are clustered together. At the end of the clustering procedure, sequences belonging to the same cluster are assumed to be substitutable or equivalent.

In preliminary experiments, we found that choice of clustering procedure made little or no difference in the performance of the algorithm. Therefore, ConText uses the following simple clustering scheme. First, it selects a sequence and iteratively compares it to subsequent sequences. Whenever another sequence is found whose distance to the current one is smaller than D, ConText adds it to the current cluster and removes it from the list of sequences. If no such sequence is found, the current sequence is maintained as a lexical item or string of lexical items. This process repeats until the list of possible sequences is empty. The order of the sequences in the list was also found to have little effect and so they are randomly ordered.

In contrast to the clustering procedure, the choice of distance metric did alter the algorithm's performance. During the development of ConText, we empirically surveyed a number of measures commonly used in NLP (e.g. Lee, 1999) on artificial corpora, and found that the angle between context vectors provided the best results. This is the measure we use to assess recall and precision, described below.

For each cluster found using the above procedure, a new non-terminal symbol is introduced into the current grammar, along with rules that rewrite it as each of the sequences in the cluster. All occurrences of these sequences in the corpus are replaced with the non-terminal symbol. The probabilities assigned to each of its rewrite rules correspond to the frequencies of each of the corresponding sequences in the training corpus. This process is repeated until no new clusters are formed.

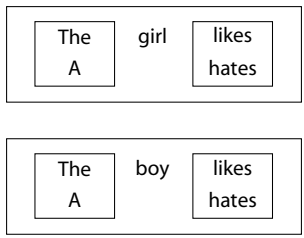
A single iteration of the algorithm on a very simple corpus is demonstrated in Figure 1. Figure 1(1) presents the initial grammar – that is, the sentences that appear in the training corpus. Figure 1(2) shows two sequences that appear at least three times in the corpus. For each sequence, the tokens to its immediate left and right are presented. Lastly, Figure 1(3) displays the new non-terminals and rewrite rules associated with each cluster, along with their associated frequencies.

Performance of ConText on CHILDES: recall

Out of the large English CHILDES collection, we chose eight corpora collected in relatively naturalistic settings (Bloom, 1970; 1973; Brown, 1973; Demuth, Culbertson & Alter, 2006; Hall, Nagy & Linn, 1984; Hall, Nagy & Nottenburg, 1981; Hall & Tirre, 1979; Higginson 1985;

(1)
S → The girl likes to play # 0.16
S → The boy hates to run # 0.16
S → A girl hates to play # 0.16
S → A boy likes to run # 0.16
S → A girl hates to run # 0.16
S → The boy likes to play # 0.16

(2)



(3)
S → D A B to C # 1.0
A → girl 0.5
A → boy 0.5
B → hates 0.5
B → likes 0.5
C → run 0.5
C → play 0.5
D → A 0.5
D → The 0.5

Fig. 1. Sample ConText grammar of a miniature artificial language. (1) The grammar is initially capable of reproducing only the sentences in the input. (2) The algorithm clusters sequences based on their distributional properties. (3) The final grammar consists of rewrite rules with their associated probabilities.

MacWhinney, 1995; Sachs, 1983; Suppes, 1974). Together, these totaled 349,187 adult utterances, comprising 2,426,578 word tokens. The resulting corpus was used for the present experiment in grammar induction, as well as for a set of detailed analyses of its discourse structure (described below).

For this grammar induction experiment, the corpus was split into three sets: a 300,000 sentence training set (203,405 unique sentences), a 15,000 sentence test set and a set of 34,000 sentences used for tuning the parameters of the learning algorithm, as described below. Each utterance in the corpus was regarded as a separate unit. Some utterances correspond to complete sentences (e.g. *Mommy's tired, honey*) and some to noun phrases, etc. (e.g. *the red ball*). The test set did not include any of the sentences in the training set and was used to calculate the resulting grammar's recall.

As noted above, ConText has two main parameters influencing its operation: K, the minimum number of times sequences must appear in the training corpus in order to participate in clusters, and D, the maximum distance between two sequences that can still be clustered together. Pilot data indicated that K has little influence on the performance of the algorithm, as long as it is not set too high. For the present study, we set it to 50. However, the algorithm is sensitive to the value of the parameter D, which provides a means to trade off between recall and precision: the higher D is, the more the algorithm generalizes – hence, the higher the recall of the resulting grammar and the lower its precision. To choose a working value for D, we conducted several runs of the algorithm, each with a different value of D, starting from 0·3 and moving up to 0·7 in 0·05 increments. After each run, we informally assessed the resulting grammar's precision on 100 generated sentences (a final assessment of precision using judgments provided by human subjects is described later). The highest value for D that still yielded a precision level above 0·5 was selected. This was attained for $D=0{\cdot}65$. (For further details on the effects of parameter settings on the performance of ConText, see the Appendix.)

The initial grammar, prior to learning, contained a single non-terminal 'S' and 203,405 rewrite rules, one for each unique sentence in the training set. Hence, this grammar could produce all of the sentences in the training set, and none other. After learning, using the final settings for K and D reported above, the non-terminal 'S' participated in 169,228 rewrite rules, including 1,202 equivalence classes (e.g. noun phrases, verb phrases, etc.). Thus, ConText meets our requirement that the grammar derived from the data be simpler than the corpus itself: the total size of the grammar reflected a compression (i.e. generalization) of roughly 17% of the original corpus. Note that this grammar contains a large number of rewrite rules, compared to the textbook notions of syntax (but not when compared to the grammar learned by U-DOP from comparable data; cf. Bod, 2009). This may result from several causes. First, the 0·22 recall suggests that learning uncovered only a small proportion of the structural regularities present in the corpus. A more sophisticated approach to learning (e.g. one that uses discourse structure found in natural caregiver–child interactions, as suggested in later sections) applied to larger training corpora may lead to smaller final grammars. Second, it is possible that the PCFG formalism does not capture well the regularities inherent in language (e.g. Joshi & Schabes, 1997).

Figure 2 provides some examples of equivalence classes inferred by ConText from CHILDES corpora. The figure reveals that ConText is sensitive to both syntactic distributions (i.e. word classes), but also, indirectly, to word meanings. Thus, *drink* and *eat* are clustered together under the equivalence class E32 because of their highly similar local contexts (L), but other verbs are not. Similarly, equivalence class E23 can be rewritten as

Rule	Equivalence classes
E11 →	we | well we
E15 →	bowl | refrigerator | oven | house |
	mirror | country | corner | sky |
	basket | living room | kitchen | barn |
	bath tub | snow | closet | carriage |
	world | box | bag | bedroom | car | sink
	air | water | movie | forest | sand | drawer
E18 →	am I | is he | is she | were you |
	were they | are you | are they
E23 →	wonderful | neat | great | good
E32 →	eat | drink
E68 →	warm | hot | cold
E89 →	to the bath room | to the store |
	to bed | to sleep
E103 →	choo | choo choo | choo choo choo
E104 →	hold on | listen
E943 →	the boy | your little girl | santa-claus …

Fig. 2. Sample ConText equivalence classes. The grammar generates new sentences by selecting elements chosen from equivalence classes and putting them into novel contexts.

a number of positive adjectives, but not other adjectives. The equivalence class E15 seems to be an exception to this, clustering together a wide variety of nouns without an obvious semantic relationship. Higher-level equivalence classes provide a glimpse of the grammatical structure inferred by ConText. Some of those correspond to classical syntactic constituents (e.g. E943, of which only a small portion is shown, roughly corresponds to NP, complete with an optional adjective preceding the noun). Note also that at this level of generalization, the semantic similarity between the nouns is far less pronounced than at lower levels. Some equivalence classes, such as E89, correctly classify multiword sequences as substitutable, yet treat the sequence as an unanalyzable whole, without inferring its internal structure.

