Learning Simple Statistics for Language Comprehension and Production: The CAPPUCCINO Model

Stewart M. McCauley (smm424@cornell.edu)
Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853 USA

Abstract

Whether the input available to children is sufficient to explain their ability to use language has been the subject of much theoretical debate in cognitive science. Here, we present a simple, developmentally motivated computational model that learns to comprehend and produce language when exposed to child-directed speech. The model uses backward transitional probabilities to create an inventory of 'chunks' consisting of one or more words. Language comprehension is approximated in terms of shallow parsing of adult speech and production as the reconstruction of the child's actual utterances. The model functions in a fully incremental, on-line fashion, has broad cross-linguistic coverage, and is able to fit child data from Saffran's (2002) statistical learning study. Moreover, word-based distributional information is found to be more useful than statistics over word classes. Together, these results suggest that much of children's early linguistic behavior can be accounted for in a usage-based manner using distributional statistics.

Keywords: Language Learning; Computational Modeling; Corpora; Chunking; Shallow Parsing; Usage-Based Approach

Introduction

The ability to produce and understand a seemingly unbounded number of different utterances has long been hailed as a hallmark of human language acquisition. But how is such open-endedness possible, given the much more limited nature of other animal communication systems? And how can a child acquire such productivity, given input that is both noisy and necessarily finite in nature? For nearly half a century, generativists have argued that human linguistic productivity can only be explained by positing a system of abstract grammatical rules working over word classes and scaffolded by considerable innate language-specific knowledge (e.g., Pinker, 1999). Recently, however, an alternative theoretical perspective on linguistic productivity has emerged in the form of usage-based approaches to language (e.g., Tomasello, 2003). This perspective is motivated by analyses of child-directed speech, showing that there is considerably more information available in the input than previously assumed. For example, distributional and phonological information can provide reliable cues for learning about lexical categories and phrase structure (for a review, see Monaghan & Christiansen, 2008). Behavioral studies have shown that children can use such information in an item-based manner (Tomasello, 2003).

A key difference between generative and usage-based approaches pertains to the granularity of the linguistic units necessary to account for the productivity of human language. At the heart of usage-based theory lies the idea that grammatical knowledge develops gradually through abstraction over multi-word utterances (e.g., Tomasello, 2003), which are assumed to be stored as multi-word 'chunks.' Testing this latter assumption, Bannard and Matthews (2008) showed not only that non-idiomatic chunk storage takes place, but also that storing such units actively facilitates processing: young children repeated multi-word sequences faster, and with greater accuracy, when they formed a frequent chunk. Moreover, Arnon and Snider (2010) extended these results, demonstrating an adult processing advantage for frequent phrases. The existence of such chunks is problematic for generative approaches that have traditionally clung to a words-and-rules perspective, in which memory-based learning and processing are restricted to the level of individual words (e.g., Pinker 1999).

One remaining challenge for usage-based approaches is to provide an explicit computational account of language comprehension and production based on multi-word chunks. Although Bayesian modeling has shown that chunk-based grammars are in principle sufficient for the acquisition of linguistic productivity (Bannard, Lieven, & Tomasello, 2009), no full-scale computational model has been forthcoming (though models of specific aspects of acquisition do exist, such as the optional infinitive stage; Freudenthal, Pine & Gobet, 2009). The scope of the computational challenge facing usage-based approaches becomes even more formidable when considering the success with which the generativist principles of words and rules have been applied in computational linguistics. In this paper, we take an initial step towards answering this challenge by presenting the 'Comprehension And Production Performed Using Chunks Computed Incrementally, Non-categorically, and On-line' (or CAPPUCCINO) model of language acquisition.

The aim of the CAPPUCCINO model is to provide a test of the usage-based assumption that children's language use may be explained in terms of stored chunks. To this end, the model gradually builds up an inventory of chunks consisting of one or more words—a 'chunkatory'—used for both language comprehension and production. The model was further designed with several key psychological and computational properties in mind: a) *incremental learning:* at any given point in time, the model can only rely on the input seen so far (no batch learning); b) *on-line processing:* input is processed word-by-word as it is encountered; c) *simple statistics:* learning is based on computing backward transitional probabilities (which 8-month-olds can track; Pelucchi, Hay, & Saffran, 2009); d) *comprehension:* the

model segments the input into chunks comparable to the output of a shallow parser; e) *production:* the model reproduces the child's actual utterances; f) *naturalistic input:* the model learns from child-directed speech; g) *cross-linguistic coverage:* the model is exposed to a typologically diverse set of languages (including Sesotho, Tamil, Estonian, and Indonesian).

