Author's personal copy

Available online at www.sciencedirect.com
ScienceDirect
Cognitive Systems Research 9 (2008) 198–213
Cognitive Systems RESEARCH
www.elsevier.com/locate/cogsys

Automatic evaluation of syntactic learners in typologically-different languages

Action editor: Gregg Oden

Franklin Chang [a,*], Elena Lieven [b], Michael Tomasello [b]

[a] *Cognitive Language Information Processing Open Laboratory, NTT Communication Sciences Laboratories, NTT Corp., 2-4 Hikari-dai, Seika-cho, Souraku-gun, 6190237 Kyoto, Japan*

[b] *Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany*

Received 7 June 2007; received in revised form 12 September 2007; accepted 6 October 2007
Available online 1 November 2007

Abstract

Human syntax acquisition involves a system that can learn constraints on possible word sequences in typologically-different human languages. Evaluation of computational syntax acquisition systems typically involves theory-specific or language-specific assumptions that make it difficult to compare results in multiple languages. To address this problem, a bag-of-words incremental generation (BIG) task with an automatic sentence prediction accuracy (SPA) evaluation measure was developed. The BIG–SPA task was used to test several learners that incorporated *n*-gram statistics which are commonly found in statistical approaches to syntax acquisition. In addition, a novel Adjacency–Prominence learner, that was based on psycholinguistic work in sentence production and syntax acquisition, was also tested and it was found that this learner yielded the best results in this task on these languages. In general, the BIG–SPA task is argued to be a useful platform for comparing explicit theories of syntax acquisition in multiple languages.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Syntax acquisition; Computational linguistics; Corpora; Syntax evaluation; Linguistic typology

1. Introduction

Children, computers, and linguists have similar challenges in extracting syntactic constraints from language input. Any system that acquires syntactic knowledge (a syntactic learner) must confront the fact that words do not come labeled with syntactic categories and the syntactic relations that can hold among these words can vary to a great extent among languages. This article presents a method for evaluating syntactic learners, that is, how well they have acquired syntactic knowledge from the input. This method, which uses a *bag-of-words incremental generation* (BIG) task and an evaluation measure called *sentence prediction accuracy* (SPA), is applied to several formally-specified learners, as well as to a new learner called the Adjacency–Prominence learner. It will be shown that the SPA measure is capable of evaluating the syntactic abilities in a variety of learners using input from typologically-different languages and it does so in a manner that is relatively free of assumptions about the form of linguistic knowledge.

Words in utterances are not labeled with syntactic categories, and there is variability in how linguistic theories characterize the syntactic constraints on an utterance. For example, constructions are a type of syntactic unit in some theories (Goldberg, 1995), but not others (Chomsky, 1981). Syntactic constraints also differ across languages, and it is difficult to adapt a particular theory of syntactic categories or constraints to typologically-different languages (Croft, 2001). For example, the adjective category is often thought to be a universal syntactic category, but in many languages, it is difficult to distinguish adjectives and stative verbs (e.g., Chinese, Li & Thompson, 1990)

* Corresponding author. Tel.: +81 774 93 5273; fax: +81 774 93 5345.
E-mail address: chang.franklin@gmail.com (F. Chang).

1389-0417/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cogsys.2007.10.002

Author's personal copy

and in some languages, there are several adjective categories (e.g., Japanese, Tsujimura, 1996). Since the labeling of corpora requires that one make particular assumptions about the nature of syntax, the evaluation of syntactic knowledge with these human-labeled corpora is both theory- and language-dependent. These evaluation methods work best for mature areas of syntactic theory, such as the evaluation of adult English syntactic knowledge, but are less suited for areas such as syntax acquisition or linguistic typology, where there is more controversy about the nature of syntax (Croft, 2001; Pinker, 1989; Tomasello, 2003).

A large number of computational approaches for learning syntactic knowledge are evaluated against human-labeled corpora. For example in part-of-speech tagging, a tagger attempts to predict the syntactic category (or tag) for each of the words in an utterance, and the system is evaluated by comparing its output against the human-labeled tag sequence associated with the test utterance (Church, 1989; Dermatas & Kokkinakis, 1995). The set of tag categories that are used to label a particular corpus is called its tagset, and different corpora, even in the same language, use different tagsets (Jurafsky & Martin, 2000). In addition, the same tagger can show different levels of performance, when evaluated against different types of corpora or different tagsets. Atwell et al. (2000) trained a supervised tagger with a single corpus that had been tagged with eight different English tagsets and found significant variation among the tagsets in test accuracy from 86.4% to 94.3%. When taggers are applied to multiple languages, there is an additional problem that the tagsets are not equated across the languages, because tagsets can vary in the specificity of the categories or in the degree that semantic or formal criteria are used for assignment of categories (Croft, 2001). For example, Dermatas and Kokkinakis (1995) found that the same Hidden Markov Model for part-of-speech tagging (HMM-TS2) with the same amount of input (50,000 words) labeled with the same set of categories (extended grammatical classes) yielded better accuracy levels for English (around 5% prediction error, EEC-law text) than for five other European languages (Greek yielded more than 20% prediction error). Since many of the relevant factors were controlled here (e.g., input size, learner, categories), the large variability in accuracy is probably due to the match between the categories and the utterances in the corpora, in this case, the match was better for English than Greek. If that is the case, it suggests that evaluating these systems with this tagset is inherently biased towards English. Other evaluation measures in computational linguistics, such as the learning of dependency structures, also seem to be biased toward English. Klein and Manning (2004) found that their unsupervised dependency model with valence plus constituent-context learner yielded accuracy results in English of 77.6% (Fig. 6 in their paper, UF_1), but German was 13.7% lower and Chinese was 34.3% lower. In addition to these biases, English corpora are often larger and more consistently labeled and together these factors help to insure that there will be a bias towards English in evaluation of computational systems. But since humans can learn any human language equally well, it is desirable to have a way to evaluate syntax that is not inherently biased for particular languages.

One area of computational linguistics that has been forced to deal with variability in syntax across languages is the domain of machine translation. In translating an utterance from a source language to a target language, these systems attempt to satisfy two constraints. One constraint is to ensure that the meaning of the source utterance is preserved in the target utterance and the other constraint is that the order of words in the target utterance should respect the syntactic constraints of the target language. In statistical approaches to machine translation, these constraints are supported by two components: the translation model and the language model (Brown, Della Pietra, Della Pietra, & Mercer, 1993). The translation model assumes that the words in the source utterance capture some of its meaning, and this meaning can be transferred to the target utterance by translating the words in the source language into the target language. Since words in some languages do not have correspondences in other languages, the set of translated words can be augmented with additional words or words can be removed from the set. This set of translated words will be referred to as a bag-of-words, since the order of the words may not be appropriate for the target language. The ordering of the bag-of-words for the syntax of the target language is called decoding, and involves the statistics in the language model. Statistical machine translation systems are not able to match human generated translations, but they are able to generate translations of fairly long and complicated utterances and these utterances can be often understood by native speakers of the target language.

In statistical machine translation, the ordering of the words in an utterance is a whole utterance optimization process, where the goal is to optimize a particular metric (e.g., the transition probabilities between words) over the whole utterance. This optimization is computationally intensive, since finding an optimal path through a set of words is equivalent to the Traveling Salesman problem and therefore is NP-complete (Knight, 1999). There is however no guarantee that humans are doing whole sentence optimization of the sort that is used in statistical machine translation. And there is experimental evidence from humans that contradicts the assumptions of whole sentence optimization and suggests instead that speakers can plan utterances incrementally. Incremental planning means that speakers plan sentences word-by-word using various scopes of syntactic and message information. Incremental planning during production predicts that words that are more accessible due to lexical, semantic, or discourse factors will tend to come earlier in utterances and there is a large amount of experimental evidence supporting this (Bock, 1982, 1986; Bock & Irwin, 1980; Bock & Warren, 1985; Bock, Loebell, & Morey, 1992; Ferreira & Yoshita,

Author's personal copy

2003; Prat-Sala & Branigan, 2000). Notice that in whole sentence planning, accessible words can be placed anywhere in the sentence and therefore there is no explanation for why they tend to go earlier in sentences. In addition to this work on accessibility, the time spent planning a sentence is not consistent with whole sentence optimization. In statistical machine translation systems that do whole sentence optimization, the time it takes to plan and initiate an utterance depends on utterance length (Germann, Jahr, Knight, Marcu, & Yamada, 2004), but in humans, sentence initiation times can be equivalent for different length sentences and this suggests that humans are only planning part of the utterance (Ferreira, 1996; Smith & Wheeldon, 1999). And furthermore, human utterances are not globally optimal in terms of transition statistics. Humans sometimes produce non-canonical structures such as heavy NP shifted structures (e.g., "Mary gave to the man the book that she bought last week"; Hawkins, 1994; Stallings, MacDonald, & O'Seaghdha, 1998; Yamashita & Chang, 2001) that violate the local transition statistics of the language ("gave to" is less frequent than "gave the"). Therefore while whole sentence optimization is an appropriate computational approach to solving the utterance-ordering problem, it may not be the most appropriate way to model the process that people use to generate utterances. Since our goal is to have an evaluation measure of syntax acquisition and use that is compatible with experimental work on syntax acquisition and sentence production, our evaluation task will be designed to accommodate incremental or greedy approaches to sentence generation.