This is probably because this multiword phrase is idiomatic in the training corpus. While a more thorough syntactic analysis is possible, it is not warranted by the corpus in the sense that other syntactically plausible variants (e.g. *to the bedroom*) do not appear.

Lastly, we note that many equivalence classes seem to cross traditional constituent boundaries. This is not surprising, as ConText was not designed with the purpose of inferring constituent boundaries as such. Indeed, in order to optimize its ability to infer a highly generative grammar, we elected not to use a constituency test. Although previous work suggests that such a criterion is beneficial when using distributed clustering for unsupervised parsing (Clark, 2001; Klein & Manning, 2001), we have found that using such a criterion severely interferes with the resulting grammar's generative capacity, which was our goal here.

Performance of ConText on CHILDES: precision

To estimate the precision performance of the grammar learned by ConText, we asked participants to judge the acceptability of 100 sentences generated by ConText, mixed randomly with 100 sentences from the CHILDES corpus.

Participants. Fourteen students, all native speakers of English at the University of Hawaii, were paid $5.

Methods. Participants were asked to rate, on a scale from 1 to 7 (7 being perfectly plausible), how likely each target utterance was to appear in normal child-directed speech. We used the R procedure LMER (e.g. Baayen, 2006) to fit a mixed-effects linear model to the ratings produced by participants, with Subject and Item as random effects and Source (CHILDES/ConText) and Phrase Length as fixed effects. Phrases up to ten words long were included in the analysis (beyond that length, the data were very sparse).

Results. The Source effect was significant at $p < 0{\cdot}033$; the interaction between Source and Phrase Length was highly significant at $p < 0{\cdot}0001$ (all significance estimates were obtained by the MCMC procedure; Baayen, 2006). The original CHILDES sentences were better on average (6·12 on a scale of 1–7, $SD = 1{\cdot}64$) compared to ConText (4·99, $SD = 2{\cdot}30$) (see Figure 3). This is especially true for longer phrases – our results suggest that speakers' judgments of ConText sentences were similar to those for CHILDES sentences for Phrase Length between 1 and 6, beyond this ConText increasingly often failed to generate grammatical utterances.

Grammar induction: concluding remarks

We have presented two grammar induction algorithms and demonstrated their usefulness for inferring a generative grammar for English given a

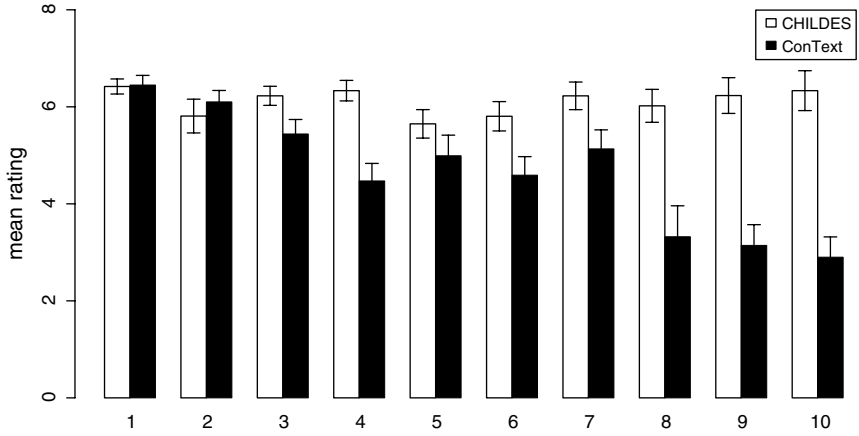


Fig. 3. Mean plausibility ratings plotted against utterance length for a sample of 100 caregiver utterances (CHILDES) and 100 utterances generated by a ConText-learned grammar. Error bars denote 95% confidence intervals.

corpus of child-directed speech. ConText, by far the simpler of the two, was developed to overcome the shortcomings of its predecessor, and was demonstrated to outperform it on a variety of corpora. On the subset of CHILDES that we studied, the performance of the two algorithms seems comparable. We believe that in order to achieve qualitative improvements in grammar induction performance, statistical cues present in child-directed speech but not currently utilized by the existing algorithms, such as temporal ordering of sentences, will have to be used. For example, ConText and ADIOS both have access to the entire corpus when making clustering decisions. Taking advantage of context-local discourse structure such as variation sets (described below) could lead to the development of algorithms that are incremental and therefore more psychologically real.

VARIATION SETS AND LEARNABILITY

Having summarized the findings of two algorithms that apply alignment and comparison to child-directed speech, we now describe observational and experimental studies relevant to our goal of developing more realistic algorithms for grammar induction. Specifically, we examine a property of spontaneous caregiver speech – the use of VARIATION SETS – and its implications for both language acquisition and language-learning algorithms. Variation sets are clusters of caregiver utterances occurring

within a conversational turn that share some (but not all) lexical items and structures:

(1) Mother 12, talking to her child aged 1;2, pushing dolls in a stroller:

You got to	*push them*	*to school.*
	Push them.	
	Push them	*to school.*
	Take them	*to school.*
You got to	*take them*	*to school.*

These adjacent utterances in caregiver input exhibit some of the properties that could be used by the processes of language structure induction outlined by Harris (1954) – namely, the ALIGNMENT of repeated parts of the utterances (e.g. *you got to; push them, take them, to school*), which in turn may facilitate the COMPARISON of those utterances. These naturally occurring groupings of related utterances are a well-known property of how parents talk to young children, yet they have received relatively little attention. Waterfall (2006; submitted) was the first to investigate variation sets longitudinally and their relation to child language. Here, we summarize the relevant results of those observational studies and briefly discuss the results of subsequent artificial language studies that also investigate variation sets (Onnis, Waterfall & Edelman, 2008).

Previous research on variation sets

The presence of partial repetitions/overlapping utterances in caregiver speech to young children has been known for quite some time (e.g. Brown, Cazden & Bellugi, 1969). These types of utterances have also been found to be relatively common in child-directed speech: roughly 20 percent of utterances in caregiver speech are in variation sets (e.g. Küntay & Slobin, 1996; Waterfall 2006; submitted). Variation sets, which involve a caregiver's partial self-repetitions, are to be distinguished from caregiver–child interaction, where the caregiver expands, repeats or corrects a previous child utterance (e.g. Sokolov 1993; Stine & Bohannon, 1983). That type of interaction, while important for language acquisition, typically involves older children (e.g. Sokolov 1993; Stine & Bohannon, 1983). Variation sets, however, occur in caregiver speech before the child starts speaking (e.g. Waterfall, 2006) and their study thus complements previous expansion research. Moreover, while expansions are necessarily related to child speech, caregivers can use structures in variation sets that the child has yet to produce and thus may lead child production of those structures (Waterfall, submitted). Similar to caregiver expansions, variation sets change over time: the proportion of speech in variation sets decreases as children age (Waterfall, 2006; submitted), and this decrease may be related to the rise in caregiver–child expansion interactions.