In what follows, we first describe the basic workings of the CAPPUCCINO model, its comprehension performance across English, German, and French, and its production ability across 13 different languages. Next, we demonstrate that the model is capable of closely fitting child data from a statistical learning study (Saffran, 2002). Finally, we discuss the limitations of the current model.

Simulation 1: Modeling Comprehension and Production in Natural Languages

The CAPPUCCINO model performed two tasks: comprehension of child-directed speech through the discovery and use of chunks, and sentence production through the use of the same chunks and statistics as in comprehension. Comprehension was approximated in terms of the model's ability to segment a corpus into phrasal units, and production in terms of the model's ability to reconstruct utterances produced by the child in the corpus. Thus, the model sought to 1) build an inventory of chunks—a chunkatory—and use it to segment out phrases, and 2) use the chunks to reproduce child utterances. We hypothesized that both problems could, to a large extent, be solved by attending to a single statistic: transitional probability (TP).

TP has been proposed as a cue to phrase structure in the statistical learning literature: peaks in TP can be used to group words together, whereas dips in TP can be used to find phrase boundaries (e.g., Thompson & Newport, 2007). The view put forth in such studies is that TP is useful for discovering phrase structure when computed over form classes rather than words themselves. We hypothesized, instead, that distributional information tied to individual words provides richer cues to syntactic structure than has been assumed previously. Because we adopted this item-based approach, we decided to examine backward transitional probability (BTP) as well as forward transitional probability (FTP). If learners compute statistics over individual words rather than form classes, the FTP between the words in phrases like *the cat* will always be low, given the sheer number of nouns that may follow any given determiner. BTPs provide a way around this issue: given the word *cat*, the probability that the determiner *the* immediately precedes it is quite high.

Corpora

Thirteen corpora were selected from the CHILDES database (MacWhinney, 2000) to cover a typologically diverse set of languages, representing 12 genera from 8 different language families (Haspelmath, Dryer, Gil, & Comrie, 2005). For each language, the largest available corpus for a single child was chosen rather than aggregating data across multiple child corpora, in order to assess what could be learned from the input available to individual children. All corpora involved interactions between a child and one or more adults. The average age of the target child at the beginnings of the corpora was 1;8, and 3;6 at the ends. The average number of words in each corpus was 168,204.

Table 1: Natural Language Corpora

Language	Genus	Family	Word Ord.
English	Germanic	Indo-European	SVO
German	Germanic	Indo-European	n.d.
French	Romance	Indo-European	SVO
Irish	Celtic	Indo-European	VSO
Croatian	Slavic	Indo-European	SVO
Estonian	Finnic	Uralic	SVO
Hungarian	Ugric	Uralic	n.d.
Hebrew	Semitic	Afro-Asiatic	SVO
Sesotho	Bantoid	Niger-Congo	SVO
Tamil	Dravidian	Dravidian	SOV
Indonesian	Sundic	Austronesian	SVO
Cantonese	Chinese	Sino-Tibetan	SVO
Japanese	Japanese	Japanese	SOV

The selected languages differed syntactically in a number of ways (see Table 1). Four word orders were represented: SVO, VSO, SOV, and no dominant order (n.d.; Haspelmath et al., 2005). The languages varied widely in morphological complexity, falling across the isolating/synthetic spectrum: while some languages had a relatively low morpheme-to-word ratio (e.g., Cantonese), others had a much higher ratio (e.g., Hungarian), and others had ratios falling between the two (e.g., Sesotho; Chang, Lieven, & Tomasello, 2008).

Corpus Preparation Each corpus was submitted to the same automated procedure whereby punctuation (including apostrophes: e.g., *it's→its*), codes, and tags were removed, leaving only speaker identifiers and the original sequence of words. Hash tags (#) were added to the beginning of each line to signal the start of the utterance.

