J. Child Lang. **37** (2010), 477–485. © Cambridge University Press 2010
doi:10.1017/S0305000910000139

Computational models of child language learning: an introduction*

BRIAN MACWHINNEY

(*First published online* 25 *March* 2010)

This special issue showcases recent work on the computational modeling of child language acquisition. Together, these nine papers provide testimony to the scientific value of corpus-driven computational modeling. Each of the papers presents a clear, mechanistic model that can be tested, refined or rejected on the basis of publicly available data and/or replicable experiments. However, for many readers, the multiple formalisms and technicalities involved in this type of work can serve as barriers to evaluating the nature of the contributions being made. Therefore, it is my goal in this introduction to summarize what I see as the important take-home messages delivered by each of the nine projects. Hopefully, this overview will encourage the reader to turn then to the details of each of the nine contributions.

Of the nine papers included here, eight base their analysis on the corpora of spontaneous adult–child interactions made available through the Child Language Data Exchange System (CHILDES). These CHILDES corpora provide two types of information crucial to the modeling enterprise. First, they document in detail the naturalistic development of language in the child. Second, these corpora provide a good sampling of the adult speech that serves as input to the child's language learning mechanisms. Given these two empirical bases, the job of the computational modeler is to determine a set of algorithms that can take the child-directed speech (CDS) as input and produce the learner's output (LO) at successive developmental levels. We can refer to this approach as input–output (I–O) modeling. In its simplest form, I–O modeling tends to view language learning as an emergent, data-driven process. However, there is ample room within this same computational framework for precise statements regarding the operation of non-emergentist innate constraints, parameters, principles and universals. Furthermore, complex features of situational context, social understandings and recent dialog history can, in principle, be quantified and coded as features of the input. More generally, as long as long as all these pieces of the Language Acquisition Device (LAD; Chomsky 1965) are fully specified, all

[*] Address for correspondence: Department of Psychology 254M Baker, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh PA 15213. Tel: 412 268-3793. Fax: 412 268-3905. E-mail: macw@cmu.edu

current theoretical approaches can be articulated and evaluated within the framework of computational modeling.

Of course, the job of specifying and coding all of the features of the input, the output, and the learning algorithms is exceedingly complex. Faced with this complexity, modelers have taken the reasonable approach of building accounts for specific linguistic subdomains, rather than the whole of language acquisition. The hope is that these separate pieces could eventually be reassembled into a more integrated view of language acquisition. The articles in this issue present models in the areas of syntax, phonology and lexical segmentation. To assist the reader in navigating through this issue, the articles are ordered in terms of increasing abstraction of the relevant linguistic skills from word segmentation up to discourse, and my discussion below follows this same order. After completing this overview, I will evaluate the overall activity in terms of the extent to which it has been illuminated our understanding of children's language learning.

The first three articles focus on the modeling of segmentation, or the processes by which children pull out words from the continuous stream of input speech, even before they have learned the meanings of these words. Each of these models relies on the CDS in the English CHILDES database. However, each model uses a different segment of CHILDES as its target. Blanchard, Heinz and Golinkoff use the Bernstein-Ratner corpus of speech to children between the ages of 1;1 and 1;11. To facilitate processing of these data, Michael Brent created a training corpus now available on the web from http://childes.psy.cmu.edu/derived/. Brent, Venkataraman, Johnson and Goldwater, and Fleck have all used this corpus to evaluate their various models. A great advantage of the use of such standard corpora is that it facilitates the replication and comparison of results and analyses across alternative models. This method of focusing on standardized corpora has led to significant progress in areas such as speech recognition, and it seems that the same method can easily apply to work on language learning. Mark Johnson has contributed a similar target corpus derived from Demuth's Sesotho corpus, which is also available at http://childes.psy.cmu.edu/derived/.