Partial self-repetitions in the speech of parents may be important in language development, although previous studies of partial repetitions typically also include expansions. Nelson (1977) experimentally manipulated input to toddlers by increasing the number of partial repetitions of adult utterances and expansions. This kind of input had a positive effect on children's production of questions with inverted auxiliaries. Hoff-Ginsberg (1985) found that alternations in maternal self-repetitions and expansions that conformed to major constituent boundaries were related to growth in children's verb use while those repetitions and expansions that altered material within a phrasal constituent aided noun phrase growth. Later studies confirmed that the frequency of self-repetitions and expansions was positively correlated with verb phrase development (e.g. Hoff-Ginsberg, 1990).

Although many researchers have noted the presence of partial repetitions in caregiver speech, Küntay & Slobin (1996) were the first to systematically examine the real-time ordering of these overlapping utterances, introducing the term 'variation sets' in their study of child-directed speech in Turkish. They hypothesized that variation sets may be useful to children for acquiring verbs and verb subcategorization frames. Waterfall (2006; submitted) investigated variation sets in English-speaking caregiver–child dyads, exploring these hypotheses.

Waterfall (2006; submitted) examined longitudinal use of variation sets by caregivers and their relation to children's development of vocabulary and syntax. The speech of twelve English-speaking caregiver–child dyads was analyzed, starting when the children were 1;2 and continuing to 2;6. Families were controlled for child birth order (six first-borns), child gender (six girls) and maternal educational level (four high-school graduates, four college graduates and four mothers with graduate degrees). Families were visited every four months for a total of five observations. Dyads were observed interacting naturally in their homes for 90 minutes.

When caregiver and child speech from the same observation was analyzed, Waterfall (2006; submitted) found that more of children's noun and verb types were related to variation sets than to ordinary child-directed speech, even when the frequency of the lexical items themselves was accounted for. Thus, variation sets seem to be related to contemporaneous child speech. When caregiver speech from earlier observations was compared to child speech at later observations, caregiver verb use from earlier variation sets was significantly correlated with later child production of verbs, indicating that variation sets are predictive of child verb production. Further, caregivers' earlier variation of certain syntactic structures in variation sets (e.g. syntactic subjects and direct objects) was predictive of later child production of those structures. These findings suggest that variation sets not only contribute to the acquisition of specific lexical items, but may also

facilitate the acquisition of lexical classes as well as larger syntactic structures.

Lastly, Onnis, Waterfall & Edelman (2008) investigated the role of variation sets in acquisition by manipulating their distribution in two artificial language experiments with adults. In each experiment, there were two conditions: a VarSet condition in which 20 percent of the data were presented in variation sets – a proportion based on observational data (Waterfall 2006; submitted); and a Scrambled condition in which there were no adjacent utterances with lexical overlap. In both conditions, participants received the exact same input: only the order of presentation differed. In Experiment 1, when given a forced-choice task, participants in the VarSet condition identified words more successfully than participants in the Scrambled condition. In Experiment 2, when given a forced-choice task, participants in the VarSet condition were more successful at identifying phrases than those in the Scrambled condition. While not directly comparable to child language acquisition, these experiments suggest that variation sets may aid the acquisition of words and multiword phrases.

The results of the observational and artificial language studies seem to suggest that Harris' initial suppositions on the role of alignment and comparison – the two key computational operations behind both the ADIOS and the ConText algorithms – were not only correct in the abstract context of grammar discovery but also applicable to everyday interactions between caregivers and children. Below, we support this conclusion by conducting a large-scale statistical analysis of variation set structure in select English CHILDES corpora.

STATISTICAL CUES TO STRUCTURE AVAILABLE IN TRANSCRIBED CHILD-DIRECTED SPEECH

Computational definition of variation sets

Variation sets can be defined according to contextual and linguistic criteria as well as computational criteria. Few computational studies have examined variation sets to date (Brodsky *et al.*, 2007; Sokolov & MacWhinney 1990). Brodsky *et al.* (2007) proposed the following definition: a variation set is a contiguous sequence of utterances produced by a single speaker in a conversation and each successive pair of utterances has a lexical overlap of at least one element (excluding a few highly frequent words and clitics*: a, an, the, 'll 'm, 're, 's, 't, 've, um*). According to this definition there is no need for the same word to appear in each of the member utterances in a variation set. Although this computational definition differs from the one employed by Waterfall (2006; submitted), the resulting percentage of utterances in variation sets in the present large-scale corpus study (about 20 percent) was similar.

The above definition of variation set may be extended by allowing for intervening utterances (or 'gaps') between utterances sharing a lexical item. Later in this section, we state the results for values of gap ranging from 0 (strictly adjacent utterances) to 2 (up to two intervening utterances).

Optimal variation

Computational work to date has focused on exploring broad characteristics of variation sets (e.g. Brodsky *et al.*, 2007), in particular the (dis)similarity between utterances within variation sets in terms of the Levenshtein (edit) distance (Ristad & Yianilos, 1998). The edit distance between two sentences is defined as the number of elementary edit operations (insertion, deletion or substitution of individual words) needed to transform one sentence into the other. Brodsky *et al.* (2007) found that the Levenshtein edit distance was significantly smaller for utterances in variation sets than for adjacent utterances not in variation sets. In addition, variation sets were analyzed for their information value – that is, the amount of structure that could be determined by analyzing a pair of sentences. Completely non-overlapping utterances or exact repetitions are not informative: there is nothing to compare or contrast. In contrast, a pair of utterances that combines some repetition with some change can be informative. The authors found that variation set use by caregivers in the Waterfall corpus (2006; submitted) was correlated with child vocabularies and that those variation sets with a novelty value of 0·487 (defined in information-theoretic terms) were the most strongly correlated with child vocabularies. This suggests that those variation sets where roughly 50 percent of the material changes may be optimally informative for children.

Variation sets in CHILDES

Using the above definition of variation sets, we conducted a set of novel analyses of caregiver speech from the CHILDES database. We chose the same eight naturalistic corpora selected for the ConText experiments (Bloom, 1970; 1973; Brown, 1973; Demuth *et al.*, 2006; Hall *et al.*, 1984; Hall *et al.*, 1981; Hall & Tirre, 1979; Higginson, 1985; MacWhinney, 1995; Sachs, 1983; Suppes, 1974). We calculated the proportion of caregiver speech in variation sets and the average length of variation sets (in utterances). We then performed two sets of structural analyses on variation sets. The first set of analyses seeks to determine characteristics of variation sets derived from the data – namely, the most frequent bi-grams and tri-grams (i.e. two- and three-word sequences) in variation sets and their position within utterances. The second set of analyses seeks to determine whether or not variation sets are informative with respect to specific linguistic phenomena (i.e. dative alternation and clausal complementation).

Methods: statistical properties of variation sets and analysis of most frequent n-grams. We first identified all the variation sets in caregiver speech for our corpora and computed five principal statistics: (1) the proportion of utterances in variation sets; (2) the proportion of word types in variation sets; (3) the average length of variation sets; (4) the Levenshtein edit distances for utterances in variation sets; and (5) the significance of utterances in variations – that is, the probability of chance alignment. The search for variation sets was controlled by two parameters: (1) the maximum allowed gap between successive utterances with partial lexical overlap, which ranged from 0 (no intervening lexically unrelated utterance) to 2 (at most two intervening lexically unrelated utterances); and (2) the number of lexical items (i.e. the length of the *n*-gram) shared by the pair of utterances under consideration, which ranged from 1 (a single shared word) to 6 (six words in the same order but possibly with intervening unrelated words).

