Computational Models of Language Acquisition

Shuly Wintner

Department of Computer Science, University of Haifa
Mount Carmel, 31905 Haifa, Israel
shuly@cs.haifa.ac.il

Abstract. Child language acquisition, one of Nature's most fascinating phenomena, is to a large extent still a puzzle. Experimental evidence seems to support the view that early language is highly formulaic, consisting for the most part of frozen items with limited productivity. Fairly quickly, however, children find patterns in the ambient language and generalize them to larger structures, in a process that is not yet well understood. Computational models of language acquisition can shed interesting light on this process. This paper surveys various works that address language learning from data; such works are conducted in different fields, including psycholinguistics, cognitive science and computer science, and we maintain that knowledge from all these domains must be consolidated in order for a well-informed model to emerge. We identify the commonalities and differences between the various existing approaches to language learning, and specify desiderata for future research that must be considered by any plausible solution to this puzzle.

1 Introduction

Language acquisition is one of Nature's most fascinating puzzles. Human languages are extremely complex systems, yet (most) children acquire them naturally, fairly quickly and seemingly effortlessly [24, p 144]. Research in language acquisition attempts to study the mechanisms of this puzzle in order to explain the very nature of language itself: the primary cognitive capacity which makes us human.

Two competing theories of language acquisition dominate the linguistic and psycholinguistic communities [60, pp. 257-258]. One, the *nativist* approach, originating in Chomsky [21, 22, 23] and popularized by Pinker [50], claims that the linguistic capacity is innate, expressed as dedicated "language organs" in our brains; therefore, certain linguistic universals are given to language learners for free, requiring only the tuning of a set parameters in order for language to be fully acquired. The other, *emergentist* explanation [8, 61, 42, 47, 43, 44, 60], claims that language emerges as a result of various competing constraints which are all consistent with general cognitive abilities, and hence no dedicated provisions for universal grammar are required. Consequently, "[linguistic universals] do not consist of specific linguistic categories or constructions; they consist of general ... cognitive abilities" [59, p. 101]. Furthermore, language is first acquired in an *item-based* pattern: "[young children] do not operate on the basis of any linguistic abstractions, innate or otherwise. Fairly quickly, however, they find some patterns in the way concrete nouns are used and form something like a category of a

A. Gelbukh (Ed.): CICLing 2010, LNCS 6008, pp. 86–99, 2010.

noun, but schematization across larger constructions goes more slowly. The process of how children find patterns in the ambient language and then construct categories and schemas from them is not well understood at this point." [59, pp. 106-107].

Computational models can shed new light on language acquisition processes and provide new insights into the nativist vs. emergentist debate. Three related fields of research address grammar induction from data [4]: "applied" grammar induction in linguistics; empirical (computational) grammar induction; and formal (mathematical or logical) grammar induction. Adriaans and van Zaanen [4] conclude that "it is time to remove the (artificial) boundaries and combine the research performed within each subfield." In this paper we survey several approaches to language learning from data and assess their contribution to a clearer picture of child language acquisition. We review existing work with an eye to consolidating the "applied" and "empirical" approaches to grammar induction. We then propose future research directions that aim at a better explanation of early stages of human language acquisition.

This paper is structured as follows. Section 2 defines the language learning task in a general way, indicating in what ways different approaches may vary. Section 3 discusses the difficult issue of evaluating the performance of models and algorithms that address such tasks. We then survey existing approaches in three classes: works whose motivation is to explain child language acquisition from a cognitive perspective (Section 4.1); works that aim to devise an efficient mechanism for inducing (formal) grammars from raw data (Section 4.2); and a few recent works that try to consolidate some aspects of the two approaches (Section 4.3). Finally, we propose in Section 5 some directions for future research that consolidate the benefits of rigorous computational models backed up by solid psycholinguistic findings.

2 The Language Learning Task

The task that we focus on is language learning: a *learner*, be it a child or a computer program, is presented with *data*, in the form of raw utterances. To approximate the fact that human learning is grounded in real-world situations, the raw data are sometimes annotated with part-of-speech (POS) categories, syntactic information or even semantic information. The learner's task is to generalize the data and induce a model of the grammatical utterances (in other words, a *grammar*). In the formal sense, a grammar is a generative device that defines a set of expressions, its *language*, and induces some structure (e.g., trees) on expressions in the language. A grammar can be expressed explicitly as a set of rules, perhaps with probabilities attached to them; or implicitly as a set of "operations" on strings, with or without "slots" (which are the equivalent of non-terminal symbols in a formal setting). The success of the learner can be evaluated by testing its grammar on new utterances. Two aspects of the grammar can be tested: its ability to generate new utterances; and its ability to assign a valid structure to the grammatical utterances.

Language learning tasks rely on the existence of large text corpora that document language use, both for training and for evaluation [25, 49, 31]. Computational linguistic tasks standardly use manually-annotated sentences from the Penn Tree Bank (PTB, Marcus et al. [48]), whose data are taken from the Wall Street Journal. Grammar induction tasks are sometimes limited to a subset of the PTB, where sentences are limited to length of 10 or less (WSJ10). Clearly, this is a genre that is not well-defined linguistically, and quite likely irrelevant for language acquisition investigations.