We propose that systems that learn syntactic constraints can be evaluated using a bag-of-words generation task that is akin to a monolingual incremental version of the task use in statistical machine translation. In our task, we take the target utterance that we want to generate, and place the words from that utterance into an unordered bag-of-words. We assume that speakers have a meaning or message that they want to convey (Levelt, 1989) and the bag-of-words is a practical way of approximating the message constraints for utterances in typologically-different languages. The syntactic learner must use its syntactic knowledge to order this bag-of-words. The generation of the sentence is incremental, where the learner tries to predict the utterance one word at a time. As the sentence is produced, the target word is removed from the bag-of-words. This means that a learner can use statistics based on the changing set of words in the bag-of-words as well as information from the previous words to help in the prediction process. By reducing the bag-of-words as the sentence is produced, this task breaks down sentence generation into a recursive selection of the first word from the gradually diminishing bag-of-words, and this makes the task more incremental than standard bag-generation approaches, and hence we will refer to this approach as the bag-of-words incremental generation (BIG) task.

To evaluate our syntactic learners, we will have them produce utterances in our corpora, and then see whether the learner can correctly predict the original order of all of the words in each of the utterances. If we average over all of the utterances in a corpus, then the percentage of complete utterances correctly produced is the Sentence Prediction Accuracy (SPA). The SPA evaluation measure differs in several respects from the evaluation measures used for language models and statistical machine translation. Evaluation of language models often uses word-based accuracy measures, often filtered through information-theoretic concepts like perplexity and entropy (Jurafsky & Martin, 2000; Chapter 6). Since the grammaticality of a sentence depends on the order of all of the words in the utterance, a word-based accuracy measure is not a suitable way to measure syntactic knowledge. For example, if a system predicted the word order for a set of 10-word utterances and it reversed the position of two words in each utterance, then its word accuracy would be 80%, even though it is possible that all of the utterances produced were ungrammatical (its SPA accuracy would be zero).

The SPA measure is similar to evaluation measures in statistical machine translation such as Bleu (Papineni, Roukos, Ward, & Zhu, 2001). The Bleu metric captures the similarity in various *n*-grams between a generated utterance and several human reference translations. Since Bleu is a graded measure of similarity, it does not make a strict distinction between a sentence that is an exact match, and therefore guaranteed to be grammatical, and a partial match, which could be ungrammatical. Even with this limitation, Bleu has transformed the field of statistical machine translation by reducing the need for laborious and expensive human evaluation of machine-generated translations, and thereby increasing the speed of system development and allowing objective comparison of different systems. The SPA metric is similar to word-prediction accuracy measures and Bleu in that it can be automatically computed from corpora, but it is stricter in that it makes a strong distinction between an exact sentence match and a partial match. In addition, perplexity or Bleu scores are not typically understood by non-computational linguists or psychologists, so SPA has another advantage in that it is transparent and can be compared directly to the average sentence accuracy in experiments or to the percentage of test sentences that are rated grammatical by a linguistic informant.

The SPA measure can be said to measure syntax in so far as the order of words in human utterances are governed by syntax. Word order is influenced by many factors such as structural, lexical, discourse, and semantic knowledge, and these factors are often incorporated into modern syntactic theories (e.g., Pollard & Sag, 1994). Syntactic theories use abstract categories and structures to encode the constraints that govern word order. For example in English, determiners tend to come before nouns and noun phrases tend to come after transitive verbs. In Japanese, noun phrases come before verbs and case-marking particles come after nouns. Hierarchical syntactic knowledge also has implications for word order. For example, the order

Author's personal copy

of elements within a subject phrase is the same regardless if it is in sentence initial position ("the boy that was hurt is resting") or after an auxiliary verb ("Is the boy that was hurt resting?") and this is captured in hierarchical theories by representing the ordering of the subject phrase elements in a subtree within the main clause tree that encodes the position of the auxiliary verb. These abstract structural constraints represent hypotheses about the internal representation for syntax and these hypotheses are tested by generating theory-consistent and theory-inconsistent word sequences that can be tested on linguistic informants. Critically, the word sequence is the link between the hypothesized syntactic theory and human syntactic knowledge. Using word sequences to evaluate syntactic knowledge is therefore a standard approach in the language sciences.

One goal of the BIG–SPA evaluation task is to bring together research from three domains that share related goals: developmental psycholinguistics, typological linguistics, and computational linguistics. Since each of these domains makes different assumptions, it is difficult to integrate these disparate approaches. For example, developmental psycholinguists assume that child-directed speech is necessary to understand the nature of syntactic development in children. Computational linguists do not often use small corpora of child-directed speech, because their data-driven algorithms require a large amount of input to yield high levels of accuracy. Instead, they tend to use large corpora like the Penn Treebank Wall Street Journal corpus, that includes economic or political news, or the Brown corpus, which includes utterances from computer manuals (e.g., IBM 7070 Autocoder Reference manual) and federal and state documents (e.g., the Taxing of Movable Tangible Property; Francis & Kucera, 1979). Since these types of corpora do not resemble the input that children receive, developmental psycholinguists might have good reasons to be skeptical about the relevance of computational linguistic results with these corpora for the study of language acquisition.

In addition, because child-directed corpora are smaller than the massive corpora that are used for computational linguistics, data-driven algorithms might not work as well with these corpora. Corpus size is linked to a variety of issues related to "the poverty of the stimulus", namely the claim that the input to children is too impoverished to insure the abstraction of the appropriate syntactic representations (Chomsky, 1980). While there is controversy about whether the input to children is actually impoverished (Pullum & Scholz, 2002; Reali & Christiansen, 2005), it is less controversial to say that the input corpora used by computational systems or researchers may not be sufficiently complete to allow them to find the appropriate abstractions. For example in computational linguistics, the input to computational systems does not always cover the test set (e.g., data sparseness, unknown words, Manning & Schütze, 1999). And in developmental psycholinguistics, the corpora that researchers use may not be big enough or dense enough to capture the phenomena of interest (Lieven, Behrens, Speares, & Tomasello, 2003; Tomasello & Stahl, 2004). Given the difficulty in creating large corpora for typologically-different languages, it is important to develop and test computational linguistic algorithms that can work with small corpora. Since the task of generating a sentence does not require the use of abstract theory-specific categories that are hard to learn from small corpora, the BIG–SPA task might be a more appropriate way to use small unlabeled corpora of child-directed speech for the study of syntax acquisition.

Another integration problem has to do with applying computational algorithms to study child-produced utterances. Developmental psycholinguists are interested in how to characterize the developing syntax in child utterances as these utterances move from simple, sometimes ungrammatical, utterances to grammatical adult utterances (Abbot-Smith & Behrens, 2006; Lieven et al., 2003; Pine & Lieven, 1997; Tomasello, 1992, 2003). Computational linguistic systems often make assumptions which make it difficult to use these algorithms with utterances in development. Many part-of-speech tagging systems require that the system know the syntactic tagset before learning begins and evaluation of these systems requires a tagged corpus or a dictionary of the words paired with syntactic categories (Mintz, 2003; Mintz, Newport, & Bever, 2002; Redington, Chater, & Finch, 1998). There is no consensus in how to build these tagsets, dictionaries, and tagged corpora for child utterances, because developmental psychologists disagree about the nature of the categories that children use at particular points in development. For example in early syntax development, Pinker (1984) argues that children link words to adult syntactic categories, while Tomasello (2003) argues that children initially use lexical-specific categories.