We then used the resulting variation set to identify the most frequent lexical bi-grams and tri-grams and the position of the bi- and tri-grams in the utterance. We also sampled a random 110 utterances for each of the five most frequent *n*-grams and determined the part of speech for the immediately following word. Note that as this is a preliminary investigation, we did not examine whether the prevalence of variation sets changed with child age or whether the prevalence changed from corpus to corpus.

Results: statistical properties of variation sets. Figure 4 presents the results of the first four principal statistics of variation sets – the proportion of utterances in variation sets; the proportion of word types in variation sets; the average length of variation sets; and the Levenshtein edit distances for utterances in variation sets. The percentage of utterances in variation sets is reported in Figure 4a for values of gap = 0, 1, 2 and values of *n*-gram $n = 1$, 2, 3, 4, 5, 6. For *n*-gram $n = 1$ (a single lexical item in common), the percentage of utterances in variation sets ranges from 50·8% for gap = 0 (no intervening unrelated utterances) to 82·5% for gap = 2 (up to 2 intervening utterances in the middle of a variation set). For two-word variation sets (*n*-gram $n = 2$) and gap = 0, the figure is 21·3%, which is very close to that reported by previous studies (e.g. Küntay & Slobin, 1996; Waterfall 2006; submitted). Increasing *n*-gram length resulted in progressively smaller proportions of variation sets.

We found that the word types that define variation sets are far from rare (Figure 4b). For all values of gap tested (0, 1, 2), the proportion of word types (relative to the total number of word types in caregiver speech) ranges from about 33% for *n*-gram $n = 1$ (single-word overlap) to about 10% for *n*-gram $n = 6$ (six-word overlap). Thus, for the most common variation sets, between one-quarter and one-third of all word types serve as anchors in a variation set.

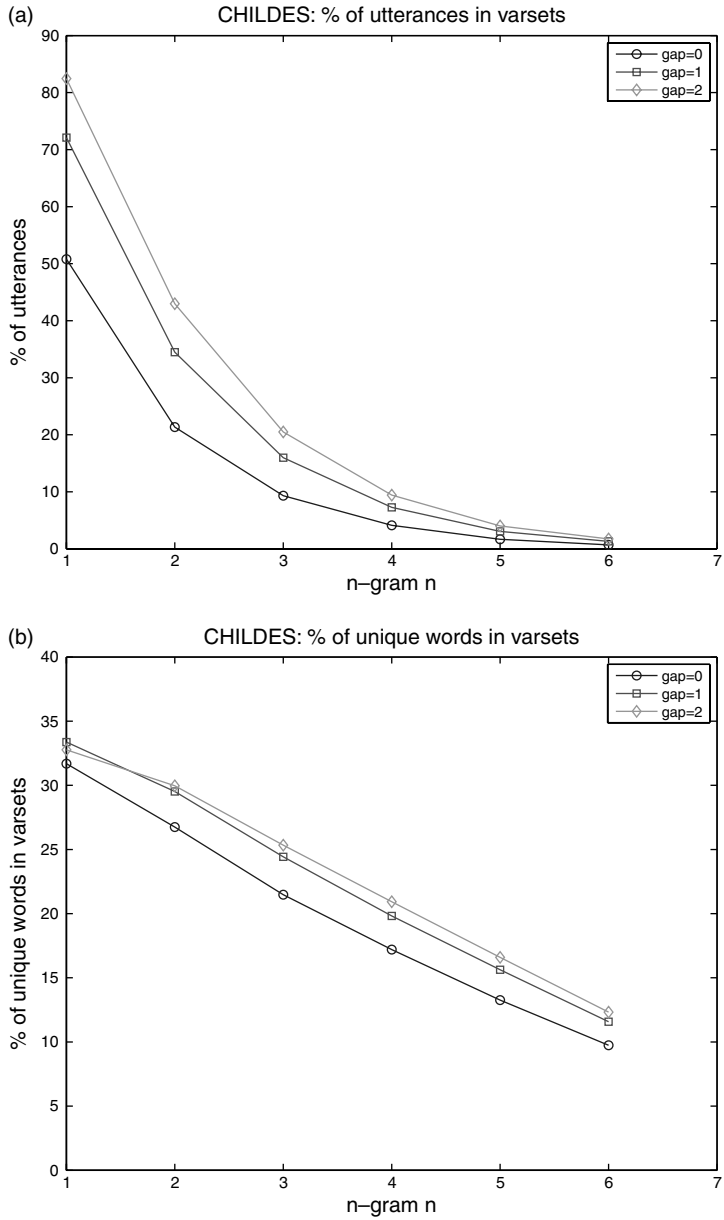

(a)
CHILDES: % of utterances in varsets
% of utterances
n–gram n
gap=0
gap=1
gap=2
(b)
CHILDES: % of unique words in varsets
% of unique words in varsets
n–gram n
gap=0
gap=1
gap=2

Fig. 4. (*Cont.*)

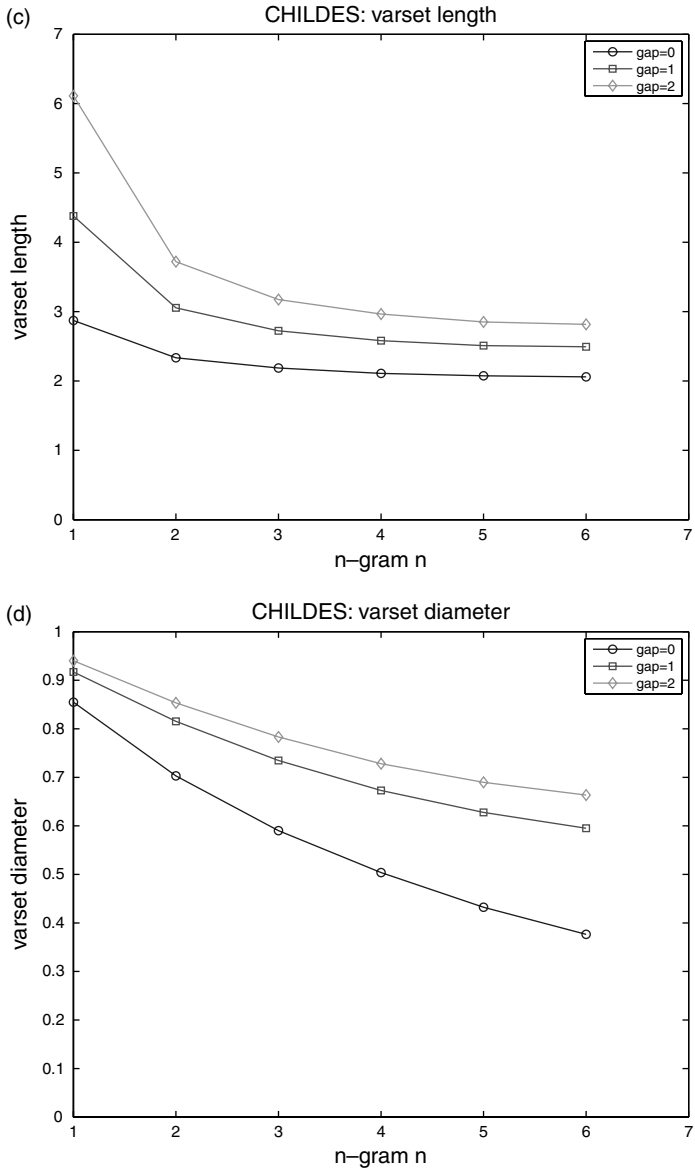


Fig. 4. Four prinicipal statistics for variation sets. (4a) Percentage of caregiver utterances in variation sets in the CHILDES corpora. (4b) Percentage of word types in variation sets the CHILDES corpora. (4c) Mean variation set length in the CHILDES corpora (in utterances). (4d) Mean variation set diameter (in terms of Levenshtein distance normalized to 1) in the CHILDES corpora. For 4a–d, we report six values of *n*-gram *n* overlap and three values of allowed intervening unrelated utterances (gap).