In order to investigate the development of child language, corpora which document linguistic interactions involving children are needed. The CHILDES database [41], a 300MB corpus in over 25 languages, contains transcripts of spoken interactions between children at various stages of language development and their caretakers. CHILDES provides vast amounts of useful data for linguistic, psychological, and sociological studies of child language development. To date, this database has served as the basis for over 1500 published articles and as a secondary resource in hundreds of other studies. Many of the CHILDES corpora are morphologically analyzed and annotated in a compatible manner, which makes it possible to compare language development across different languages. Recently, the English CHILDES database has been annotated with syntactic structures in the form of *grammatical relations* [52, 53, 54]; similar efforts are currently underway for several other languages. The CHILDES database thus provides a perfect environment for investigating language development and for evaluating psycholinguistic hypotheses.

Computational approaches to language learning from data, and in particular the works surveyed in Section 4, can be distinguished along the following axes:

The data. What data are presented to the learner? Cognitively-motivated approaches (Section 4.1) assume that the data are raw texts, usually of child-directed speech (but sometimes including also child speech, i.e., utterances produced earlier by the same child). Sometimes, these data are accompanied by some form of annotation, aiming to reflect the grounding of language in real-life situations. Grammar induction algorithms (Section 4.2) often do not assume that the data are linguistic; and when they are, they are often annotated with POS tags, and sometimes only the sequences of POS tags are considered, ignoring the actual words. Training material in this paradigm consists mainly of the WSJ, and in particular WSJ10.

The task. What is the learner required to learn? Is it a language as a set of strings, or are these strings augmented by (usually, tree-) structures?

The grammar. What formalism is the grammar expressed in? Grammar induction algorithms are usually formal and explicit in their definition of the class of models that they attempt to learn. These can be deterministic finite-state automata (FSA) or Hidden Markov Models (HMMs) or Probabilistic context-free grammars (PCFGs) or Tree substitution grammars, with a variety of probabilistic models. Cognitive models are more vague on this point, and may represent the grammar implicitly and informally, sometimes in a rather ad-hoc manner.

Evaluation. Grammar induction algorithms are evaluated on annotated data; fundamentally, they are expected to learn the bracketing (and, sometimes, also the labels) of manually annotated corpora. In contrast, works in the cognitive tradition evaluate on child language data, which is crucially not annotated. We elaborate on evaluation in the next section.

3 Evaluation

Several factors make the evaluation of language learning systems difficult. First, especially when child-data are concerned, the training data provided to the system are extremely limited: even with high-density corpora, it is assumed that the corpus reflects less than 10% of the utterances the child was exposed to during a very short period (see Rowland et al. [51]). Second, it is unclear whether the task should be evaluated by testing the strings generated by the grammar, or also the structures that the model induces on them. The latter task is more demanding, and it is usually unclear what the "correct" structures are that the grammar should produce. Clearly, the PTB is inappropriate to investigate child language; and WSJ10 in particular is a very artificial genre. Finally, while it is relatively easy to measure the portion of the target utterances that the system properly generated, it is much harder to asses the proportion of the utterances generated by the model that are indeed grammatical.

In the computational linguistics community, similar tasks are standardly evaluated using two measures adopted from Information Retrieval: *precision* and *recall*, and their harmonic mean, *f-score*. Informally, recall measures the ability of the grammar to account for new utterances: it is roughly the proportion of the strings in the test data that can be correctly generated by the grammar. Precision, on the other hand, measures the extent to which grammar-generated strings are observed in the test data. In the extreme case, one can always maximize either of the two measures: if a grammar generates nothing its precision is 100% (but its recall is 0); if it generates everything its recall is 100% (but its precision is very low). The f-score therefore balances between the two. For language learning tasks, however, while recall is relatively easy to asses, precision is much harder: for it is possible to present sentences from a test corpus to a learner and verify that the learner accepts them as grammatical; but it is more difficult to ask the learner to generate utterances, and then verify that they are correct.

van Zaanen and Geertzen [66] identify four types of approaches to the evaluation of learning algorithms, each with is own problems. The *looks good to me* approach (namely, informal evaluation by the author of a system) is obviously subjective and unreliable. An alternative is *rebuilding a-priory known grammars*; here; the authors of a system construct a small grammar whose sentences are used as input for the language learning system. This is again subjective, and in addition only toy grammars can be build in this fashion. The *language membership* method amounts to measuring the precision and recall of a learner with respect to a test corpus, which is problematic as explained above; finally, *comparison against a treebank* is particularly problematic for child language, and is additionally very brittle [66].