The other two models of segmentation in this special issue choose to use other segments of the CHILDES database, each for different reasons. Rytting, Brew and Fosler-Lussier used 13,433 utterances from the Brent-Siskind database of speech from mothers to their children between the ages of 0;9 and 1;3. Crucially, these data were recorded using high standard audio devices, making the recordings plausible candidates for automatic speech recognition (ASR). The speech stream was then broken into phones, using adult acoustic models and an adult phonemic lexicon. Although one might argue that such adult knowledge is outside the scope of the infant's abilities, one must remember that the goal here is not to model

segmentation into phones, but to model the segmentation of a string of phones into words. In this regard, the analysis by Rytting *et al*. constitutes a significant step forward in terms of linking the word segmentation task to the actual input available to the child, although it achieves this advance by introducing other, less problematic, assumptions.

In their PUDDLE model, Monaghan and Christiansen took CDS from six corpora from the English segment of CHILDES and passed them through the Festival speech synthesizer to derive a set of phonemes. Although this method cannot create the level of phonetic detail in the ASR approach of Rytting *et al*., it provides a convenient way of bringing the representation of the CDS more in line with the actual input.

The PHOCUS and PUDDLE models share three important assumptions. First, both models maintain a lexicon that is used to record information from previously processed utterances. In this regard, they differ radically from the ASR model of Rytting *et al*., which relies instead on a simple recurrent net (SRN) that can only record lexical items indirectly if at all. This may be the reason that PHOCUS and PUDDLE outperform the ASR-based system in terms of their ability to correctly segment CDS. It would be interesting to know whether the success of learning from ASR-processed material could be improved through use of such lexicon-based models. Second, both PHOCUS and PUDDLE assume that items in the lexicon are accessed in order of frequency. This is an obviously reasonable assumption. However, it may be more important for processing functor words than content words, as suggested by Table 3 in Monaghan and Christiansen. It makes sense to believe that children store functors as high-frequency sounds, even without knowing their meaning and without yet using them in their own productions. For content words, it is more likely that they are acquired as full meaningful lexical items. However, current models of lexical learning seldom provide a way of distinguishing between stored auditory forms and actual lexical items.

The third assumption shared by the PHOCUS and PUDDLE models relates to the importance of phonotactics. In PHOCUS, phonotactics plays a crucial role in defeating over-segmentation. The idea is that you do not want the system to extract 'n' as a possible word by locating the sound 'oh' inside 'no'. PHOCUS achieves this by invoking a universal constraint that words must have at least one syllabic element, and 'n' would fail in some versions of this constraint. In addition, PHOCUS allows for a complete incremental computation of the transitional probabilities between pairs of phonemes. PUDDLE computes phonotactics by encoding the initial and final segments of words that have been already stored in the lexicon and then using these as guesses about how words can begin or end. If phonotactics are computed and stored independently of lexical representations, the approach used in PHOCUS would seem more defensible.

However, it is not yet clear whether phonotactics are indeed stored in some long-term memory independently from the lexicon.

The success of corpus-based computational models of segmentation raises a more general issue. It may be possible to create a system that would achieve a fairly high level of segmentation accuracy based on use of strong assumptions, extensive abstracted input, a large memory and intensive processing. But do children or beginning second language learners actually achieve a high level of accurate segmentation before lexical learning kicks in to bootstrap the process and eventually solve the problem? If we rely on our intuitions derived from second language learning, most of us would answer this question in the negative. At the same time, we know that learners end up solving the segmentation problem nearly perfectly. Although phonotactics and stress patterns may play an initial role in segmentation, it appears that the bulk of the solution stems from the growth of the lexicon. Both PHOCUS and PUDDLE provide a role for the lexicon, but not for a lexicon that involves true lexical learning. Bringing these two issues together in the context of the more realistic input derived from ASR would seem to be the next challenge for this line of research.