With *n*-gram set to 2, the mean length of a variation set in utterances ranged from 2·33 ($SD = 0{\cdot}004$) for gap = 0 to 3·72 ($SD = 0{\cdot}032$) for gap = 2 (for the full range of data for gap = 0, 1, 2 and *n*-gram $n = 1, 2, 3, 4, 5, 6$, see Figure 4c). The first of these figures is similar to that of Waterfall (2006; submitted), who found that caregiver speech variation sets were on average 2·24 utterances long ($SD = 0{\cdot}13$).

We also identified the diameter of variation sets, which we quantified in terms of normalized Levenshtein distance (Ristad & Yianilos, 1998). To allow comparison of edit distance across sentences of different lengths, we normalized it by dividing the raw edit distance by the length of the longest of the two sequences, which brings it into the range between 0 and 1. For the present corpus, with *n*-gram $n = 2$ and gap = 0, we found that the average diameter of a variation set in terms of normalized edit distance was 0·703 ($SD = 0{\cdot}001$). Not surprisingly, for higher-overlap variation sets the diameter shrank, to 0·376 ($SD = 0{\cdot}006$) for *n*-gram $n = 6$ and gap = 0 (see Figure 4d).

Lastly, to assure safe generalization, any corpus-based inference about structure entertained by the learning algorithm needs to pass a test of statistical significance (Edelman & Waterfall, 2007). Given a variation set, the null hypothesis is that of a chance partial alignment of utterances (Onnis *et al.*, 2008). The learner may test the null hypothesis by comparing the distance between the utterances participating in a variation set to a baseline value – e.g. the cumulative average dissimilarity for the corpus at hand. Figure 5 provides a sample analysis of the data on which such a test could be based: a plot of edit distances between successive sentences in the beginning of our corpus (the first 400 utterances). Specifically, the figure plots edit distances d_n between successive utterances (n and $n+1$) in the corpus, against utterance index n. The solid line in the plot indicates the cumulative average $d_{avg} = (1/n)\sum_{i=1}^{n} d_i$. An utterance pair for which the between-utterance distance is significantly smaller than the cumulative average according to a *t*-test ($d_n < d_{avg}$, indicated in the plot by an asterisk) immediately becomes a candidate for a significant variation set (as opposed to a chance alignment). A learner can rely on this feature of the training corpus in distinguishing between significant and spurious patterns in structure discovery. In Figure 5, which is representative of the corpus at large, this distinction is very easy, the relevant mean distances being many standard deviations apart.

Results: most frequent n-grams in variation sets. Having outlined five principal statistics of variation sets, we now address empirically motivated characteristics of variation sets that focus on particular words and phrases. Table 1 shows the results of the most frequent bi-grams and tri-grams in variation sets in the CHILDES corpora. First, with the exception of *what you*, the words in *n*-grams are contiguous within an utterance. Furthermore, the phrases typically appear at the beginning of the utterance. Roughly 50%

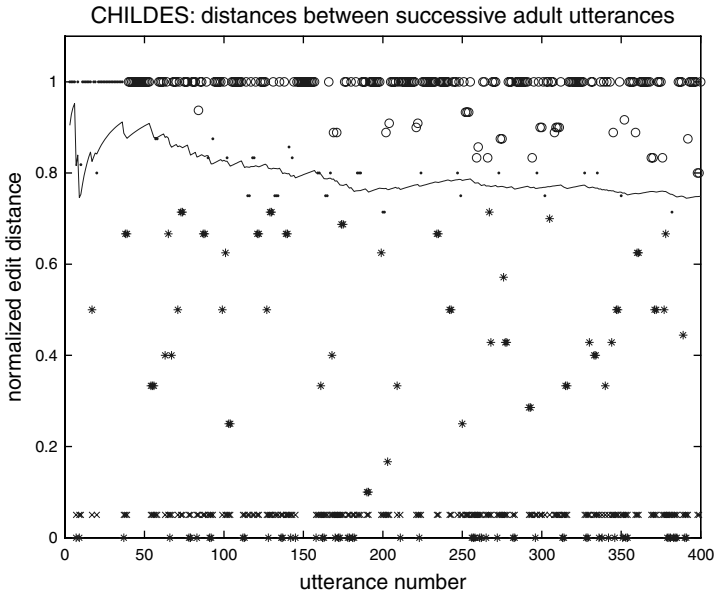


Fig. 5. Edit distances d_n between successive utterances (n and $n+1$) in the CHILDES corpora, against utterance index n, for the first 400 utterances. Solid line shows the cumulative average $d_{avg} = (1/n)\sum_{i=1}^{n} d_i$. Crosses (×) mark utterance pairs for which d_n and d_{avg} do not differ significantly according to a 2-sided t-test. Asterisks (*) mark pairs for which $d_n < d_{avg}$. Circles (o) mark pairs for which $d_n > d_{avg}$.

of the most frequent bi- and tri-grams appear utterance-initially. Approximately 76% of the bi-and tri-grams are either the first or second word in the utterance. This is substantially different from chance, suggesting that the constant part of a variation set is remarkably consistent in position within utterances. These characteristics, contiguity and utterance position, may facilitate children's ability to identify the repetitive part of a variation set while also serving to highlight the novel part (see Lieven, Pine & Baldwin (1997) for similar results on child speech using different analyses).

When the word immediately following the most frequent bi- or tri-gram is analyzed for part of speech (Table 2), it is clear that the frames are highly predictive of specific word classes – this finding parallels one of the key observations behind the development of highly successful statistical part-of-speech taggers in the past two decades (e.g. Charniak, 1997). For example, the word immediately following *are you* is a verb for 90% of the utterances analyzed while *what is* is followed by a pronoun for 63% of the utterances analyzed. This suggests that variation sets, with their rapid succession of related utterances, could be related to the acquisition of specific lexical classes.

TABLE 1. *Frequent* n-grams *and their position in utterances in variation sets*

bi-grams	Utterances in VS	% 1st position	% 2nd position	% 1st+ % 2nd
did you	2410	49%	36%	85%
what you	3520	74%	13%	87%
are you	1924	35%	47%	83%
what is	1633	64%	12%	76%
that's a	1064	43%	26%	69%
you want	2448	35%	33%	68%
you have	2338	34%	29%	63%
it's a	1107	33%	23%	57%
you know	1665	32%	34%	66%
you can	1749	40%	26%	66%
I think	1414	65%	22%	86%
tri-grams				
I don't know	243	69%	20%	88%
what are you	447	83%	12%	95%
you want to	963	35%	36%	71%
you're going to	547	29%	28%	56%
you have to	576	40%	24%	64%
I don't think	217	62%	30%	92%
do you think	257	47%	40%	86%
what did you	378	78%	13%	91%
Average		50%	26%	76%
Standard Deviation		0·176	0·1	0·126

NOTE: VS = variation set; 1st position = first word in *n*-gram is the first word in the utterance; 2nd position = first word in *n*-gram is the second word in the utterance.