Comprehension Task

Child language comprehension was approximated in terms of the model's ability to segment the corpus into phrasal units. The model's performance was evaluated against a shallow parser, a tool (widely used in the field of natural language processing) which identifies and segments out non-embedded phrases in a text. The shallow parsing method was chosen because it is consistent with the relatively underspecified nature of human sentence comprehension (Sanford & Sturt, 2002) and provides a reasonable approximation of the item-based way in which children process sentences (cf. Tomasello, 2003).

For reasons explained above, we focused on BTP as a cue to phrasal units. The model discovered chunks by tracking the peaks and dips in BTP between words, using high BTPs to group words into phrases and low BTPs to identify phrase boundaries. Chunks learned in this way were then used to help process and learn from subsequent input. We tested the

model on the corpora for which an automated scoring method was available: English, German, and French.

Model The model discovered its first chunks through simple sequential statistics. Processing utterances on a word-by-word basis, the model learned frequency information for words and word-pairs, which was used on-line to track the BTP between words and maintain a running average BTP for previously encountered pairs. When the model calculated a BTP that was greater than expected, based on the running average, it grouped the word-pair together such that it would form part (or all) of a chunk; when the calculated BTP met or fell below the running average, a boundary was placed and the chunk thereby created (consisting of one or more words to the left) was added to the chunkatory.

Once the model discovered its first chunk, it began using its chunkatory to assist in processing the input on the same word-to-word basis as before. The model continued learning the same low-level distributional information and calculating BTPs, but also used the chunkatory to make on-line predictions as to which words would form a chunk, based on previously learned chunks. When a word-pair was encountered, it was checked against the chunkatory; if it had occurred at least twice as a complete chunk or as part of a larger chunk, the words were grouped together and the model moved on to the next word. If the word-pair was not represented strongly enough in the chunkatory, the BTP was compared to the running average, with the same consequences as before. Thus, there were no *a priori* limits on the number or size of chunks that could be learned.

As an example, consider the following scenario in which the model encounters the phrase *the blue doll* for the first time and its chunkatory includes *the blue car* and *blue doll* (with counts greater than 2). When processing *the* and *blue,* the model will not place a boundary between these two words because the word-pair is already strongly represented in the chunkatory (as in *the blue car*). The model therefore predicts that this bigram will form part of a chunk. Next, when processing *blue* and *doll*, the model reacts similarly, as this bigram is also represented in the chunkatory. The model thereby combines its knowledge of two chunks to discover a new, third chunk, *the blue doll*, which is added to the chunkatory. As a consequence, the (sub)chunk, *the blue*, becomes even more strongly represented in the chunkatory, as there are now two chunks in which it appears.

Scoring The model was scored against shallow parsers: the Illinois Chunker (Punyakanok & Roth, 2001) was used for English, and TreeTagger (Schmid, 1994) was used for French and German. After shallow parsing the corpora, phrase labels (VP, NP, etc.) were removed and replaced with boundary markers of the sort produced by the model.

Each boundary marker placed by the model was scored as a *hit* if it corresponded to a boundary marker created by the shallow parser, and as a *false alarm* otherwise. Each boundary placed by the shallow parser but which was not placed by the model was scored as a *miss*. Thus, accuracy was calculated by *hits / (hits + false alarms)*, and completeness by *hits / (hits + misses)*.

Alternate Distributional Models As previous work in the statistical learning literature has focused on FTP as a cue to phrase structure (e.g., Thompson & Newport, 2007), an alternate model was created to compare the usefulness of this cue against the BTPs used by CAPPUCCINO. This model was identical to the original model, but used FTPs in place of the BTPs. We refer to this as the FTP-chunk model. To assess the usefulness of variable-sized chunks, two additional alternate models were created which lacked chunkatories, relying instead on either FTPs or BTPs computed over stored trigrams (in the case of the former, if the FTP between the first bigram and the final unigram of a trigram fell below the average, a boundary was inserted). We refer to these models as the FTP-3G and BTP-3G alternates, respectively.