A third integration difficulty has to do with claims about the universality of syntax acquisition mechanisms. Developmental psycholinguists have proposed that distributional learning mechanisms, akin to those used in computational linguistics, might be part of the syntactic category induction mechanism in humans (Mintz, 2003; Redington et al., 1998). But since these proposals were only tested in English (and a few other languages; e.g., Chemla, Mintz, Bernal, & Christophe, in press; Redington et al., 1995), we do not know about the relative efficacy of these methods in languages with different language typologies. To make claims about the universal character of syntax acquisition, a mechanism must be tested on a wide number of typologically-different languages. But the problem is that standard evaluation measures, such as those used by the above researchers, require language-dependent tagsets and this is a problem when comparing across languages. For example, Czech corpora sometimes have more than 1000 tags (Hajič & Vidová-Hladká, 1997) and tagging this type of language would be a challenge for algorithms that are designed or tuned for smaller tagsets. Another issue is that linguists in different languages label corpora differently and this creates variability in the evaluation measures used.

Author's personal copy

For example, it has been found that words in Chinese corpora have more part-of-speech labels per word than words in English or German corpora, and this difference can contribute to the difficulty in part-of-speech tagging (Tseng, Jurafsky, & Manning, 2005). Since SPA does not use syntactic categories for evaluation, it is less sensitive to differences in the way that linguists label different languages.

In this paper, we will use the SPA measure with the BIG task to evaluate several algorithms of the sort that have been proposed in computational linguistics and developmental psycholinguistics. We used corpora of adult–child interactions, which include utterances that children typically use to learn their native language, from 12 typologically-different languages, which is large enough to allow some generalization to the full space of human languages. What follows is divided into three sections. First, the corpora that were used will be described (*Typologically-Different Corpora*). Then several *n*-gram-based learners will be compared and evaluated with BIG–SPA (*BIG–SPA evaluation of n-gram-based learners*). Then a new psycholinguistically-motivated learner (*Adjacency–Prominence learner*) will be presented and compared with several simpler learners (*BIG–SPA evaluation of Adjacency–Prominence-type learners*).

2. Typologically-different corpora

To have a typologically-diverse set of corpora for testing, we selected 12 corpora from the CHILDES database (MacWhinney, 2000): Cantonese, Croatian, English, Estonian, French, German, Hebrew, Hungarian, Japanese, Sesotho, Tamil, Welsh. In addition, two larger English and German-Dense corpora from the Max Planck Institute for Evolutionary Anthropology were also used (Abbot-Smith & Behrens, 2006; Brandt, Diessel, & Tomasello, in press; Maslen, Theakston, Lieven, & Tomasello, 2004). These languages differed syntactically in important ways. German, Japanese, Croatian, Hungarian, and Tamil have more freedom in the placement of noun phrases (although the order is influenced by discourse factors) than English, French, and Cantonese (Comrie, 1987). Several allowed arguments to be omitted (e.g., Japanese, Cantonese). Several had rich morphological processes that can result in complex word forms (e.g., Croatian, Estonian, Hungarian, see "Number of Cases" in Haspelmath, Dryer, Gil, & Comrie, 2005). Four common word orders were represented (e.g., SVO English; SOV Japanese; VSO Welsh; No dominant order, Hungarian; Haspelmath et al., 2005). Seven language families were represented (Indo-European, Uralic, Afro-Asiatic, Dravidian, Sino-Tibetan, Japanese, Niger-Congo; Haspelmath et al., 2005). Eleven genera were represented (Chinese, Germanic, Finnic, Romance, Semitic, Ugric, Japanese, Slavic, Bantoid, Southern Dravidian, Celtic; Haspelmath et al., 2005). All the corpora involved interactions between a target child and at least one adult that were collected from multiple recordings over several months or years (see appendix for details). For each corpus, the *child* utterances were the target child utterances for that corpus, and the *adult* utterances were all other utterances. Extra codes were removed from the utterances to yield the original segmented sequence of words. The punctuation symbols (period, question mark, exclamation point) were moved to the front of the utterances and treated as separate words. This was done because within the BIG task, we assumed that speakers have a message that they want to convey and therefore they know whether they were going to make a statement, a question, or an exclamation, and this knowledge could help them to generate their utterance. If an utterance had repeated words, each of the repeated words was given a number tag to make it unique (e.g., you-1, you-2), since in a speaker's message, the meaning of these repeated words would have to be distinctly represented. These tags were placed on words starting from the last word, but with the last word unmarked. For example, the utterance "normally when you press those you get a nice tune, don't you?" would be "? normally when you-2 press those you-1 get a nice tune don't you" for learning and testing (example utterances in this paper come from either the English or English-Dense corpora). Using this system for marking repeated words allowed learners to learn reliable statistics between the different forms of the same word (e.g., "you-2" tends to come before "you") and they might even be able to capture different statistical regularities for each word. For example, since "when" signals an embedded clause, it might be followed by "you-2" more than "you". These words were kept distinct in the statistics and during generation of utterances at test, but for calculation of SPA, any form of the word was treated as correct (e.g., "you-1" or "you-2" were equivalent to "you"). This method of marking repeated words is the most appropriate method for the BIG–SPA task, because of its use of recursive prediction on a gradually diminishing bag-of-words.

3. BIG–SPA evaluation of *n*-gram learners

To show that an evaluation measure is a useful tool for comparing syntactic learners, one needs to have a set of learners that can be compared. Since *n*-gram statistics, which use the frequency of sequences of n adjacent words, are popular in both developmental psycholinguistics (Thompson & Newport, 2007) and in computational approaches to syntax acquisition (Reali & Christiansen, 2005), we compared several learners that use these types of statistics. The simplest learners were a *Bigram* (two adjacent words) and a *Trigram* (three adjacent words) learner using maximum likelihood estimation equations (Manning & Schütze, 1999). In language modeling, it is standard to combine different *n*-grams together in a weighted manner to take advantage of the greater precision of higher *n*-gram statistics with the greater availability of lower *n*-gram statistics (this is called *smoothing*). Therefore, several smoothed *n*-gram learners were also tested: *Bigram+Trigram* learner and *Unigram+Bigram+Trigram* learner. In addition to these

Author's personal copy

learners, we created a *Backoff Trigram* learner, which tried to use trigram statistics if available, and backed-off to bigram statistics if the trigrams are not available, and finally backed-off to unigram statistics if the other two statistics were not available. Parameters were not used to weight the contribution of these different statistics in these learners, because parameters that are fitted to particular corpora make it harder to infer the contribution of each statistic over all of the corpora. In addition, we also created a Chance learner whose SPA score estimated the likelihood of getting a correct sentence by random generation of the utterance from the bag-of-words. Since an utterance with n words had $n!$ possible orders for those words, the Chance performance percentage for that utterance was $100/n!$ (notice that the average length of utterances in a corpus can be derived from the Chance learner's score). The learners differed only in terms of their Choice function, which was the probability of producing a particular word from the bag-of-words at each point in a sentence, and the Choice functions for the learners are shown below.