In the fourth article, Jarosz examines the extent to which frequency can modulate the effects of implicational markedness in determining the order of acquisition of accurate production of syllable types. Jarosz begins by noting that five constraints that have figured in the literature on Optimality Theory (OT) can also be expressed in terms of an implicational hierarchy. This hierarchy encodes claims such as 'any language that permits VC syllable types also permits the less marked V, CVC and CV syllable types'. In cases where two syllable types are both allowed by the implicational hierarchy, frequency will play a secondary, but determining role. For example, in Dutch, once the CV syllable type has been learned, the next possibilities are CVC, CCV, CVCC and V. Here, the fact that over 50 percent of the syllables in Dutch CDS are CVCs is enough to determine their acquisition before the other three. Although the Dutch data support the frequency hypothesis, it is also possible that additional markedness relations not yet fully explored by typological theory could explain these patterns. To further assess this issue, Jarosz looks at CHILDES data from the Weist Polish corpus. Examining the CDS in this Polish corpus, Jarosz finds that word-initial clusters are three times as frequent as word-final clusters. Moreover, this difference correctly predicts the order of accurate production of syllables with either complex onsets or codas in Polish. Jarosz then contrasts three computational models of this process: Stochastic OT, MLG (Maximum Likelihood Learning of Lexicons and Grammars) and HG (Harmonic Grammar). He states that all three models incorporate the same core claims of the frequency hypothesis regarding the primacy of markedness. This may be true. However, in terms of their inner workings

and assumptions the models are radically different. For example, Stochastic OT relies on learning through error, whereas MLG learns only through exposure to input data. All of the models are able to account for the basic frequency ordering effects but, as Jarosz notes, this particular domain may tend to minimize our ability to discern important differences. As we push these models to account for additional languages and more detailed aspects of acquisition, we also up the ante on our need for accurately encoded phonological data in CHILDES. Hopefully, the introduction of the new PhonBank project and the Phon analysis program will allow researchers to continue to explore these interesting interactions between markedness and frequency.

The final five articles in the special issue all examine aspects of syntactic development. This group begins with an article by Perfors designed to show how Hierarchical Bayesian Models (HBM) can be used for the learning of dative constructions. The goal is to formalize a mechanism that is capable of learning on the basis of positive evidence without relying exclusively on negative evidence or conservative item-based learning to avoid possible overgeneralizations. As such, this paper contributes to the ongoing discussion of multiple potential solutions to the Logical Problem of Language Acquisition. In a target article in this journal a few years ago (MacWhinney, 2004), I argued that mechanisms such as conservatism, preemption, and entrenchment are best understood in terms of the notion of competition. In the case of Perfors' examples, the competition is between the correct item-based (MacWhinney, 1975) construction 'confess X to Y' and the feature-based double object construction that would incorrectly allow 'confess Y X'. In this case, there is no support in the input for the item-based pattern 'confess Y X'. Therefore, children are very unlikely to produce *John confessed Bill the truth*. The beauty of the HBM approach is that it provides a way of formalizing these observations. Moreover the HBM formalism was not originally created for this type of problem, but for word and category learning, suggesting that it represents a rather general cognitive mechanism. Perfors demonstrates the adequacy of the HBM model by applying it to experimental data from Wonnacott *et al.* (2008), as well as to the CDS speech of Brown's Adam corpus in CHILDES. Of course, it would be good to extend this type of analysis to other children, constructions and languages. But what is important is how Perfors has converted informal observations into a programmable formalism. In this way, her analysis represents a major step forward that deserves to be studied carefully.

Freudenthal, Pine and Gobet present a comparison of two formalized accounts of syntactic learning. Specifically, the authors compare their MOSAIC model with the VLM (Variational Learning Model) of Legate and Yang in terms of the ability to accurately predict optional infinitive

(OI) errors in English, Dutch, German, French and Spanish. This work, like other studies that have examined OIs, relies on corpora available from the CHILDES database. Because each of the five languages being studied has a very different system for marking verb morphology, the exact shape of OI errors varies markedly. In an English error, such as **Mommy go to work*, the form *go* can be viewed as the core of the infinitival form or it can be viewed as a verb with the wrong agreement marking. However, in the French error **La poupée dormir* it is clear that the infinitive is replacing a tensed verb. Both VLM and MOSAIC assume that language learning involves changes in the grammar as a result of probabilistic features of the input. Although both models use probabilities, MOSAIC relies far more extensively on the statistics of individual verbs as they are recorded in edge-based patterns. It is this reliance on item- or phrase-specific data that provides evidence in favor of MOSAIC over VLM. Given these advantages in favor of MOSAIC, one wonders why a generativist analysis such as VLM could not just be extended to include some of the detailed phrasal information available to MOSAIC. If these successive refinements to MOSAIC and VLM could be iterated with the goal for each being to make a closer and closer match to the data, one wonders whether, in the end, the models would not end up being very similar indeed.