Word Class Corpora A great deal of work in computational linguistics has assumed that statistics computed over form classes are superior to word-based approaches for learning about syntax (hence the widespread use of tagged corpora). This assumption is also present throughout the statistical learning literature (e.g., Thompson & Newport, 2007; Saffran, 2002), but is at odds with the present model, which relies on statistics computed over individual words rather than classes. To evaluate the usefulness of word-based transitional probabilities against those calculated over word classes, we ran the model and alternates on separate versions of each of the three corpora, in which words were replaced by the names of their lexical categories. For English, this process was automatically carried out using the tags in the original corpus. The untagged French and German corpora were tagged using TreeTagger (Schmid, 1994) before undergoing the same process. Across all three corpora, the same 13 categories were used (noun, verb, adjective, numeral, adverb, determiner, pronoun, preposition, conjunction, interjection, abbreviation, infinitive marker, and proper name). Unknown words (e.g., transcribed babbling) were marked as such.

Results and Discussion The results are displayed in Figure 1. Chi-square tests were performed separately for accuracy and completeness on each language/model pair, contrasting BTP vs. FTP, chunks vs. 3G, and words vs. classes. All differences observable in the graph were highly significant ($p<.001$), with the exceptions of non-significant differences in accuracy when using words vs. classes for the FTP-chunk model (German) and both 3G models (English).

When exposed to words, CAPPUCCINO offered the best combination of accuracy and completeness; for each language, it scored highest on both measures, with the exception of a 1%-point accuracy difference from the French class-based FTP-3G alternate (note, however, that CAPPUCCINO had a better completeness score by 23%-points). The 3G alternates displayed higher accuracy when exposed to classes rather than words (with the exceptions of FTP-3G for English and BTP-3G for German), but lower completeness; in all cases, completeness was far lower for

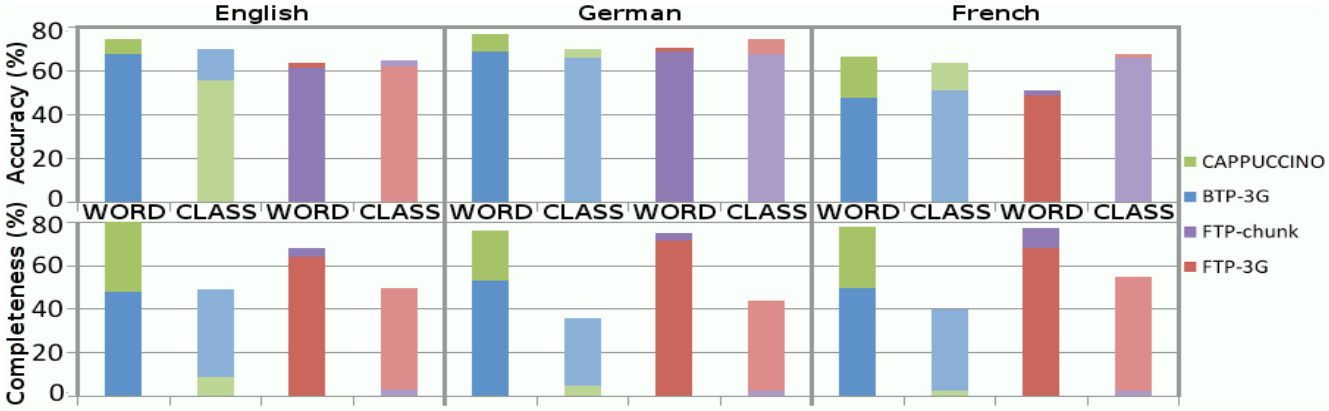

Fig. 1: Accuracy and completeness scores for CAPPUCCINO and the FTP-chunk/3G alternates

class-based models (with the exception of the 3G-BTP alternate for English). More generally, the best performance was achieved using BTP-based chunks for all three languages, despite syntactic differences between them; though two of the languages were SVO, the third (German) had no dominant word order.

Thus, CAPPUCCINO was able to approximate the performance of a shallow parser by learning in an on-line, incremental fashion from a single distributional cue. This was only the case, however, when the model was exposed to individual words rather than lexical categories. In addition to highlighting the wealth of distributional information available in the input, these results suggest that item-specific knowledge of phrase structure may be more useful to early learners than abstract knowledge.

Sentence Production Task

The production task was inspired by the bag-of-words incremental generation task used by Chang et al. (2008), which offers a method for automatically evaluating syntactic learners on any language corpus. In our task, the model made its way through the corpus incrementally, collecting statistics and discovering chunks in the service of comprehension (as described above). Each time the model encountered a multi-word child utterance, however, it was required to recreate the utterance using only chunk information discovered in the previously encountered input.