Definition of statistics used in learners	
$C(w_{n-k}\ldots w_n)$	Frequency of n-gram $w_{n-k}\ \ldots\ w_n$ in input set for $k = 0$, 1, or 2
NW	Number of word tokens
Equations for five different learners	
Bigram	$\text{Choice}(w_n) = C(w_{n-1}, w_n)/C(w_{n-1})$
Trigram	$\text{Choice}(w_n) = C(w_{n-2}, w_{n-1}, w_n)/C(w_{n-2},\ w_{n-1})$
Bigram+Trigram	$\text{Choice}(w_n) = C(w_{n-1}, w_n)/C(w_{n-1}) + C(w_{n-2}, w_{n-1}, w_n)/C(w_{n-2},\ w_{n-1})$
Unigram+Bigram+Trigram	$\text{Choice}(w_n) = C(w_n)/\text{NW} + C(w_{n-1}, w_n)/C(w_{n-1}) + C(w_{n-2}, w_{n-1}, w_n)/C(w_{n-2}, w_{n-1})$
Backed-off Trigram	$\text{Choice}(w_n) = C(w_{n-2}, w_{n-1}, w_n)/C(w_{n-2}, w_{n-1})$ if $C(w_{n-2}, w_{n-1}) > 0$
	$\text{Choice}(w_n) = C(w_{n-1}, w_n)/C(w_{n-1})$ if $C(w_{n-2}, w_{n-1}) == 0$ and $C(w_{n-1}) > 0$
	$\text{Choice}(w_n) = C(w_n)/n\text{words}$ if $C(w_{n-2}, w_{n-1}) == 0$ and $C(w_{n-1}) == 0$

If the denominator in the Choice equation was zero at test (i.e., unknown words), then the Choice function returned zero. Normally, the optimization of the probability of a whole sequence involves the multiplication of probabilities and this can lead to numerical underflow. Therefore in language modeling, it is standard to use a log (base 2) transformation of the probabilities and this yields an additional computational advantage for whole sentence optimization since multiplication of probabilities can be done with addition in log space. But since the BIG–SPA task does not involve computation of whole sequence probabilities, there is no computational advantage in using log-transformed probabilities. Instead, to deal with numerical underflow, all of the Choice functions were multiplied by 10^7 and computation was done with integers. We also tested versions of these learners that used log-transformed probabilities and compared to the learners that we present below, the results were similar although slightly lower, since log probabilities compress the range of values.

There were two main parts to the BIG–SPA task (see pseudocode below): collecting statistics on the input, predicting the test utterances. In the first part, statistics that were appropriate for a particular learner were collected. In the second part, the system generated a new utterance *newu* incrementally for each bag-of-words b from each utterance u in the test set. This was done by calculating the Choice function at each position in a sentence, and adding the word with the highest Choice value, the winner *win*, to the new utterance *newu*. After removing the actual next word *nw* from the bag-of-words, the same procedure was repeated until the bag-of-words was empty. If the resulting utterance was the same as the target utterance, then the SPA count was incremented. The SPA accuracy score was the SPA count divided by the number of test utterances. One-word utterances were excluded from testing since there is only one order for one-word bag-of-words. If two words in the bag-of-words had the same Choice score, then the system chose the incorrect word. This insured that the SPA accuracy was not strongly influenced by chance guessing.

```
Pseudocode for BIG–SPA task:
collection statistics from the input
For each utterance u in input set
    For each word w_n in utterance u
      Collect statistics C(w_{n-k}...w_n) for k = 0, 1, 2
## predicting the test utterances
Initialize SPA count to 0.
For each utterance u in test set
   Create bag-of-words b from utterance u
   Initialize newu to empty string
   For each word nw in u
      For each word w in b
         Calculate Choice(w)
      win = word with highest Choice value
      Add win to newu
      Remove word nw from bag-of-words b
   If u is the same as newu, then increment SPA count by 1
```

The five learners were tested in two different testing situations: Adult–Child and Adult–Adult. The Adult–Child situation matched the task that children perform when they extract knowledge from the adult input and use it in sequencing their own utterances. This task required the

Author's personal copy

ability to generalize from grammatical adult utterances (e.g., "Well, you going to tell me who you've delivered letters and parcels to this morning?") to shorter and sometimes ungrammatical child utterances (e.g., "who this?"). But since the child utterances were relatively simple, this testing situation did not provide a good measure of how well a learner would do against more complex adult utterances. Therefore, an Adult–Adult situation was also used, where 90% of the adult utterances were used for input, and 10% of the adult utterances were held out for testing (an example test sentence that was correctly produced was the 14 word utterance "do you remember when we were having a look at them in didsbury park?"). This situation showed how well the system typically worked on adult utterances when given non-overlapping adult input.

Paired *t*-tests were applied to compare the SPA accuracy for the different learners using the 14 corpora as a sample from the wider population of human languages. If a learner is statistically different from another learner over these 14 corpora, then it is likely that this difference will show up when tested on other languages that are similar to those in this sample. For example, our sample did not include Dutch utterances, but since we have several similar languages (e.g., English, German, French), a significant *t*-test over our sample would suggest that the difference between those learners would also generalize to Dutch. Fig. 1 shows the average sentence prediction accuracy over the corpora. *T*-tests were performed on the means for the different learners for each corpus, because the means equated for the differences in the size of different test sets. But since the differences in the means averaged over corpora can be small, Fig. 1 also shows the total number of correctly produced utterances for each condition to the right of each bar to emphasize that small differences in the means can still can amount to large differences in the number of utterances correctly predicted (the rank order of the total and mean percentage do not always match because of the way that correct utterances were distributed over corpora of different sizes).

The Chance learner was statistically lower than both the Bigram learner (Adult–Child $t(13) = 9.5$, $p < 0.001$; Adult–Adult, $t(13) = 10.9$, $p < 0.001$) and the Trigram learner (Adult–Child $t(13) = 8.5$, $p < 0.001$; Adult–Adult, $t(13) = 9.8$, $p < 0.001$), which suggested that the *n*-gram statistics in these learners were useful for predicting word order within the BIG task. The unsmoothed Bigram was better than the unsmoothed Trigram learner (Adult–Child $t(13) = 8.7$, $p < 0.001$; Adult–Adult, $t(13) = 6.7$, $p < 0.001$) and this was likely due to the greater overlap in bigrams between the input and test set in the small corpora that were used (e.g., the bigram "the man" was more likely to overlap than the trigram "at the man"). The combined Bigram+Trigram learner yielded an improvement over the Bigram learner (Adult–Child $t(13) = 4.4$, $p < 0.001$; Adult–Adult, $t(13) = 4.3$, $p < 0.001$) and the Trigram learner (Adult–Child $t(13) = 9.0$, $p < 0.001$; Adult–Adult, $t(13) = 10.8$, $p < 0.001$), which suggested that the trigram statistics, when available, did improve the prediction accuracy over the plain bigram and this was likely due to the greater specificity of trigrams, as they depended on more words than bigrams. Adding the unigram frequency (Unigram+Bigram+Trigram) seemed to reduce the average SPA score compared to the Bigram+Trigram, although non-significantly over the sample (Adult–Child $t(13) = 0.6$, $p = 0.56$; Adult–Adult, $t(13) = 1.5$, $p = 0.15$). Finally, we found no significant

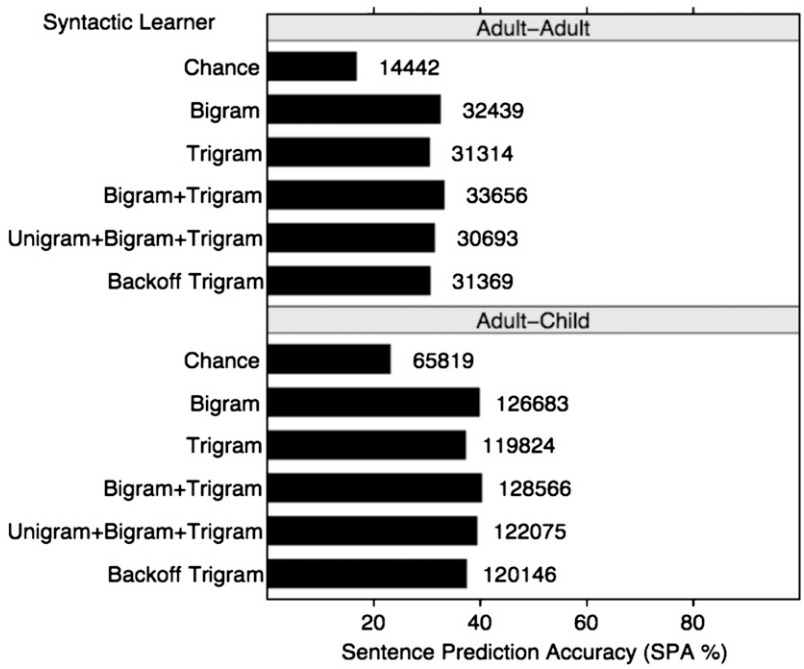

Fig. 1. Average SPA scores (%) for *n*-gram learners in Adult–Adult and Adult–Child prediction (counts of correct utterances are placed to the right of each bar).