To address the problem of evaluation, Chang et al. [20] propose a measure, called *sentence prediction accuracy (SPA)*, which basically quantifies the extent to which a learner (be it a child or a computer program) can correctly order the words in a target utterance, when these words are unordered. While SPA overcomes some problems (e.g., it is independent of the language and of any underlying linguistic theory), it is a very inaccurate measure of grammaticality. It is extremely strict, in the sense that a mistake in the placement of a single word renders the entire utterance unaccounted for; and it also implies that if two grammatical utterances differ in their word order (e.g., because an adverbial is shifted in a sentence), only one of them will be counted as correct by the SPA measure. Finally, the measure was only tested on extremely short utterances. [1] We conjecture that it will not scale up to longer sentences (indeed, the results of Chang et al. [20] indicate that the SPA measure correlates better with the specific corpus used for learning than with the learning algorithm).

An alternative method for estimating both precision and recall is proposed by Brodsky et al. [18]. Based on the observation that two constrained models that are trained on disjoint corpora are unlikely to agree on the grammaticality of any given sentence, it uses large corpora to train language models that are then used to assess the probability of the test sentences. A full evaluation of this method is still underway.

Finally, Kol et al. [36] recently proposed a first approximation for assessing the level of over-generation in a learner. They measure the precision of a learner by training it on a large corpus. Specifically, working with CHILDES data, in which files are ordered chronologically, they train the model on early files, plus 90% of the (child-directed *and* child) utterances in the current file; they then test on the remaining 10% of the child utterances in the current file. To assess over-generation, they repeat the same procedure, training on the same data but evaluating on child utterances (longer than one word) *in reverse word order*. Ideally, a learner should perform well on the first task and very poorly on the second.

4 Existing Approaches

4.1 Cognitively-Motivated Computational Approaches to Child Language Acquisition

Within the cognitive linguistics paradigm, computational approaches to language learning investigate the degree to which the utterances a child is exposed to can be used to determine the multi-word expressions the same child will produce during early language development. Lieven et al. [40] suggest that "a lexically-based positional analysis can account for the structure of a considerable proportion of children's early multi-word utterances." This is tested on eleven children aged between 1;0 and 3;0. On average, 60% of all the children's multi-word utterances are defined as frozen by the analysis. These results are replicated by Lieven et al. [38], this time focusing on one child, but using a high-density corpus consisting of 5 hours of recordings per week (together with a maternal diary for the previous 6 weeks.) The findings are that only one third of the multi-word utterances of the child are novel, and three quarters of those can be accounted for by one operation only on the basis of previous utterances. Five types of "operations" are defined which the child can use to construct a new utterance from fragments of previously-heard
utterances.

Dąbrowska and Lieven [26] identify two problems with the above method: first, "the method does not provide an explicit description of the child's linguistic knowledge." In

[1] While this piece of data is missing in Chang et al. [20], it can be deduced that the average length of an utterance in their corpus was less than 3. These figures are specified in Dąbrowska and Lieven [26], p. 446, presumably referring to the same corpus. There, the number of words per utterance ranges between 1.56 (Brian, age 2;0) and 3.15 (Annie, age 3;0). Even in the adult speech the average utterance length is 4.44).

other words, no explicit model of linguistic knowledge, or *grammar*, is defined. Second, "the method is too unconstrained since the five operations defined by the authors made it possible, in principle, to derive any utterance from any string." In other words, the above models are *over-generating*. To overcome the problems, they propose *two* operations: juxtaposition and superimposition. Working with a dense corpus of two children at ages 2;0 and 3;0, they divide the corpora into two parts: approximately the first 80% of the utterances in each corpus are defined as the *main* corpus, and the remainder are called *test*. The focus is syntactic questions; for each such utterance in the test corpus, called the *target*, they extract relevant *component units* from the main corpus. These are utterances that share lexical material with the target. They then determine whether the target can be produced from the extracted utterances by means of juxtaposition (i.e., concatenation) or superimposition. The latter operation is loosely defined; it amounts to identifying similarities among patterns in the main corpus, and generalizing such patterns to *schemas with slots*. Superimposition allows slots to be filled by lexical material. Crucially, the corpora used in this research were manually annotated with semantic information. Superimposition is then constrained by the semantic type of the slot, such that only fillers of the same type can be used. The results show that as much as 75% of the questions at age 2;0 are immediate imitations of previous questions (this figure goes down to as low as 21% at age 3;0); and all the rest can be generated with few (at most 4) operations from previously-heard material.

Lieven et al. [39] adapt the previous research, extending it to four children. The method, which is referred to as the *traceback procedure* here, is basically the same. Again, the data are assumed to be semantically annotated and the semantic tags are used in the definition of 'slots'. The actual algorithm is not defined in sufficient detail (for example, it is unclear how schemas with more than one slot are generated, or whether there is an upper bound to the number of slots in a schema).