Model We began with the assumption that the overall message which the child wanted to convey could be approximated by treating the utterance as a randomly-ordered set of words: a 'bag-of-words.' The task for the model, then, was to place these words in the correct order (as originally produced by the child). Following usage-based approaches, the model utilized its chunkatory to reconstruct the child's utterances. In order to model retrieval of stored chunks during production, the bag-of-words was filled by comparing parts of the child's utterance against the chunkatory. E.g., consider a scenario in which the model encounters the child utterance *the dog chased a cat* and the largest chunk in the chunkatory consists of 3 words. To begin, the first 3 words are checked for storage as a single chunk. As this is not found in the chunkatory, *the dog* is checked. This check succeeds, so the words are removed from the utterance and placed in the bag as a single chunk. Next, *chased a cat* is checked, unsuccessfully, followed by *chased a*, also without success. The word *chased* is placed in the bag. Then *a cat* is checked, and so on. Crucially, however, this procedure was only used to find chunks that the model already knew (i.e., that were in the chunkatory) and would be likely to use as such (e.g., *the dog*). Once in the bag, the order of chunks was randomized.

During production, the model had to reproduce the child's utterance using the unordered chunks in the bag. We modeled this as an incremental, chunk-to-chunk process rather than one of whole-sentence optimization. Thus, the model began by removing from the bag the chunk with the highest BTP given the # tag (which marked the beginning of each utterance in the corpus), and producing it as the start of its new utterance. The chunk was removed from the bag before the model selected and produced its next chunk, the one with the highest BTP given the most recently produced chunk. In this manner, the model used chunk-to-chunk BTPs to incrementally produce the utterance, adding chunks one-by-one until the bag was empty. In rare cases where two or more units in the bag-of-words were tied for the highest BTP, one of them was chosen at random.

Scoring Method For each utterance the model produced correctly, it received a score of 1; if the utterance did not match the corresponding child utterance completely, a score of 0 was assigned. The overall percentage of correctly produced utterances was then used as a measure of sentence production performance for a given corpus.

Alternate Models The alternate FTP-chunk and 3G models used in the comprehension task were again used as baselines. The FTP-chunk model performed production in an identical manner to CAPPUCCINO (but used FTPs). As the 3G alternates lacked chunk inventories, they relied on TPs between unigrams and the start-of-utterance marker to select the first word in an utterance before using TPs based on trigram statistics for every subsequent word.

Results and Discussion The average sentence production score for all 13 corpora (see Figure 2) was 59.8% for CAPPUCCINO, compared to 52.3%, 49.6%, and 54.2% for the FTP-chunk, FTP-3G, and BTP-3G alternates, respectively. The model scored higher than the alternates on all corpora. A 2 (Unit Type: Chunk vs. 3G) x 2 (Direction:

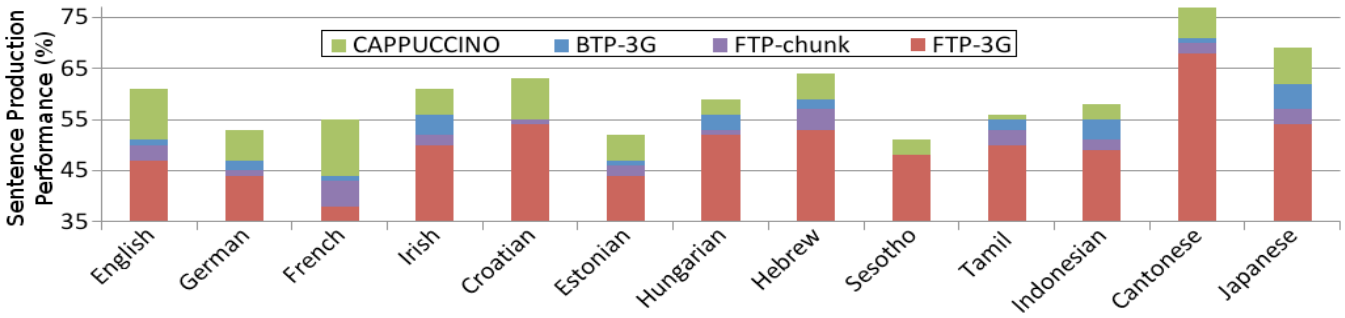

Fig. 2: Sentence production performance for CAPPUCCINO and the FTP-chunk/3G alternates

FTP vs. BTP) ANOVA confirmed the observable model differences in Figure 2, yielding main effects of Unit Type ($F(1,12)=26.2$, $p<.001$) and Direction ($F(1,12)=14.9$, $p<.01$), and a Unit Type x Direction interaction ($F(1,12)=24.6$, $p<.001$).