Author's personal copy

difference between the Unigram+Bigram+Trigram learner and the Backoff Trigram learner (Adult–Child $t(13) = 1.3$, $p = 0.20$; Adult–Adult, $t(13) = 0.7$, $p = 0.48$), which suggested that these algorithms may not differ across typologically-different languages.

To understand these results, it is useful to compare them to other systems. The closest comparable results from a statistical sentence generation system are the results in the Halogen model (Langkilde-Geary, 2002). This model used *n*-gram type statistics within a whole sentence optimization sentence generation system. They were able to predict 16.5% of the English utterances in their corpora when tested under similar conditions to our learners (condition "no leaf, clause feats", where only lexical information was given to the system). This result was lower than our results with similar *n*-grams, but this is expected as their test corpora had longer utterances. But they also used most of the Penn Treebank Wall Street Journal corpus as their input corpus, so their input was several magnitudes larger than any of our corpora. Therefore compared to learners that use massive English newspaper corpora in a non-greedy sentence generation system, our *n*-gram learners yielded similar or higher levels of accuracy in utterance prediction with input from small corpora of adult–child interactions in typologically-different languages.

In addition to looking at the means averaged over corpora, it is also useful to look at the SPA results for each corpus (Adult–Adult test, Fig. 2), as long as one remembers that the differences in the corpora were not just due to language properties, but also reflected properties of the particular speakers and the particular recording situation. One interesting finding in the Adult–Adult prediction results was that the Unigram+Bigram+Trigram learner had lower results than the Bigram+Trigram learner in Cantonese, English, English-Dense, and Japanese. One possible reason that unigram frequencies might be detrimental in these languages could be due to the analytic nature of these languages (low ratio of words to morphemes). Analytic languages use separate function words to mark syntactic relationships (e.g., articles like "the" or auxiliary verbs like "is") and since these words are separated and occur at different points in a sentence, the high unigram frequency of these function words can be problematic if unigram frequency increases the likelihood of being placed earlier in sentences. Normally, Japanese is thought to be a synthetic language, because of its high number of verb morphemes, but in the CHILDES segmentation system for Japanese (Miyata, 2000; Miyata & Naka, 1998), these morphemes were treated as separate words (since these affixes were easy to demarcate and were simple in meaning, e.g., the verb "dekirundayo" was segmented as "dekiru n da yo") and this means that this Japanese corpus was more analytic than synthetic. These results suggested that unigram frequency could have a negative influence on prediction with analytic languages.

To test this hypothesis statistically, we need to divide the languages into those that are more analytic and those that are more synthetic. But since this typological classification depends on several theory-specific factors (e.g., number of morphemes in a language) as well as corpus-specific factors (e.g., word segmentation), we will approximate the subjective linguistic classification with an objective classification based on the ratio of unique word types to total word tokens in the corpus. A synthetic language will have a high type/token ratio, because a word token will tend to be a unique combination of morphemes and hence a unique word type, while in an analytic language, many of the word tokens will come from a relatively small set of word types. When these

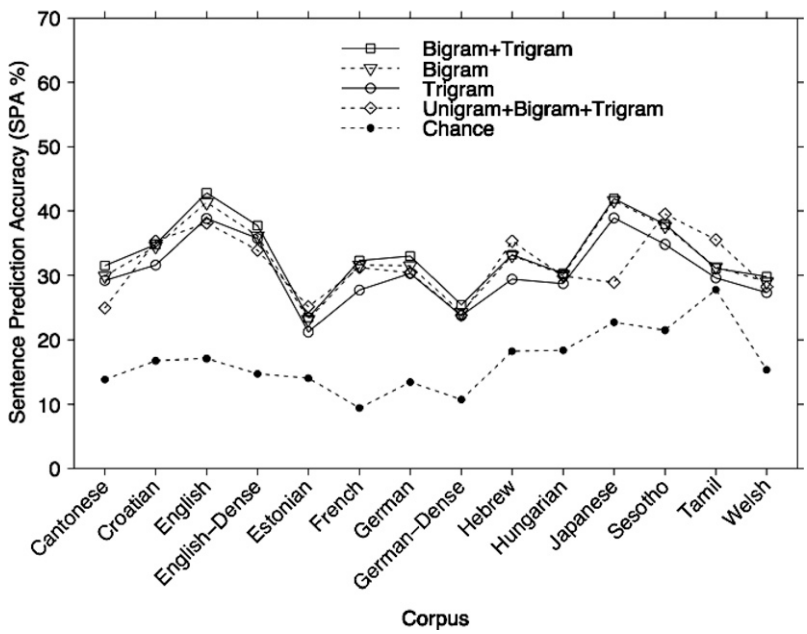

Fig. 2. SPA scores (%) for *n*-gram learners in Adult–Adult prediction by corpus.

Author's personal copy

ratios were computed for our corpora, the six languages with high ratios included the languages that are thought to be synthetic (Croatian, 0.07; Estonian, 0.08; Hebrew, 0.12, Hungarian, 0.14; Sesotho, 0.08; Tamil, 0.21; Welsh, 0.07), while the six languages with low ratios included the relatively more analytic languages (Cantonese, 0.03; English, 0.02; English-Dense, 0.01, French, 0.04; German, 0.03, German-Dense, 0.02; Japanese, 0.05). French, German, and Japanese are sometimes labeled as being synthetic languages, since they are more synthetic than English, but they are less synthetic than rich morphological languages like Croatian (Corbett, 1987), where noun morphology depends on gender (masculine, feminine, neuter), number (singular, plural), and case (nominative, vocative, accusative, genitive, dative, locative, instrumental). To test the hypothesis about the role of unigram statistics in different language typologies, we computed the difference between the SPA score for the Unigram+Bigram+Trigram learner and the Bigram+Trigram learner for each corpus, and then did a Welch two-sample t-test to compare the analytic and synthetic group. The difference in the SPA score for the analytic group (-4.77) was significantly lower than the difference for the synthetic group (1.19, $t(8.5) = 3.5$, $p = 0.007$), which suggests that the unigram frequencies did negatively reduce the accuracy of prediction in analytic languages.

Testing these n-gram-based learners in the BIG–SPA task yielded results that seem comparable to results with other evaluation measures. Although, a similar systematic comparison of n-gram-based learners in typologically-different languages with other evaluation measures has not been done, the results here are consistent with the intuition that there is a greater likelihood of input-test overlap for bigrams than trigrams, and trigrams are likely to be more informative than bigrams when available, and therefore algorithms that are smoothed with several statistics (Bigram+Trigram) are better able to deal with data sparseness than unsmoothed algorithms (Bigram learner). An unexpected result was that a smoothed trigram (Unigram+Bigram+Trigram) learner was numerically worse (although not significantly so) than the Bigram+Trigram learner. This seemed to be due the lower SPA scores for the Unigram+Bigram+Trigram learner in analytic languages, which suggests that unigram frequencies in certain language typologies might have a negative impact on word ordering processes. Since the BIG–SPA task made it possible to test multiple typologically-different languages, it allowed us to ask questions about how well the differences between learners generalized to a wider space of languages and whether there were typological biases in a set of learners.

4. BIG–SPA evaluation of Adjacency–Prominence-type syntactic learners

One goal of the BIG–SPA task is to allow comparison of learners from different domains. In this section, we examined a psychological account of syntax acquisition and compared it with one of the language models that we presented earlier. Psychological accounts of syntax acquisition/processing assume that multiple different factors or constraints (e.g., semantic, syntactic, lexical) influence processing choices at different points in a sentence (Bock, 1982; Hirsh-Pasek & Golinkoff, 1996; MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell, Sekerina, Hill, & Logrip, 1999). The computational complexity of these theories often means that models of these theories can only be tested on toy languages (Chang, Dell, & Bock, 2006; Miikkulainen & Dyer, 1991; St.-John & McClelland, 1990), while systems that are designed for real corpora tend to use simpler statistics that can be used with known optimization techniques (e.g., Langkilde-Geary, 2002). Since the BIG–SPA task incorporates features that are important in psycholinguistic theories, e.g., incrementality, it might be easier to implement ideas from psychological theories within this task.