In fact, it may be necessary to provide more structure to MOSAIC. Currently, it learns by building up a recorded stock of utterance-final and utterance-initial sequences. The system is biased to prefer the former over the latter. In addition, it is able to link the two. This linkage is needed to account for errors such as **where he go*. An alternative approach, suggested by MacWhinney (1982) and others is to treat *where* as the center of an item-based construction for which *he go* is the filler. It would be interesting to see a comparison of the item-based learning approach with edge-based learning for the OI data. As the authors note, another factor that must be considered is the role of the verbal conjugational paradigm. Both VLM and MOSAIC have a hard time predicting the high level of OI errors in English. However, if one notes that the English present tense verb and the imperative use the same form as the infinitive everywhere except the third person singular, then the high level of OI errors in English becomes unsurprising. In fact, morphological ambiguity seems to correctly predict the order English > German > Dutch > French > Spanish. But the authors show convincingly why paradigm-based ambiguity cannot be the whole story, although it may modulate effects from edge-based learning.

Of the various papers in this special issue, the contribution by Waterfall, Sandbank, Onnis and Edelman could be viewed as the most ambitious. Its goal is the construction of a complete, data-driven, empirical generative model of the learning of syntactic patterns. The authors argue that the target of a generative grammar should not be producing all and only the

acceptable sentences of a language, but rather producing a probability distribution over sentences that matches their actual occurrence in corpora. In particular, probabilistic context-free grammars (PCFGs) match this requirement and have been shown to be capable, in principle, of inducing grammars without relying on negative evidence. The authors present a method for building such grammars, called ConText, which they then apply to a 300,000-sentence training corpus from the CDS in the English corpora in CHILDES. The resulting grammar induces a variety of new syntactic classes based on co-occurrence patterns in the input. To evaluate the accuracy of this induction, the authors asked undergraduates to judge the acceptability of 100 sentences produced by ConText, mixed randomly with 100 utterances from the CDS in CHILDES. For sentences up to six words in length, subjects judged ConText sentences to be only somewhat worse than actual adult sentences. However, for longer sentences, the acceptability of ConText productions declined markedly.

Despite the success of learning in ConText, the authors recognize that the overall precision and accuracy of the system is inadequate. They suggest that additional cues available in the discourse structure of interactions may provide information crucial for improving performance. In particular, they explore the role of sequences of sentences in CDS called variation sets. In these variation sets, parents repeat stretches of words across utterances, thereby facilitating alignment. An example would be the sequence: *You got to push them to school. Push them to school. You got to take them to school.* In this sequence, children can learn to align units such as *push them* or *you got to*. Extraction of these aligned segments can lead to the formation of new classes and the clearer processing of subsequent sentences. The authors extracted the variation sets found in their corpus of 300,000 CDS utterances and showed that they were rich in structures that would be helpful in guiding syntactic learning. Although this work is still very much in progress, it is shining a light on a very promising path that links together basic algorithms for pattern extraction with utilization of specific additional cues from discourse and interactional structure.

Sagae, Davis, Lavie, MacWhinney and Wintner provide an update on a project seeking to construct automatic grammatical relation taggers for CHILDES corpora in various languages. The current article only reports on progress for English and Spanish. However, there is also progress on taggers for Japanese, Mandarin and Hebrew. The hope is that, eventually, all CHILDES corpora will be tagged for both part of speech and grammatical relations. This work has two central goals. First, once tagged, the corpora will greatly facilitate the testing of theories about language acquisition. A clear case in point is the comparison between VLM and MOSAIC as reviewed by Freudenthal *et al*. Once fully tagged grammatical corpora become available, it will be possible to run these two algorithms and others,

such as ConText, automatically across the entire database for a given language to evaluate their relative success rates. Because this process will be automatic, it will be possible to make the methods of analysis fully public to promote testing and replication. Second, it will be possible to develop automatic versions of tests of syntactic development like DSS, IPSyn or LARSP using these grammatical relations tags. Automated versions of these tests will greatly improve and expand the use of spontaneous speech samples for clinical and research study of language disorders.