Bannard et al. [7] augment the traceback procedure by a *trace forward* procedure. Here, the task of the learner is better defined: given a main corpus to learn from, the learner has to extract a formal grammar by generalizing the utterances in the main corpus. The grammar induction procedure is not described with sufficient rigor, but it is clear that the emerging grammar is a context-free grammar with a single non-terminal symbol. Rules are generated based on utterances that share lexical material, as above, but the details are not specified. Out of the infinitely many possible grammars that fit the data, Bannard et al. [7] select those that are most probable given the data, using a Bayesian model with simple independence assumptions for optimizing the likelihood, and minimum description length (MDL) assumptions of the prior. The results show that the extracted grammars perform well both in terms of their recall and, in lieu of a precision evaluation, in terms of their *perplexity* (defined informally as "how surprised the model is by the data").

In a series of works, Freudenthal et al. [28, 29, 30] develop the MOSAIC (Model of Syntax Acquisition in Children) paradigm. This model takes as input corpora of transcribed child-directed speech and learns to produce as output utterances that become progressively longer as learning proceeds. The model is based on a hierarchical network in which more deeply embedded nodes represent longer utterances, and where links connect nodes to form certain generalizations. Crucially, the same corpus is given to the learner several times. While MOSAIC has been shown to properly simulate several phenomena associated with early language acquisition in several languages, in part due to its inherent bias towards learning from the edges of the utterance, it is not viewed as a realistic model of the language acquisition process itself, but rather as one possible implementation of inherent biases in learning.

4.2 Computational Grammar Induction

A different line of research falls under the category of *grammar induction* (or, more specifically, *computational grammar inference*). Here the goal is to devise algorithms that can learn accurate, compact models for identification of language (i.e., grammars) from finite sets of examples [4]. Such approaches are usually not cognitively motivated; Klein and Manning [32], p. 35, for example, explicitly mention that "the presented system makes no claims to modeling human language acquisition," and Borensztajn et al. [17] add that their approach "has no pretense of being a model for language acquisition"; but the relation to the works discussed above is obvious. Formally, a finite set of examples is consistent with infinitely many different grammars, and thus different approaches must somehow constrain or bias the set of hypotheses from which grammars can be drawn [27].

The EMILE model [1, 2, 3] attempts to learn the grammatical structure of a language from positive examples, without prior knowledge of the grammar. It is based on the idea that expressions of the same (syntactic) type can be substituted in the same context, and hence it searches for clusters of expressions and contexts in the input, interpreting them as grammatical types. The model then generalizes the sample and learns rules of a context-free grammar. A related approach is Alignment-based Learning (ABL) [63, 64, 62]. Given a corpus of sentences, an alignment learning phase first finds possible constituents by aligning pairs of sentences and identifying parallel strings. Strings that are unequal in a pair of sentences are considered *hypotheses*. Then, non-terminal types are assigned to hypotheses, merging different non-terminals that occur in the same context. The result is an induced context-free grammar. While ABL and EMILE are implemented differently, and EMILE only extracts a rule when sufficient support is available in the corpus whereas ABL stores all possible rule candidates and selects the best ones [65], the two systems are similar in spirit. In particular, both can learn recursive structures. The two systems were evaluated on Dutch corpora; the metric was unlabeled bracketing f-score. Whereas EMILE reached an f-score of 0.25-0.41 (depending on the corpus), ABL's performance was much higher, at 0.39-0.62 [65].

Stolcke and Omohundro [58] propose a technique called *Bayesian Model Merging (BMM)*: first, strings that are observed in the data are incorporated by adding ad-hoc rules to form an initial grammar; then, the grammar is made more concise by merging some of the rules. Stolcke and Omohundro [58] discuss two incarnations of their technique, one in which the models are probabilistic context-free grammars (PCFGs), and another in which they are hidden Markov models (HMMs). In the former, rules are merged by identifying non-terminal symbols A and B if the rule $A \rightarrow B$ is in the grammar; this leads to (over-) generalizations, and renders the grammar more compact. In the latter, two HMM states are merged to a state that inherits the union of their transitions (and emission probabilities). In both cases, the prior probabilities are optimized by minimizing their description length. Stolcke and Omohundro [58] discuss the application of BMM to natural language learning, but do not provide quantitative evaluation results.

In a series of works, Klein and Manning [32, 33, 35] present the *constituent-context model (CCM)*. Here the task is to determine the correct bracketing of sentences in the input: the assumption is that the input is tagged with parts of speech (POS); in fact, the algorithm ignores the actual words and works on POS tag sequences. The output is a tree structure without the labels of non-terminal symbols. The model is a generative one; first, an initial bracketing is chosen from some distribution and a sentence is generated given the bracketing, assuming that the context and yield of each span are independent of each other. Then, an EM algorithm is run on the model to induce structure, assuming that the sentence is observed but the bracketing is not. This model was evaluated on the PTB WSJ10 subset, resulting in an f-score of 0.71, reducing about a quarter of the errors of a trivial (right-bracnhing trees) baseline (which yields 0.60). This result improves to 0.776 when the model is combined with a dependency parsing model in subsequent work [34].