Here we examined a corpus-based learner that was based on an incremental connectionist model of sentence production and syntax acquisition called the Dual-path model (Chang, 2002; Chang et al., 2006). The model accounted for a wide range of syntactic phenomena in adult sentence production and syntax acquisition. It learned abstract syntactic representations from meaning–sentence pairs and these representations allowed the model to generalize words in a variable-like manner. It accounted for 12 data points on how syntax is used in adult structural priming tasks and six phenomena in syntax acquisition. And lesions to the architecture yielded behavioral results that approximate double dissociations in aphasia. Since it could model both processing and acquisition phenomena, it provided a set of useful hypotheses for constructing a corpus-based learner that could both learn syntactic knowledge from the input and use that knowledge in sentence generation.

The Dual-path model had two pathways called the sequencing and meaning pathways (Fig. 3). The sequencing pathway incorporated a simple recurrent network (Elman, 1990) that learned statistical relationships over sequences, and this part of the architecture was important for modeling behavior related to abstract syntactic categories in production. The meaning pathway had a representation of the message that was to be produced, but it was completely dependent on the sequencing pathway for sequencing information. Hence, the meaning system instantiated a competition between the available concepts in the speaker's message. By having an architecture with these two pathways, the resulting model learned different types of information in each pathway and also learned how to integrate this information in production. The dual-pathways architecture was critical then to the model's ability to explain how abstract syntax was learned and used in sentence production.

The dual-pathways architecture suggested that a corpus-based syntax learner should have separate components that focus on sequencing constraints and meaning-based constraints. The sequencing component of this learner was

Author's personal copy

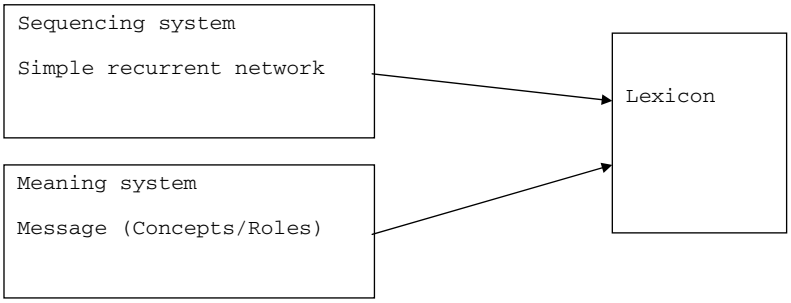

Fig. 3. The architecture of the Dual-path Model (Chang, 2002).

implemented with an *n*-gram adjacency statistic like the learners that we tested earlier. The meaning component of this learner was based on the message-based competition in the meaning system in the Dual-path model. One way to view the operation of the Dual-path's meaning system is that it instantiated a competition between the elements of the message and more prominent elements tended to win this competition and were therefore placed earlier in sentences. This message-based competition can be modeled by constructing a prominence hierarchy for each utterance. Since we used the bag-of-words to model the constraining influence of the message, our prominence hierarchy was instantiated over words and it was implemented by recording which words preceded other words in the input utterances, on the assumption that words that come earlier in utterances are more prominent on average than words that come later in utterances. The learner that incorporated the adjacency statistic and the prominence hierarchy was called the Adjacency–Prominence learner.

To illustrate how these statistics were collected, the example sentence "Is that a nice home for the bus?" will be used (Fig. 4). To represent adjacency information in this learner, a bigram frequency was collected (rightward arrows on the top side of Fig. 4). To model the prominence hierarchy, a *prominence* frequency was collected, which encoded how often a word preceded the other words in the sentence separated by any number of words (leftward arrows on the bottom side of Fig. 4). To normalize these frequency counts, they were divided by the frequency that the two words occurred in together in the same utterance in any order (this was called the paired frequency). When the bigram frequency was divided by the paired frequency, it was called the *adjacency* statistic and when the prominence frequency was divided by the paired frequency, it

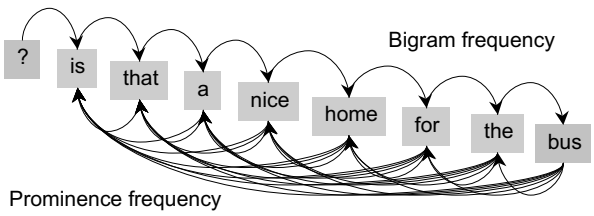

Fig. 4. Bigram and prominence frequencies for the utterance "Is that a nice home for the bus?".

was called the *prominence* statistic. While it is possible to use a smoothed trigram statistic as the adjacency statistic, the adjacency statistic that was used was kept as a simple bigram to emphasize the role that the prominence statistic and the paired frequency might play in the behavior of the learner. The Adjacency–Prominence learner combined the adjacency and prominence statistics together to incrementally pick the next word in a sentence.

To demonstrate how these statistics were used by the Adjacency–Prominence learner, we will work through an example test utterance "is that nice?" (Fig. 5). In this example, we assume that adjacency and prominence statistics have been collected over an English corpus. To start the production of the test sentence, the previous word is set to the punctuation symbol ("?" in top left box in Fig. 5) and the bag-of-words is set to the words "is", "nice", and "that" (bottom left box in Fig. 5). For each of the words in the lexicon, a Choice score is collected, which represents the combined activation from the adjacency and prominence statistics (right box in Fig. 5). Since questions tend to start with words like "is" more than words like "that" or "nice" (e.g., "Is that a nice home for the bus?"), the Choice score for "is" will be higher due to the adjacency statistics (arrows from "?" to "is"). And since "is" and "that" can occur in both orders in the input (e.g., "that is nice", "is that nice"), the prominence statistics will not pull for either order (there are arrows to both "is" and "that" from the prominence statistics in Fig. 5). The word that is produced is the word with the highest Choice score. Since "is" has the most activation here (three arrows in Fig. 5), it is produced as the first word in the sentence. Then, the process is started over again with "is" as the new previous word and the bag-of-words reduced to just the words "nice" and "that". Since "is" is followed by both "that" and "nice" in the input, the adjacency statistics might not be strongly biased to one or the other word. But since "that" tends to occur before "nice" in general (e.g. "does that look nice?"), the prominence statistics will prefer to put "that" first. Since "that" has the strongest Choice score, it is produced next, and then the process starts over again. Since there is only one word "nice" in the bag-of-words, it is produced. Since the produced utterance ("is that nice") matches the target utterance, the SPA score for this one sentence corpus is 100%.

Author's personal copy

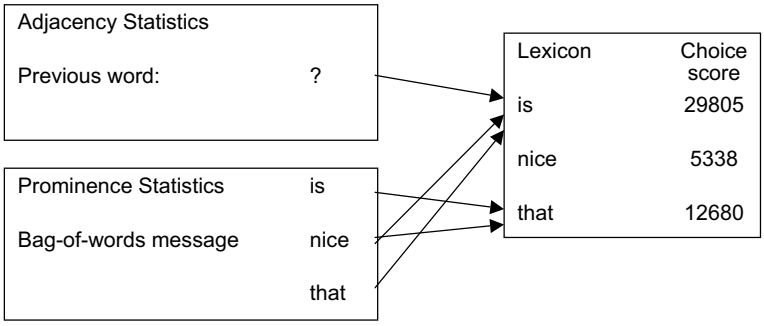

Fig. 5. Example of production of the first word of "is that nice?" in Adjacency–Prominence learner.

In order to understand the behavior of the Adjacency–Prominence learner, we also created learners that just used the adjacency statistics (Adjacency-only) or just used the prominence statistics (Prominence-only). We also included the Chance learner as a baseline and the Bigram learner from the previous section, because the adjacency statistic in the Adjacency-only learner differed from the equation in the standard Bigram learner. The adjacency statistic and the bigram statistic had the same numerator (e.g., frequency of the adjacent words A and B in the input), but they had different denominators (unigram frequency of word A vs. paired frequency of both words). Unlike the bigram statistic, the paired frequency took into account the unigram frequency of the word B. The comparison of the Adjacency-only learner with the Bigram learner will allow us to determine which approach for adjacency statistics provides a better account of utterance prediction. The statistics and Choice functions for the learners are defined below.