Methods: linguistic phenomena in variation sets. In addition to investigating the statistical properties and *n*-gram patterns of variation sets, we examined variation sets for the presence of well-studied linguistic phenomena: dative alternation verbs and object-complement verbs. As with our data-driven analyses, we first identified all the variation sets in the caregiver speech portion of the corpus, using the same two parameters: gap length between successive utterances in a variation set and the number of lexical items shared by a pair of utterances in a variation set.

For this set of analyses, we first determined the proportion of variation sets containing dative alternation verbs and object-complement taking verbs. Next, we calculated the frequency of a list of common verbs for our syntactic phenomena in variation sets. For the five most frequent dative alternation verbs, we determined how frequently each verb occurred with either a double-object dative construction (e.g. *give Mommy the ball*) or a prepositional object construction (e.g. *give the ball to Mommy*). For the five most frequent clausal-complement verbs, we analyzed the following word for part of speech as well as the presence of an object complement clause or

TABLE 2. *Part of speech following frequent* n-grams *in variation sets*

bi-grams	Coded utterances	Followed by: V.	Art.	Pro.	Adj.	N.	Wh/C.
did you	110	82%	–	–	–	–	–
are you	110	90%	–	–	–	–	–
what is	110	–	13%	63%	–	–	–
that's a	110	–	–	–	32%	40%	–
You want	110	58%	11%	25%	–	–	–
tri-grams							
I don't know	110	–	2%	12%	–	3%	35%
What are you	110	97%	–	–	–	–	–
you want to	110	68%	–	23%	–	–	–
you're going to	110	89%	–	–	–	–	–
you have to	110	26%	–	–	–	–	–
I don't think	110	–	5%	54%	–	–	2%

NOTE: V. = verb, Art. = article, Pro. = pronoun, N. = noun; Wh/C. = wh-word or complementizer (e.g. *that*).

a noun phrase direct object. In short, we determined the degree to which the utterances in variation sets were informative of dative alternation structures and object-complement clauses.

Results: linguistic phenomena in variation sets. Approximately 12 percent of sentences in variation sets contain dative alternation verbs (gap = 0, $n = 6$). The proportion is highest for n-grams where $n = 6$ with a gap value of 0, suggesting that dative alternation verbs occur most frequently in variation sets where a large portion of the utterance remains the same and the utterances are immediately adjacent in caregiver speech. As indicated in the section above, the consistency between utterances and their immediate juxtaposition may make these verbs and their accompanying syntactic frames particularly salient to children.

In analyzing the structures following the verb (Table 3), we find that individual dative alternation verbs differ with respect to whether they primarily occur with a double-object dative or a prepositional object. For example, *bring* occurs roughly equally frequently with both structures, while *make* occurs predominately with a prepositional object and *tell* overwhelmingly occurs in a dative-shift construction. It may be the case that different distributions of words within the same paradigm could facilitate the acquisition of those words. For example, the adjacent presentation of these verbs within a variation set might highlight these distributions (e.g. *you have to tell him a story/tell him the whole story*). Alternatively, alternations of structure within a variation set might facilitate children's use of these verbs in both frames (e.g. *give me that/give it to me*). Much more research

TABLE 3. *Linguistic phenomena in variation sets*

Structure Dative alternation	Utterances in VS	Analyzed utterances	Occurs with: % dative structures	% PP/ dative structures	% shift/dative structures
make	239	239	7%	76%	24%
tell	223	223	14%	6%	94%
give	129	129	81%	20%	80%
show	91	91	45%	22%	78%
bring	70	70	27%	42%	58%
Object complement			% VP	% NP	% OC	% Formulaic
see	598	100	0%	72%	14%	0%
want	457	100	53%	36%	10%	0%
know	456	100	0%	34%	41%	0%
think	348	100	5%	6%	59%	0%
say	302	100	2%	36%	19%	29%

NOTE: VS = variation set; PP = prepositional dative (*give the ball to Mary*); shift = dative shift structures (*give Mary the ball*); VP = verbs and verb phrases; NP = noun phrases, pronouns; OC = object-complement clauses; Formulaic = formulaic speech (e.g. *thank you, please, moo*, etc.).

will be needed to determine whether variation sets play a role during the acquisition of these structures. Specifically, it would be important to understand the distributional statistics of these same verbs occurring outside variation sets. It may be the case that variation sets provide more frame alternations, thereby possibly facilitating children's use of multiple syntactic frames with a particular word. On the other hand, variation sets may provide more frame consistency than the rest of the corpora, thereby facilitating children's use of a particular word's most frequent syntactic frame. Thus, future work will have to address whether the statistics of variation sets are different from the corpora as a whole and whether this affects language acquisition.

The results for object complement verbs are similar to those of dative alternations. Approximately 27% of sentences in variation sets contain object complement verbs. Once again, the highest proportion is for *n*-grams where $n = 6$ and gap = 0, suggesting that when caregivers produce variation sets involving object-complement verbs, much of the utterance remains the same and the utterances occur sequentially in speech (e.g. *Who do you think started the whole thing?/Who do you think started it?*). Our analyses of the structures following the verbs in question (Table 3) further indicate that not all object-complement-bearing verbs are used equally with object complements. For example, *see, say* and *want* are followed by an embedded

clause in fewer than 20% of uses, while *think* and *know* are followed by clauses in over 40% of uses. There are also uses that are unique to particular words: *want* is followed by *to* + a verb phrase in roughly 50% of uses, while *say* is followed by formulaic speech (e.g. *yes, no, please, thank you*, etc.) in nearly 30% of uses. Because these diverse uses occur within variation sets, they may be attentionally highlighted or salient for the child. As noted above, in order to evaluate the role of variation sets in the acquisition of verbs and their complements, it is important that future work also examine the distributional statistics of these same verbs occurring outside variation sets. Once again, it is not yet known whether variation sets provide more or less frame consistency than the corpus at large. This question will have to be addressed in order to understand what role, if any, variation sets play in acquisition.

CONCLUSIONS

As suggested in the 'Introduction', the culmination of an empirically minded computational inquiry into language acquisition would be a grammar that is (1) fully generative, (2) algorithmically learnable from realistic data, (3) quantifiably successful and (4) psychologically real. In this paper, we reported some progress in this research program. Specifically, we described two algorithms that, when exposed to the transcribed child-directed speech from a subset of the English CHILDES, acquired a generative grammar. We also detailed findings from longitudinal and experimental studies that explore the role of variation sets in acquisition. Lastly, we provided novel analyses of variation sets using the CHILDES database. We determined that variation sets could be useful not only for acquiring high-frequency *n*-grams (collocations) but also for investigating traditional linguistic phenomena such as the dative alternation.

Our goal for the immediate future is to link the computational and behavioral findings surveyed in this paper. In re-examining ADIOS and ConText, we note that although the patterns ('rules') comprising the acquired grammars subsequently proved capable of transcending the original corpus as measured by the standard quantitative means (recall and precision), testing their psychological reality would be a vast undertaking. We believe that this undertaking, while necessary, should be postponed until a computational model of grammar induction aimed explicitly to account for human developmental psycholinguistics becomes available. Although our algorithms were loosely based on the general principles of language discovery intimated by linguists and developmental psychologists, neither of them made use of certain potentially very important characteristics of child-directed speech, such as the variation set patterns revealed by our corpus studies. We now list some suggestions for the design of a novel algorithm that would

integrate the behavioral and computational findings and that may eventually come close to meeting the research criteria (1–4) laid out in the 'Introduction', thereby contributing to the development of a comprehensive computational model of human language acquisition.