Data-oriented parsing (DOP, Bod et al. [15]) is a paradigm for supervised parsing that differs from other approaches in that it considers *all* the possible structures given in a training corpus, and estimates their likelihood from the data. It can then be used to assign a structure to a new utterance by combining sub-trees from the training corpus. In its unsupervised version [11, 12, 13], called U-DOP, the algorithm initially assigns all possible unlabeled binary trees to an un-annotated training set, and then employs a probabilistic model to determine the most likely tree for a new utterance (various probabilistic models were investigated). The best results outperform the previous model, with an f-score of up to 0.80 on WSJ10.

Various works address the issue of inducing labels for the unlabeled trees. Notably, Borensztajn and Zuidema [16] extend the BMM model of Stolcke and Omohundro [58], but they assume that the input is already bracketed. Their algorithm then proceeds by merging nonterminal labels to maximize a Bayesian objective function. The algorithm is evaluated on the PTB WSJ10 subset, and shows best performance on the labeling task (although when used only for bracketing, it is much inferior to competing algorithms).

While many grammar induction algorithms start with strings of POS tags, this is not the case with Seginer [55], who uses lexical information (and does not assume known POS tags). While other algorithms resort to unsupervised learning of POS tags, which amounts to clustering, here the algorithm collects lists of labels for each word, based on its neighbors, and uses these labels to parse. The parser is incremental, local and greedy, and hence quite efficient. Evaluated on WSJ10, the results are an f-score of almost 0.76 when parsing begins from plain text.

Note that algorithms for inducing part of speech categories from raw data (i.e., unsupervised POS tagging) abound, both in the cognitive linguistic literature (e.g., Li et al. [37] and references therein) and in the computational linguistic literature (e.g., Banko and Moore [5], Smith and Eisner [56]).

4.3 Consolidating the Two Approaches

Most of the works described above fall into one of two classes: either the motivation is to explain child language acquisition from a cognitive perspective (Section 4.1); or it is to devise an efficient mechanism for inducing (formal) grammars from raw data (Section 4.2). Very few works, and only recently, try to consolidate some aspects of the two approaches.

The ADIOS system [57] implements a novel algorithm that learns a complex context-free grammar from raw data. Based on a graph representation, the algorithm performs segmentation and generalization of the input simultaneously. The system was applied to several types of data, both linguistic (including CHILDES data) and non-linguistic (protein sequences). Recall was evaluated automatically, while to assess precision human judgements were used. The results show that ADIOS is superior to other grammar induction algorithms that can learn from raw data. In a subsequent work, Berant et al. [9] observe that the algorithm does not deal well with complex texts and improve it by applying a two-stage learning technique: first, sentences in the input are split to sub-sentences on the basis of conjunctions in the text; then, the resulting simpler corpus is processed as above. Precision was evaluated by feeding the sentences generated by the learner to an alternative parser, and f-score varied between 0.24 and 0.39, depending on the precise task. Brodsky et al. [18] apply ADIOS to the full (English section of the) CHILDES corpus. Training on 300,000 utterances and testing on 500, the system reached a recall of 0.5 and precision of 0.63. Precision was again evaluated manually by humans judging the grammaticality of 100 generated utterances.

Borensztajn et al. [17] use the DOP paradigm as a vehicle for investigating psycholinguistic hypotheses. Specifically, they use the syntactically-anotated Brown corpus of CHILDES [54] to learn DOP-style structures. These tree fragments are then used to induce structure on utterances in the test corpus. This is an automatic approach to identifying the most probable multi-word units (constructions) in children's utterances. The main result is that *abstraction*, defined as the ratio between non-terminal and terminal leaves in the tree fragments that represent constructions, increases with age. One of the main drawbacks of this approach is that the grammar is induced from POS-tagged *and* syntactically-annotated corpora; cognitively, this amounts to assuming that children have access to the syntactic structure of the utterances they are exposed to, which cannot be the case in early language acquisition.

5 Directions for Future Research

According to Edelman and Waterfall [27], p. 265, "of the three goals of linguistic theory... the most promising one at present is, in fact, an algorithmic discovery procedure for grammar." Similarly, Adriaans and van Zaanen [4], p. 200 observe that "researchers within the several sub-fields [linguistic, empirical and formal grammatical inference] seem to have created certain boundaries between the fields" and conclude that "it is time to remove the (artificial) boundaries and combine the research performed within each sub-field." We propose that future research should indeed consolidate well-established findings of psycholinguistics with developments in computational linguistics to yield a research program that is on one hand informed by our understanding of early language acquisition, and on the other hand is rigorously defined and robustly evaluated.

We list below some desiderata for the kind of computational models that we envision.

Data. Unlike much work in the area of grammar induction algorithms (Section 4.2), research concerned with child language acquisition must be trained (and evaluated) on dedicated corpora, of the kind exhibited by CHILDES. Ideally, they should be tested on more than one language. Models can be trained on both child and child-directed speech.

Task. We suggest that computational models of language acquisition focus on the easier task of learning language as a set of strings, leaving the induction of syntactic structures to future research.