These results offer substantial, cross-linguistic support for CAPPUCCINO, and, more broadly, for the view that simple distributional statistics can capture a considerable part of early linguistic behavior. A single distributional cue, BTP, was used by the model to discover chunks as well as combine them to construct over half of the child utterances in each corpus (in one case, over 75%). Importantly, this approach was effective across the isolating/synthetic spectrum, yielding high performance despite the morphological complexity of the language learned by the model. This was true for all four word orders represented.

The results also serve to demonstrate that a single source of information is sometimes useful for learning about structure at multiple levels: the same distributional statistic (BTP) can be used to segment words when calculated over syllables (e.g., Pelucchi et al., 2009), to discover phrase structure when calculated over words (as in the comprehension task), and to construct utterances when calculated over multi-word chunks.

Simulation 2: Modeling Child AGL

Artificial grammar learning (AGL) studies provide a means to study learning from language-like stimuli in a controlled setting. As such, they provide a rich source of psycholinguistic data for constraining computational accounts of learning from distributional information. We therefore tested CAPPUCCINO's ability to model data from a child AGL experiment (Saffran, 2002).

This particular study was chosen because it demonstrated the use of predictive dependencies on the part of the learner to group words into phrases. Subjects (aged 7;6 to 9;8) were trained on nonsense sentences generated by one of two artificial grammars. Each grammar consisted of a set of rewrite rules used to generate an artificial language: one incorporated predictive dependencies between words within phrases (Language P), while the other lacked this cue (Language N). When tested on grammatical/ungrammatical item pairs, children exposed to Language P outperformed those exposed to Language N.

Method

The model was identical to that used with natural languages. Importantly, this meant that the model continued learning during exposure to test items. For each language, 15 simulations were performed, corresponding to the 30 child subjects from Saffran (2002). The model received the same amount of exposure to the exact same stimuli as the human subjects did (for each language, 50 sentences repeated 8 times for a total of 400 training items, followed by 24 test item pairs). Each test item pair consisted of one sentence that was grammatical, and one that was ungrammatical. Saffran's languages were created such that the same set of test items could be used for both language exposure conditions. We hypothesized that the human responses in this study were primarily based on sensitivity to the phrase-like structure of the test stimuli. The model was therefore evaluated against a version of the test items that contained the correct phrase boundaries, as defined by the rewrite rules used to generate the sentences in Saffran's study. Boundaries were placed between phrases in a non-embedded fashion that emulated the shallow parsing technique used to evaluate the model's performance on natural languages in Simulation 1.

To further contrast the usefulness of item-specific vs. class-based distributional information, a separate set of simulations was performed after each word had been replaced by the corresponding class symbol from the rewrite rules in the original Saffran (2002) study.

Scoring To model the two-alternative forced choice (2AFC) task from Saffran (2002), each item in a given test pair was scored according to the number of correctly placed phrase boundaries. The item with the highest score was then selected as the model's response. If the model produced the same number of hits (including zero) for both items, a choice was made at random, allowing individual differences to appear across simulations.

Results and Discussion

The child subjects in Saffran (2002) had overall correct response rates of 71.8% for Language P and 58.3% for Language N. The model provided a close quantitative fit, with overall correct response rates of 70.5% for Language P and 57.5% for Language N. When the model was given information on word classes instead of concrete words,

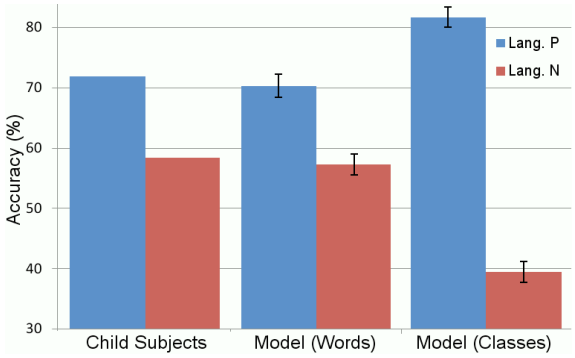

Fig. 3: 2AFC task accuracy (%) for subjects and CAPPUCCINO

however, it provided a poor fit to the child data, with 81.1% accuracy for Language P and 39.7% for Language N.