Focus on the conversation

Inspired by the approach pioneered by Zellig Harris (e.g. 1954), we have argued that variation sets should be beneficial to learning the patterns of substitution classes and syntactic constituents. As noted by several researchers (e.g. Pickering & Garrod, 2004; Szmrecsanyi, 2005), natural-language speech, even that between adult interlocutors, involves many kinds of coordination, including partially aligned strings, all of which can be used to scaffold the acquisition of linguistic constructions. By admitting as primary linguistic data the coordinated utterances produced by ALL the participants in an 'overheard' conversation, a grammar induction algorithm intended to model human language acquisition can be made both more realistic and, presumably, more powerful – that is, capable of learning from more impoverished data. In particular, when learning from a corpus of child-directed speech, the model should make use of variation sets and expansions, just as human infants and adults appear to do.

Incremental learning

One significant limitation of the current algorithms such as ConText stems from the amount of statistical power necessary to determine the significance of candidate patterns. Specifically, such algorithms have difficulties with smaller corpora (and more generally with corpora in which the ratio of lexicon size to the number of utterances is high), because they must compute equivalence classes from the corpus before they can take advantage of the highly significant information inherent in alignment. It may be possible to address this issue by replacing equivalence classes with banks of *n*-gram patterns. As sentences pass through the learning mechanism, each one will be simultaneously represented by a set of unigrams, bi-grams, tri-grams, etc., as well as *n*-grams that contain gaps. For example, the phrase *a furry marmot* when considered as an instance of a tri-gram with a gap in position 2 would be analyzed as *a __ marmot*; this would render *a furry marmot* and *a cuddly marmot* equivalent. Each *n*-gram and gapped *n*-gram will maintain a tally of its occurrences and its co-occurrences with other *n*-grams. A co-occurrence of two distinct *n*-grams is almost always significant, given their a priori probability. Because aligned utterances in variation sets contain the same set of co-occurrences more than once, the baseline probability of this event being due to chance is squared (if not raised to a higher power),

greatly reducing the likelihood of the null hypothesis in the significance test for an individual construction that the algorithm encounters.

Moreover, the use of multiple *n*-grams and gapped *n*-grams creates an instant 'template' where the gaps can be filled by leveraging the transitional probabilities already known for their flanking members. In addition, this approach allows the process of constructing the grammar to be recursive: transitional probabilities and gaps can reference other sets of *n*-grams and gapped *n*-grams, much as equivalence classes in learned by ADIOS or ConText are currently capable of referencing other equivalence classes. In other words, traditional units like nouns, verbs, noun phrases, verb phrases and clauses can be built up by referencing *n*-grams and their clustering. Additional possible uses of the new insights from developmental psycholinguistics in devising more powerful and psychologically relevant algorithms for empirical generative grammar learning are the subject of ongoing research (Edelman, 2010).

In conclusion, we propose that a key principle for linking computational structure-acquiring algorithms and insights from behavioral data is to rely explicitly on naturally occurring discourse structures that facilitate alignment, thereby reducing the power needed to encounter statistically significant patterns and allowing the algorithms to operate in a more psychologically real way.

REFERENCES

Adriaans, P. (2001). Learning shallow context free languages under simple distributions. In A. Copestake & K. Vermeulen (eds), *Algebras, diagrams and decisions in language, logic and computation*, 1–35/ Stanford, CA: CSLI/CUP.

Baayen, R. (2006). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baker, L. & McCallum, A. (1998). Distributional clustering of words for text classification. In *SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, Melbourne*, 96–103. New York: ACM Press.

Batchelder, E. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition* **83**, 167–206.

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

Bloom, L. (1973). *One word at a time: The use of single-word utterances before syntax*. The Hague: Mouton.

Bod, R. (2009). Constructions at work or at rest. *Cognitive Linguistics* **20**, 129–34.

Brodsky, P., Waterfall, H. & Edelman S. (2007). Characterizing motherese: On the computational structure of child directed language. In D. McNamara & J. Trafton (eds), *Proceedings of the 29th Cognitive Science Society Conference, Nashville*, 833–38. Austin, TX: Cognitive Science Society.

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

Brown, R., Cazden, C. & Bellugi, U. (1969). The child's grammar from one to three. In J. P. Hill (ed.), *Minnesota symposium on child development, Volume 2*, 28–73. Minneapolis, MI: University of Minnesota Press.