Grammar. Cognitive works (Section 4.1) tend to be more vague on the formal properties of the class of languages admitted by the models they suggest. A good model must be explicit on this point. We believe that a reasonable language class for early language (up to, say, three years of age) is a proper subset of the regular languages. Models that can learn unrestricted context-free languages, for example, miss an important point and are likely to over-generate.

Model. Works that are specifically designed to model early child language acquisition should incorporate into the framework biases that reflect psycholinguistic models of acquisition processes [46, 42]. These include item-based learning, rote learning [45], left-edge biases [30], etc.

Evaluation. Clearly, evaluation is still an unsolved problem (Section 3), and much work is still needed in this area. Still, and in contrast to some of the approaches described in section 4.1, any computational model of language acquisition must be rigorously evaluated on real data.

To further emphasize this last point, Kol et al. [36] conducted an alternative evaluation of the *traeback* model [38, 26, 39, 6, 67]. They show that the original evaluation scheme in these works is lacking, as it focuses on recall but completely ignores precision. As a measure of over-generation, Kol et al. [36] apply the traceback method not just to child utterances in the test corpus, but also to the same utterances in reverse order. While the model can generate 64-64% of the genuine child utterances (showing reasonable recall), it can also generate 42-50% of the reverse utterances, indicating a serious over-generation problem.

One of the reasons for this problem is that the traceback model is not defined in a sufficiently rigorous way. The sets of operations allowed for traceback is not fixed, and changes from one work to another. Much of the work involved in applying the model to data is done manually in a way that prohibits computational re-implementation.

In contrast, works such as those discussed in Section 4.2 are often evaluated on WSJ10, a clearly inadequate corpus for assessing child language development. Some of them assume that the data are already annotated with parts of speech or even syntactic information, an obviously unacceptable assumption when child language is concerned. Language learning from sequences of POS-tags is a particularly bad example of how to model child language acquisition.

Clearly, then, the benefits of these different approaches must be consolidated in order for a formal, computational, linguistically- and cognitively-informed model of language acquisition to emerge. Such a model must be rigorously defined, in a way that lends itself to computational implementation; formally, it should exhibit highly-retricted computational expressivity; it should employ biases that correspond to established observations of child language research (such as item-based learning, rote learning [45], left-edge biases [30], adherence to stages of acquisition [19, 10], etc.) Only future works that will correspond to such considerations will properly address the criticism of Bod [14], whereby “almost any current linguistic theory ... has given up on the construction of a precise, testable model of language use and language acquisition.”

Acknowledgements

I am grateful to Sheli Kol, Alon Lavie, Brian MacWhinney and Bracha Nir for their continuous support. This research was supported in part by Grant No. 2007241 from the United States-Israel Binational Science Foundation (BSF) and by the National Science Foundation (NSF) under grant IIS-0414630.

References

[1] Adriaans, P.: Language Learning from a Categorial Perspective. PhD thesis, Universiteit van Amsterdam (1992)

[2] Adriaans, P.: Learning shallow context-free languages under simple distributions. In: Copestake, A., Vermeulen, K. (eds.) Algebras, Diagrams and Decisions in Language, Logic and Computation. CSLI/CUP, Stanford (2001)

[3] Adriaans, P., Vervoort, M.: The EMILE 4.1 grammar induction toolbox. In: Adriaans, P.W., Fernau, H., van Zaanen, M. (eds.) ICGI 2002. LNCS (LNAI), vol. 2484, pp. 293–295. Springer, Heidelberg (2002)

[4] Adriaans, P.W., van Zaanen, M.M.: Computational grammatical inference. In: Holmes, D.E., Jain, L.C. (eds.) Innovations in Machine Learning. Studies in Fuzziness and Soft Computing, vol. 194, ch. 7. Springer, Heidelberg (2006)

[5] Banko, M., Moore, R.C.: Part of speech tagging in context. In: COLING 2004: Proceedings of the 20th international conference on Computational Linguistics, Morristown, NJ, USA, p. 556. Association for Computational Linguistics (2004)

[6] Bannard, C., Lieven, E.: Repetition and reuse in child language learning. In: Corrigan, R., Moravcsik, E., Ouali, H., Wheatley, K. (eds.) Formulaic Language. John Benjamins, Amsterdam (2009)

[7] Bannard, C., Lieven, E., Tomasello, M.: Early grammatical development is piecemeal and lexically specific. Proceedings of the National Academy of Science 106(41), 17284–17289 (2009)

[8] Bates, E., MacWhinney, B.: Competition, variation, and language learning. In: [46], ch. 6, pp. 157–193 (1987)

[9] Berant, J., Gross, Y., Mussel, M., Sandbank, B., Edelman, S.: Boosting unsupervised grammar induction by splitting complex sentences on function words. In: Proceedings of the 31st Boston University Conference on Language Development, pp. 93–104. Cascadilla Press (2007)

[10] Berman, R.A.: Between emergence and mastery: The long developmental route of language acquisition. In: Berman, R.A. (ed.) Language development across childhood and adolescence. Trends in Language Acquisition Research, vol. 3, pp. 9–34. John Benjamins, Amsterdam/Philadelphia (2004)