The final article, by van Rij, van Rijn and Hendriks, is unlike the other eight in that it uses computer programs to model experimental rather than corpus data. The particular target of the model is the comparison of Principles A and B of the binding theory. Principle A holds that reflexives must be bound within their governing domain. Children learn this by age 3;0. Principle B holds that the pronoun in a sentence such as *The penguin is hitting him with a pan* cannot be reflexive. Children fail to pick up on this until after age 6;0, resulting in the Delay of Principle B Effect (DPBE). Following OT analysis, the authors claim that Principle A is a faithfulness constraint that applies to both comprehension and production, whereas Principle B is a markedness constraint that only applies to production. This explains why children use pronouns and reflexives correctly in production, while making errors on pronouns in comprehension. In order to implement Principle B for comprehension, older children must impose bidirectional optimization, which is a formal way of saying that they must learn to take the listener's perspective. The authors' ACT-R simulation models the basic DPBE by allowing Principle A to compete with the AVOID PRONOUNS constraint when children are listening to pronouns. This leads to not only greater error, but also slower reaction times, since more steps are involved in processing the OT relationships. In principle, children could correct this by applying the second perspective-switching phrase, but that would take time and would place excessive demands on their computational mechanism. Over time, children overcome this limitation by compiling productions that implement perspective switching. To test this, the authors conducted a study using a slowed speech rate that was predicted to help in overcoming DPBE. The results of this test were complex and I will leave the reader to evaluate the status of the predictions. However, apart from the empirical test, this model breaks new ground in terms of linking theoretical syntax to an implementation within a generalized cognitive architecture with important developmental and processing implications.

I hope that my discussion of these nine articles has given the reader a sense of the importance of this work. Although I am clearly an enthusiastic proponent of the computational analysis of child language corpora,

I nonetheless recognize that this work faces a serious challenge. These models have tended to focus on a very narrow range of the overall problem of language acquisition. The hope is that, by making this narrow focus sufficiently precise, the models will illuminate important details of the language learning process. A strength of this approach is that it can be used to provide a clear characterization of the relevant pieces of the input in CDS. We can have clear information about the distribution of optional infinitives, syllabic patterns, dative constructions or phonotactic regularities, simply by studying the CDS in the CHILDES corpora. The problem I see is that this approach eliminates any real role for interactive and incremental learning. On the one hand, it makes assumptions that are too strong by treating the child as a computer that can repeatedly scan over millions of sentences of input in batch mode. On the other hand, it makes assumptions that are too weak by failing to detect the importance of methods – such as variation sets, replacement sequences and fine-tuning – that parents and the children use collaboratively to bootstrap learning. In order to correct these problems, it seems to me that the best solution will be to integrate models across linguistic levels. Segmentation models will have to be linked more closely to lexical learning; lexical learning will have to be linked to syntactic processing; and everything from pronominal processing to collocation learning will have to be embedded in discourse context. Building these more complex models will take time and energy, but the results will provide increasingly interesting and valuable insights into language learning.

REFERENCES

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

MacWhinney, B. (1975). Pragmatic patterns in child syntax. *Stanford Papers And Reports on Child Language Development* **10**, 153–65.

MacWhinney, B. (1982). Basic syntactic processes. In S. Kuczaj (ed.), *Language acquisition: Volume 1. Syntax and semantics*, 73–136. Hillsdale, NJ: Lawrence Erlbaum.

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

Wonnacott, E., Newport, E. & Tanenhaus, M. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology* **56**, 165–209.