Definition of statistics used in learners

$C(w_n)$	Frequency of unigram w_n (unigram frequency)
$C(w_{n-1}, w_n)$	Frequency of bigram w_{n-1} w_n (bigram frequency)
$\mathrm{P}(w_a, w_b)$	Frequency that word w_a occurs before w_b in an utterance at any distance (prominence frequency)
$\mathrm{Pair}(w_a, w_b)$	Frequency that words w_a and w_b occur in the same utterance in any order (paired frequency)
length	Number of words in the bag-of-words

Equations for four different learners

Bigram	$\mathrm{Choice}(w_n) = C(w_{n-1}, w_n)/C(w_{n-1})$
Adjacency-only	$\mathrm{Choice}_{\mathrm{adjacency}}(w_n) = C(w_{n-1}, w_n)/\ \mathrm{Pair}(w_{n-1}, w_n)$
Prominence-only	$\mathrm{Choice}_{\mathrm{prominence}}(w_n) = \sum_{W_b} \mathrm{P}(w_n, w_b)/\mathrm{Pair}(w_n, w_b)$ for all w_b in the bag-of-words, except w_n
Adjacency–Prominence	$\mathrm{Choice}(w_n) = \mathrm{length} * \mathrm{Choice}_{\mathrm{adjacency}}(w_n) + \mathrm{Choice}_{\mathrm{prominence}}(w_n)$

Fig. 6 shows the results for the Adult–Child (adult input, child test) and Adult–Adult (90% adult input, 10% adult test). One question is whether bigram frequency should be divided by unigram frequency of the previous word (Bigram learner) or paired frequency of both words (Adjacency-only learner). We found that the Adjacency-only learner was better than Bigram learner in both testing situations (Adult–Child, $t(13) = 5.0$, $p < 0.001$; Adult–Adult, $t(13) = 7.8$, $p < 0.001$). An example of the difference between these two learners can be seen with the Adult sentence "? do you want me to draw a cat", which the Adjacency-only learner correctly produced and the Bigram learner mistakenly produced as "? do you want to to draw a cat". The reason that the learner incorrectly produced "to" instead of "me" was because the standard bigram equation had an artificially strong statistic for "want" → "to", because it did not recognize that "to" was a very frequent word by itself (the denominator only has the unigram frequency of "want"). In the Adjacency-only learner, the adjacency statistic was the frequency that "want" proceeds "to" divided by the paired frequency that "want" and "to" occurred in the same sentence in any order. The adjacency statistic was weaker for the word "to" after "want", because "want" and "to" were often non-adjacent in an utterance, and this allowed the word "me" to win out. This suggests that for word order prediction, the frequency that both words occur in the same utterance is an important constraint for adjacent word statistics.

Another question is whether there is evidence that supports the assumption of the Dual-path model that a syntax acquisition mechanism will work better if it combines separate statistics for sequencing and meaning. Since we have demonstrated that sequencing statistics like the Adjacency-only or *n*-gram statistics are useful, the main question is whether the prominence statistics, that depend on our bag-of-words simulated message, will augment or interfere with the predictions of the sequencing statistics. We found that in both testing situations, Adjacency–Prominence was better than Adjacency-only (Adult–Child, $t(13) = 7.4$, $p < 0.001$; Adult–Adult, $t(13) = 10.5$, $p < 0.001$) and Prominence-only (Adult–Child, $t(13) = 12.2$, $p < 0.001$; Adult–Adult,

Author's personal copy

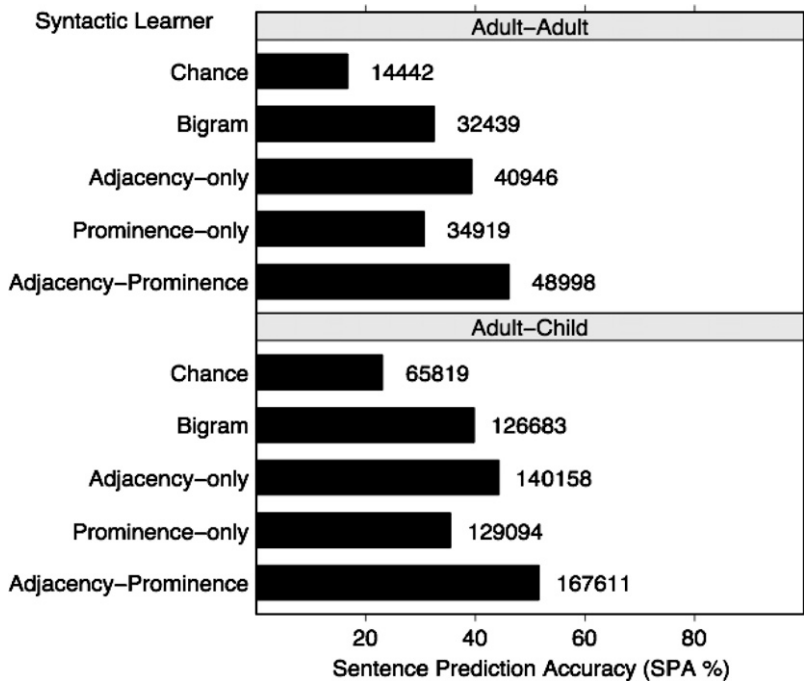

Fig. 6. Average SPA scores (%) for five learners in Adult–Adult and Adult–Child prediction (counts of correct utterances are placed to the right of each bar).

$t(13) = 17.8$, $p < 0.001$). The Adjacency–Prominence learner correctly predicted 27,453 more utterances than the Adjacency-only learner over the corpora in the Adult–Child situation, and 38,517 more than the Prominence-only learner.

These results suggest that the adjacency and prominence statistics capture different parts of the problem of word order prediction and these statistics integrate together without interfering with each other. This is partially due to the way that the Adjacency–Prominence learner used each statistic. The influence of the adjacency statistics came from the past (the previous word), while the influence of the prominence statistics depended on the future (the words to be produced in the bag-of-words message). Also these two statistics have different scopes, where the adjacency statistics captured linear relationships between words, while the prominence statistics handled some of the hierarchical relationships between words. For example, the Adjacency–Prominence learner was able to predict a sentence with multiple prepositional phrases like "you've been playing with your toy mixer in the bathroom for a few weeks" in the Adult–Adult test, because the adjacency statistics recorded the regularities between the words in the sequences "in the bathroom" and "for a few weeks" in other sentences in the input, while the prominence statistics recorded the fact that "in" preceded "for" more often than the other way around (e.g., "put those **in** the bin **for** mummy please"). In addition to capturing relations of different scopes, these two statistics also differed in their availability and their reliability. Since prominence statistics were collected for all the pairs of words in an input utterance at any distance, they were more likely to be present at test than the adjacency statistic which only existed if that particular pair of words in that order occurred in the input. These two statistics worked together well, because the prominence statistics were likely to overlap between input and test but only encoded general position information, while the adjacency statistics, when they existed, were guaranteed to predict only grammatical transitions.

The results were broken down for each individual corpus (Fig. 7). The significant difference between the means for the Bigram, Adjacency-only, and Adjacency–Prominence learners was evident in each of the individual languages. Only the Prominence-only learner had a different pattern. The prominence statistics seemed to have a typology-specific bias, since they seemed to be more useful in analytic languages (e.g., Cantonese, English, English-Dense, Japanese) than in synthetic languages (e.g., Croatian, Estonian, Hebrew, Hungarian, Sesotho, and Tamil). The effect of prominence statistics was evident in the difference between the Adjacency–Prominence learner and the Adjacency-only learner. This difference was significantly higher for analytic languages (8.70%) than for synthetic languages (4.97%, $t(11.8) = 4.54$, $p < 0.001$) suggesting that the prominence statistics improved performance over adjacency statistics more in analytic languages. Prominence statistics recorded all pairwise relationships between words in a sentence, and these types of statistics could make use of the greater contextual information associated with frequent words. So while the frequent words in analytic languages can be problematic for systems that use unigrams, they can be beneficial for systems that use prominence statistics.