[11] Bod, R.: An all-subtrees approach to unsupervised parsing. In: ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Morristown, NJ, USA, pp. 865–872. Association for Computational Linguistics (2006a)

[12] Bod, R.: Unsupervised parsing with U-DOP. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), New York City, pp. 85–92. Association for Computational Linguistics (2006b)

[13] Bod, R.: Is the end of supervised parsing in sight? In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 400–407. Association for Computational Linguistics (2007)

[14] Bod, R.: Constructions at work or at rest? Cognitive Linguistics 20(1) (2009)

[15] Bod, R., Sima'an, K., Scha, R. (eds.): Data-Oriented Parsing. CSLI Publications, Stanford (2003)

[16] Borensztajn, G., Zuidema, W.: Bayesian model merging for unsupervised constituent labeling and grammar induction. ILLC Prepublication PP-2007-40, ILLC, University of Amsterdam (2007)

[17] Borensztajn, G., Zuidema, J., Bod, R.: Children's grammars grow more abstract with age — evidence from an automatic procedure for identifying the productive units of language. In: Proceedings of CogSci 2008 (2008)

[18] Brodsky, P., Waterfall, H., Edelman, S.: Characterizing motherese: On the computational structure of child-directed language. In: Proceedings of the 29th Cognitive Science Society Conference. Cognitive Science Society (2007)

[19] Brown, R.: A first language: the Early stages. Harvard University Press, Cambridge (1973)

[20] Chang, F., Lieven, E., Tomasello, M.: Automatic evaluation of syntactic learners in typologically-different languages. Cognitive Systems Research 9(3), 198–213 (2008)

[21] Chomsky, N.: Aspects of the theory of syntax. MIT Press, Cambridge (1965)

[22] Chomsky, N.: Language and Mind. Harcourt Brace Juvanovich, New York (1968)

[23] Chomsky, N.: Rules and representations. Behavioral and Brain Sciences 3, 1–61 (1980)

[24] Chomsky, N.: Reflections on Language. Pantheon, New York (1975)

[25] Church, K.W., Mercer, R.L.: Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics 19(1), 1–24 (1993)

[26] Dąbrowska, E., Lieven, E.: Towards a lexically specific grammar of children's question constructions. Cognitive Linguistics 16(3), 437–474 (2005)

[27] Edelman, S., Waterfall, H.: Behavioral and computational aspects of language and its acquisition. Physics of Life Reviews 4(4), 253–277 (2007)

[28] Freudenthal, D., Pine, J.M., Gobet, F.: Modelling the development of children's use of optional infinitives in Dutch and English using MOSAIC. Cognitive Science 30, 277–310 (2006)

[29] Freudenthal, D., Pine, J.M., Gobet, F.: Understanding the developmental dynamics of subject omission: the role of processing limitations in learning. Journal of Child Language 34(01), 83–110 (2007)

[30] Freudenthal, D., Pine, J.M., Gobet, F.: Simulating the referential properties of Dutch, German, and English root infinitives in MOSAIC. Language Learning and Development 5, 1–29 (2009)

[31] Kennedy, G.: An introduction to corpus linguistics. Addison Wesley, Reading (1998)

[32] Klein, D., Manning, C.D.: Natural language grammar induction using a constituent-context model. In: Dietterich, T.G., Becker, S., Ghahramani, Z., Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) NIPS, pp. 35–42. MIT Press, Cambridge (2001)

[33] Klein, D., Manning, C.D.: A generative constituent-context model for improved grammar induction. In: ACL, pp. 128–135 (2002)

[34] Klein, D., Manning, C.D.: Corpus-based induction of syntactic structure: Models of dependency and constituency. In: ACL, pp. 478–485 (2004)

[35] Klein, D., Manning, C.D.: Natural language grammar induction with a generative constituent-context model. Pattern Recognition 38(9), 1407–1419 (2005)

[36] Kol, S., Nir, B., Wintner, S.: Acquisition of abstract slot-filler schemas: Computational evaluation. Presented at the COGSCI 2009 Workshop on Psychocomputational Models of Human Language Acquisition (2009)

[37] Li, P., Farkas, I., MacWhinney, B.: Early lexical development in a self-organizing neural network. Neural Networks 17(8-9), 1345–1362 (2004)

[38] Lieven, E., Behrens, H., Speares, J., Tomasello, M.: Early syntactic creativity: a usage-based approach. Journal of Child Language 30(2), 333–370 (2003)

[39] Lieven, E., Salomo, D., Tomasello, M.: Two-year-old children's production of multiword utterances: a usage-based analysis. Cognitive Linguistics 20(3), 481–507 (2009)

[40] Lieven, E.V., Pine, J.M., Baldwin, G.: Lexically-based learning and early grammatical development. Journal of Child Language 24(1), 187–219 (1997)

[41] MacWhinney, B.: The CHILDES Project: Tools for Analyzing Talk, 3rd edn. Lawrence Erlbaum Associates, Mahwah (2000)