Charniak, E. (1997). Statistical techniques for natural language parsing. *AI Magazine* **18**, 33–44.
Chater, N. & Vitányi, P. (2007). Ideal learning of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology* **51**, 135–63.
Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.
Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
Christiansen, M. & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences* **5**, 82–88.
Clark, A. (2001). Unsupervised language acquisition: Theory and practice. Unpublished PhD thesis, School of Cognitive and Computing Sciences, University of Sussex.
Clark, A. (2006). PAC learning unambiguous NTS languages. In *Proceedings of ICGI*, 59–71. Tokyo: Springer-Verlag.
Demuth, K., Culbertson, J. & Alter, J. (2006). Word-minimality, epenthesis, and coda licensing in the acquisition of English. *Language & Speech* **49**, 137–74.
Edelman, S. (2010). On look-ahead in language: Navigating a multitude of familiar paths. In M. Bar (ed.), *Prediction in the brain* (to appear). New York: Oxford University Press.
Edelman, S. & Waterfall, H. (2007). Behavioral and computational aspects of language and its acquisition. *Physics of Life Reviews* **4**, 253–77.
Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.
Goldsmith, J. (2007). Towards a new empiricism. In J. B. de Carvalho (ed.), *Recherches linguistiques à Vincennes*, *Volume 36*.
Goodman, J. (2001). A bit of progress in language modeling. *Computer Speech and Language* **15**, 403–434.
Hall, W., Nagy, W. & Linn, R. (1984). *Spoken words: Effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Erlbaum.
Hall, W., Nagy, W. & Nottenburg, G. (1981). *Situational variation in the use of internal state words*. Champaign, IL: University of Illinois.
Hall, W. & Tirre, W. (1979). *The communicative environment of young children: Social class, ethnic and situational differences*. Champaign, IL: University of Illinois.
Harris, Z. (1954). Distributional structure. *Word* **10**, 140–62.
Higginson, R. (1985). Fixing-assimilation in language acquisition. Unpublished doctoral dissertation, Washington State University.
Hoff-Ginsberg, E. (1985). Some contributions of mothers' speech to their children's syntactic growth. *Journal of Child Language* **12**, 367–85.
Hoff-Ginsberg, E. (1990). Maternal speech and the child's development of syntax: A further look. *Journal of Child Language* **17**, 85–99.
Joshi, A. & Schabes, Y. (1997). Tree-adjoining grammars. In G. Rozenberg and A. Salomaa (eds), *Handbook of formal languages, 3*, 69–124. Berlin: Springer.
Klein, D. & Manning, C. (2001). Distributional phrase structure induction. In W. Daelemans & R. Zajac (eds), *Proceedings of the Conference on Natural Language Learning (CoNLL) 2001*, 113–20. Toulouse: ACL.
Klein, D. & Manning, C. (2002). Natural language grammar induction using a constituent-context model. In T. G. Dietterich, S. Becker & Z. Ghahramani (eds), *Advances in neural information processing systems 14*, 35–42. Cambridge, MA: MIT Press.
Küntay, A. & Slobin, D. (1996). Listening to a Turkish mother: Some puzzles for acquisition. In D. Slobin, J. Gerhardt (eds), *Social interaction, social context, and language: Essays in honor of Susan Ervin-Tripp*, 265–86. Hillsdale, NJ: Lawrence Erlbaum Associates.
Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th ACL*, 25–32, College Park, MD: ACL.
Lieven, E., Pine, J. & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language* **24**, 187–219.
MacWhinney, B. (1995). *The CHILDES Project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Volume 1: Transcription format and programs. Volume 2: The Database*. Mahwah, NJ: Erlbaum.
MacWhinney, B. & Snow, C. (1985). The Child Language Exchange System. *Journal of Computational Linguistics* **12**, 271–96.
Nelson, K. (1977). Facilitating children's syntax acquisition. *Developmental Psychology* **13**, 101–107.
Onnis, L., Waterfall, H. & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition* **109**, 423–30.
Pereira, F., Tishby, N. & Lee, L. (1993). Distributional clustering of English words. In *Meeting of the Association for Computational Linguistics (ACL)*, 183–90. ACL.
Pickering, M. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* **27**, 169–225.
Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science* **22**, 425–69.
Ristad, E. & Yianilos, P. (1998). Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 522–32.
Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent–child discourse. In K. E. Nelson (ed.), *Children's language, Vol. 4*, 1–28. Hillsdale, NJ: Lawrence Erlbaum Associates.
Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.
Sokolov, J. (1993). A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology* **29**, 1008–1023.
Sokolov, J. & MacWhinney, B. (1990). The CHIP framework: Automatic coding and analysis of parent–child conversational interaction. *Behavior Research Methods, Instruments & Computers* **2**, 151–61.
Solan, Z., Horn, D., Ruppin, E. & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Science* **102**, 11629–11634.
Stine, E. & Bohannon III, J. (1983). Imitations, interactions, and language acquisition. *Journal of Child Language* **10**, 589–603.
Stolcke, A. & Omohundro, S. (1994). Inducing probabilistic grammars by Bayesian model merging. In R. C. Carrasco & J. Oncina (eds), *Grammatical inference and applications*, 106–118. Berlin: Springer.
Suppes, P. (1974). The semantics of children's language. *American Psychologist* **29**, 103–114.
Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* **1**, 113–49.
Valiant, L. (1984). A theory of the learnable. *Communications of the ACM* **27**, 1134–1142.
Waterfall, H. (2006). A little change is a good thing: Feature theory, language acquisition and variation sets. Unpublished doctoral dissertation, University of Chicago.
Waterfall, H. (submitted). Relation of variation sets to noun and verb development. Manuscript submitted for publication.
Wolff, J. (1988). Learning syntax and meanings through optimization and distributional analysis. In Y. Levy, I. M. Schlesinger & M. D. S. Braine (eds), *Categories and processes in language acquisition*, 179–215. Hillsdale, NJ: Lawrence Erlbaum.

APPENDIX

We explored the performance of ConText for a range of settings of its three parameters by estimating recall and precision of the learned grammars (Figure A1). In estimating precision, because of the large number of sentences (a total of 1,800 sentences in this experiment) a single native speaker of English rated 100 sentences for acceptability on a scale of one to seven.

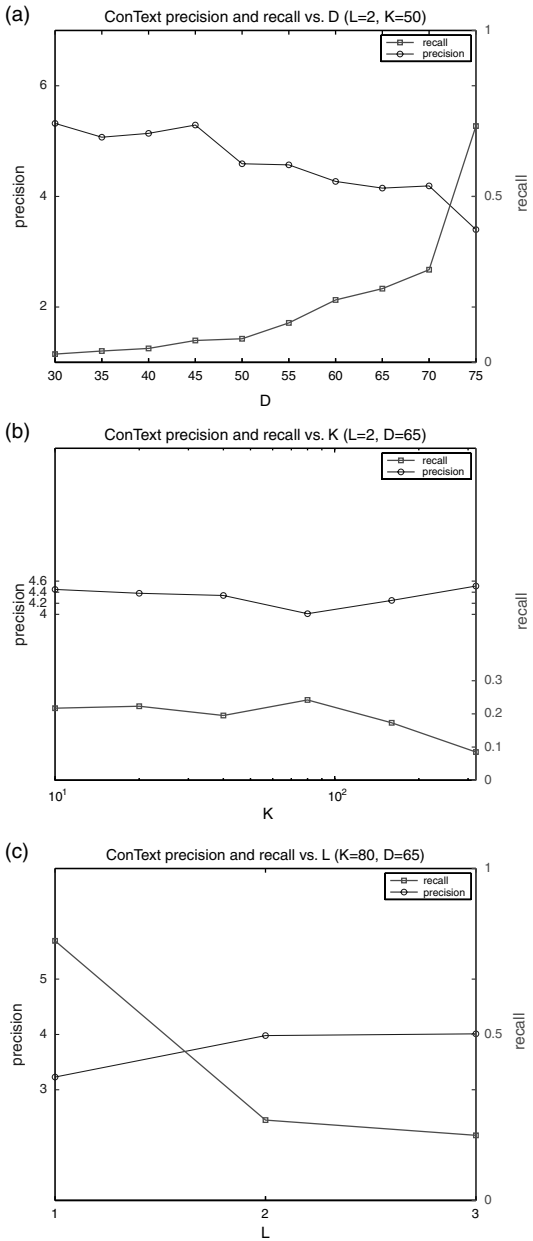


Fig. A1. ConText performance for a range of values of K, L, and D parameters. A1a. The precision and recall of grammars inferred by the ConText algorithm for a range of values of the parameter *K* (the minimum number of times a sequence must appear in the corpus for

The first parameter, K, controls the minimum number of times a sequence must appear in the corpus to be considered for clustering. As K grows larger, the number of sequences not captured by the grammar increases. Therefore we expected recall decrease, with a corresponding increase in precision, as indicated by Figure A1a. Note that $L=2$ and $D=0{\cdot}65$, described below.

As the value of the parameter L, which controls the size of the context window around each sequence, grows larger, we expected lower recall and higher precision, as the vector representation of each word sequence becomes more fine-grained. As Figure A1b shows, this is precisely what happened when L changed from 1 to 2. Increasing L further made no differences, presumably because of the relatively small average sentence length of this corpus. Note that $K=80$, $D=0{\cdot}65$.

Increasing the value of the third parameter, D, which controls the maximum distance between two sequences that can still be clustered together, allows more sequences to be clustered together, leading to more, larger equivalence classes. Thus, the resulting grammar is expected to be more generative, corresponding to higher recall. At the same time, precision is expected to decrease with a larger D: as the criterion for substitutability becomes more lenient, more sequences are erroneously clustered together. Both these trends are clearly visible in Figure A1c. Note that $K=50$, $L=2$.

clustering). Note that the x-axis is logarithmic. A1b. The precision and recall for a range of values of the parameter L (the size of the context window around the sequence under consideration). A1c. The precision and recall for a range of values of the parameter D (the maximum distance between two sequences that can still be clustered together).