Thus, the model provides a close fit to these psycholinguistic data, suggesting that the ability to group words into larger units can indeed account for subject performance in the original study. While predictive dependencies between word classes were a potentially useful cue, the calculation of statistics over classes in the present model could not account for subject performance as well as the word-based approach. This resonates with the superior performance of the model on natural languages when working with words as opposed to lexical categories.

General Discussion

Our CAPPUCCINO model has demonstrated that incremental learning and on-line processing based on a single distributional cue, BTP, can capture a considerable part of children's early linguistic behavior. In addition to approximating the performance of a shallow parser with high accuracy and completeness, the model was able to reproduce the majority of the child utterances in each of a typologically diverse set of 13 corpora, and closely fit AGL data from child subjects. In line with usage-based approaches to language (e.g., Tomasello, 2003), the model's superior comprehension performance and ability to fit child AGL data when exposed to words as opposed to lexical categories suggests that knowledge of concrete words and chunks may be more important to early language acquisition than abstract rules operating over word classes. Of course, there is more to comprehension than shallow parsing—e.g., meaning is not taken into account—but it is encouraging to see just how well the model can reconstruct children's utterances based on distributional information alone.

As an initial step towards a chunk-based account of children's comprehension and production of language, CAPPUCCINO is not without limitations. Firstly, although it closely fits child data from an artificial language learning study, it is important to determine whether our model can also account for specific patterns of natural language acquisition (e.g., similar to MOSAIC's match with cross-linguistic data regarding the optional infinitive stage; Freudenthal et al., 2009). Secondly, our model learns from already segmented speech and does not address the ways in which word segmentation may impact on chunk discovery; a child may discover its earliest chunks before segmentation of the component words has taken place. Finally, the unitization account offered by the model is oversimplified; psycholinguistic work suggests that there is no frequency 'threshold' beyond which collocations are unitized, but instead that the processing advantage for chunks increases as a function of frequency (Arnon & Snider, 2010). In future work, we will thus aim to extend CAPPUCCINO by exposing it to unsegmented corpora, by making its chunk processing more graded, and by applying it to specific patterns of language acquisition.

Acknowledgments

Thanks to Jen Misyak and Rick Dale for helpful comments.

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62, 67-82.

Bannard, C., Lieven, E.V., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences, 106*, 17284-17289.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, 19, 241-248.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research*, 9, 198-213.

Freudenthal, D. Pine, J.M. & Gobet, F. (2009). Simulating the referential properties of Dutch, German, and English Root Infinitives in MOSAIC. *Language Learning and Development, 5,* 1-29.

Haspelmath, M., Dryer, M.S., Gil, D., & Comrie, B. (2005). *The world atlas of language structures*. Oxford: OUP

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Vol. II: The database*. Mahwah, NJ: LEA.

Monaghan, P. & Christiansen, M.H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Trends in corpus research: Finding structure in data* (pp. 139-163). Amsterdam: J. Benjamins.

Pelucchi, B., Hay, J.F., & Saffran, J.R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition,* 113, 244-247.

Pinker, S. (1999). *Words and rules.* New York: Basic Books.

Punyakanok, V., & Roth, D. (2001). The use of classifiers in sequential inference. In Dietterich, G., Becker, S., & Ghahramani, Z. (Eds.), *Proceedings of the Conference on Advances in Neural Information Processing Systems* (pp. 995-1001). Cambridge, MA: MIT Press

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, 47, 172-196.

Sanford, A.J. & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences, 6,* 382-386.

Schmid, H. (1995, March). *Improvements in part-of-speech tagging with an application to German.* Paper presented at the EACL-SIGDAT Workshop, Dublin, Ireland.

Thompson, S.P., & Newport, E.L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, 3, 1-42.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: HUP.