[42] MacWhinney, B.: Models of the emergence of language. Annual Review of Psychology 49, 199–227 (1998)

[43] MacWhinney, B.: A multiple process solution to the logical problem of language acquisition. Journal of Child Language 31, 883–914 (2004a)

[44] MacWhinney, B.: A unified model of language acquisition. In: Kroll, J., De Groot, A. (eds.) Handbook of bilingualism: Psycholinguistic approaches. Oxford University Press, Oxford (2004b)

[45] MacWhinney, B.: Rules, rote, and analogy in morphological formations by Hungarian children. Journal of Child Language 2, 65–77 (1975)

[46] MacWhinney, B. (ed.): Mechanisms of language acquisition. Lawrence Erlbaum Associates, Hillsdale (1987)

[47] The emergence of language. In: MacWhinney, B. (ed.) Carnegie Mellon Symposia on Cognition. Lawrence Erlbaum Associates, Mahwah (1999)

[48] Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn treebank. Computational Linguistics 19(2), 313–330 (1993)

[49] McEnery, A., Wilson, A.: Corpus Linguistics. Edinburgh University Press, Edinburgh (1996)

[50] Pinker, S.: The Language Instinct. William Morrow and Company, New York (1994)

[51] Rowland, C.F., Fletcher, S.L., Freudenthal, D.: Repetition and reuse in child language learning. In: Behrens, H. (ed.) Corpora in Language Acquisition Research: History, methods, perspectives, pp. 1–24. John Benjamins, Amsterdam (2008)

[52] Sagae, K., MacWhinney, B., Lavie, A.: Automatic parsing of parent-child interactions. Behavior Research Methods, Instruments, and Computers 36, 113–126 (2004)

[53] Sagae, K., Davis, E., Lavie, A., MacWhinney, B., Wintner, S.: High-accuracy annotation and parsing of CHILDES transcripts. I. In: Proceedings of the ACL-2007 Workshop on Cognitive Aspects of Computational Language Acquisition, Prague, Czech Republic, pp. 25–32. Association for Computational Linguistics (2007)

[54] Sagae, K., Davis, E., Lavie, A., MacWhinney, B., Wintner, S.: Morphosyntactic annotation of CHILDES transcripts. Journal of Child Language (to appear)

[55] Seginer, Y.: Fast unsupervised incremental parsing. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp. 384–391. Association for Computational Linguistics (2007)

[56] Smith, N.A., Eisner, J.: Annealing techniques for unsupervised statistical language learning. In: ACL 2004: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, p. 486. Association for Computational Linguistics (2004)

[57] Solan, Z., Horn, D., Ruppin, E., Edelman, S.: Unsupervised learning of natural languages. Proceedings of the National Academy of Sciences of the United States of America 102(33), 11629–11634 (2005)

[58] Stolcke, A., Omohundro, S.M.: Inducing probabilistic grammars by bayesian model merging. In: Carrasco, R.C., Oncina, J. (eds.) ICGI 1994. LNCS, vol. 862, pp. 106–118. Springer, Heidelberg (1994)

[59] Tomasello, M.: On the different origins of symbols and grammars. In: Christiansen, M.H., Kirby, S. (eds.) Language Evolution. Studies in the Evolution of Language, ch. 6, pp. 94–110. Oxford University Press, Oxford (2003)

[60] Tomasello, M.: Acquiring linguistic constructions. In: Kuhn, D., Siegler, R. (eds.) Handbook of Child Psychology, pp. 255–298. Wiley, New York (2006)

[61] Tomasello, M.: Language is not an instinct. Cognitive Development 10, 131–156 (1995)

[62] van Zaanen, M.: Implementing alignment-based learning. In: Adriaans, P.W., Fernau, H., van Zaanen, M. (eds.) ICGI 2002. LNCS (LNAI), vol. 2484, pp. 312–314. Springer, Heidelberg (2002)

[63] van Zaanen, M.: ABL: alignment-based learning. In: Proceedings of the 18th conference on Computational linguistics, Morristown, NJ, USA, pp. 961–967. Association for Computational Linguistics (2000)

[64] van Zaanen, M.: Bootstrapping Structure into Language: Alignment-Based Learning. PhD thesis, University of Leeds, Leeds, UK (2002a)

[65] van Zaanen, M., Adriaans, P.: Alignment-Based Learning versus EMILE: A comparison. In: Proceedings of the Belgian-Dutch Conference on Artificial Intelligence (BNAIC), Amsterdam, The Netherlands, pp. 315–322 (2001)

[66] van Zaanen, M., Geertzen, J.: Problems with evaluation of unsupervised empirical grammatical inference systems. In: Clark, A., Coste, F., Miclet, L. (eds.) ICGI 2008. LNCS (LNAI), vol. 5278, pp. 301–303. Springer, Heidelberg (2008)

[67] Vogt, P., Lieven, E.: Verifying theories of language acquisition using computer models of language evolution. In: Adaptive Behavior Special issue on Language Evolution: Computer models for Empirical Data (forthcoming)