Author's personal copy

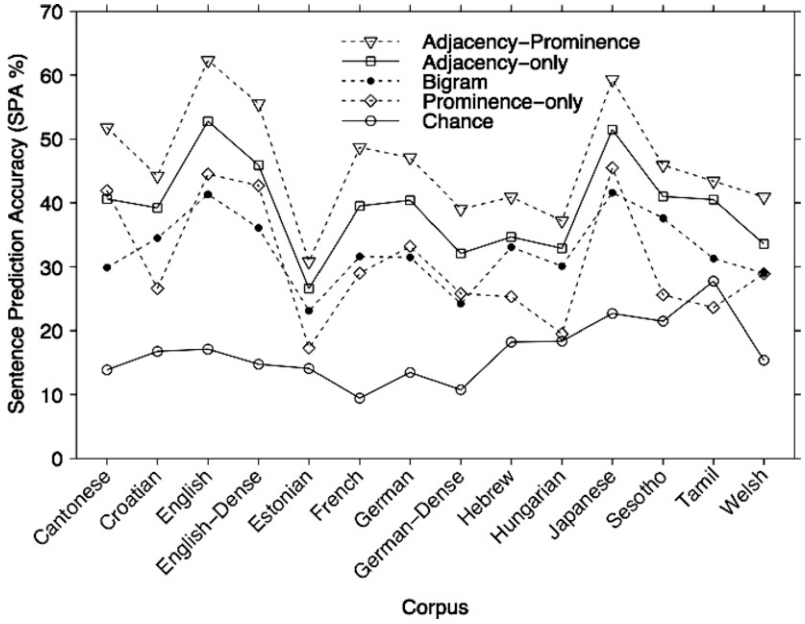

Fig. 7. SPA scores (%) for five learners in Adult–Adult prediction by corpus.

In this section, we compared a learner that made use of statistics that are commonly used in computational linguistics (Bigram learner) with a learner that was inspired by psychological accounts of human syntax processing (Dual-path model → Adjacency–Prominence learner). We found that the Adjacency–Prominence learner worked better than the Bigram learner across the 14 corpora, and this was both because it modulated its statistics with information about the set of possible words (Adjacency-only vs. Bigram comparison) and it combined two statistics that captured different aspect of the problem of generating word order (Adjacency–Prominence vs. Adjacency-only and Prominence-only). In addition, the SPA results broken down by corpus suggested that prominence statistics were biased for analytic languages and this suggests that a typologically-general approach for syntax acquisition should pay attention to the analytic/synthetic distinction.

5. Conclusion

Machine translation was transformed by the incorporation of statistical techniques and the creation of automatic evaluation measures like BLEU. Likewise, explicit theories of human syntax acquisition might also be improved by having an automatic evaluation task that does not depend on human intuitions and which can be used in different languages, and the BIG–SPA task is one method for accomplishing this. Although the BIG–SPA task is similar to statistical machine translation tasks, it differs in some important ways. The SPA measure is a stricter sentence level evaluation measure which is more appropriate for the evaluation of syntactic knowledge. The BIG task is closer to psychological theories of language production, because it does utterance generation in an incremental manner from a constrained set of concepts (as encoded by the bag-of-words). If theories of syntax acquisition were made explicit and tested with BIG–SPA, it would be easier to compare them with learners from other domains, such as computational linguistics, and this might allow a greater cross-fertilization of ideas.

Although many computational linguistics algorithms use combinations of *n*-grams, there has been relatively little work systematically comparing different *n*-gram learners in a large set of typologically-different languages. While the differences between different combinations of *n*-gram learners in the BIG–SPA task matched our expectations, the overall accuracy of these *n*-gram learners was fairly low (<45% SPA). This is because SPA is a challenging metric, where 100% accuracy requires that all of the words in all of the utterances in a particular corpus are correctly sequenced, and therefore it is not expect that *n*-gram learners trained on small input corpus will be able to achieve high accuracy on this measure. Rather, these *n*-gram models can be seen as default or baseline learners that can be used for comparison with learners that incorporate more sophisticated learning mechanisms.

To improve a syntactic learner, researchers often embed some constraints of the language or the task into their system to improve its performance. But this is made more difficult when testing typologically-different languages, since one cannot embed properties of a particular language (e.g., its tagset) into the learner. And incorporating abstract syntactic universals into a learner is difficult because these universals often depend on linguistic categories (e.g., noun, phrasal head) and it is difficult to label these linguistic categories in an equivalent

Author's personal copy

way across typologically-different languages. Another approach for improving learners is to incorporate knowledge about the task into the learner. Since BIG–SPA task mimics the task of sentence production, we used ideas from a psycholinguistic model of sentence production to develop the Adjacency–Prominence learner and it was found to have the highest accuracy for utterance prediction of all the systems tested. This can be attributed to the fact that it used its Adjacency and Prominence statistics in very different ways. In particular, the influence of the Prominence statistics changed as the set of words in the bag-of-words diminished. This kind of dynamically-changing statistic is not typically used in computational linguistic approaches to bag generation, since these approaches do not normally view sentence planning as an incremental process that adjusts to both the words that have been produced, but also to the set of message concepts that the speaker has yet to produce. The BIG task emphasizes the way that information changes over a sentence, and therefore this task might be a useful platform for comparing learners that use more dynamic learning approaches.

Since the BIG–SPA task does not require a gold standard for syntax, it can be used to compare syntactic learners in typologically-different languages. By using a typologically-diverse sample of languages, one can do statistics across the sample that allow generalization outside of the sample. This helps to insure that any hypothesized improvements in a syntactic learner are not simply optimizations for particular languages or particular corpora, but actually characterize something shared across the speakers of those languages. BIG–SPA task can also be used to look for typological biases in particular algorithms and that can help in the search for a syntax acquisition algorithm that can work on any human language. Since work in developmental psycholinguistics and computational linguistics is still predominately focused on a few major languages (European languages, Chinese, Japanese), it is still unclear whether many standard algorithms and theories would work equally well on all human languages (most of the 2650 languages in the World Atlas of Language Structures have never been tested, Haspelmath et al., 2005). Making theories explicit and testing them within the BIG–SPA task on a larger set of languages is one way to move towards a more general account of how humans learn syntax.

Acknowledgements

We would like to thank Dan Jurasky, David Reitter, Gary Dell, Morten Christiansen, and several anonymous reviewers for their comments on this work. Early versions of this manuscript were presented at the Cognitive Science Society Conference in 2005 (Stressa), 2006 (Vancouver), and the 2006 Japanese Society for the Language Sciences Conference (Tokyo).

Appendix

Table of corpora used.

Age of the child is specified in year; months. The utterance counts do not include single word utterances.

Corpora	Child	Database	Age	# of Child Utt.	# of Adult Utt.
Cantonese	Jenny	CanCorp (Lee et al., 1996)	2;8-3;8	8174	18,171
Croatian	Vjeran	Kovacevic (Kovacevic, 2003)	0;10-3;2	12,396	27,144
English	Anne	Manchester (Theakston, Lieven, Pine, and Rowland, 2001)	1;10-2;9	11,594	27,211
English-Dense	Brian	MPI-EVA (Maslen et al., 2004)	2;0-3;11	106,059	270,575
Estonian	Vija	Vija (Vihman and Vija, 2006)	1;7-3;1	23,667	20,782
French	Phil	Leveillé (Suppes, Smith, and Leveillé, 1973)	2;1-3;3	10,498	17,587
German	Simone	Nijmegen (Miller, 1976)	1;9-4;0	14,904	62,187
German-Dense	Leo	MPI-EVA (Abbot-Smith and Behrens, 2006)	1;11-4;11	68,931	198,326
Hebrew	Lior	Berman Longitudinal (Berman, 1990)	1;5-3;1	3005	6952
Hungarian	Miki	Réger (Réger, 1986)	1;11-2;11	4142	8668
Japanese	Tai	Miyata-Tai (Miyata, 2000)	1;5-3;1	19,466	29,093
Sesotho	Litlhare	Demuth (Demuth, 1992)	2;1-3;2	9259	13,416
Tamil	Vanitha	Narasimhan (Narasimhan, 1981)	0;9-2;9	1109	3575
Welsh	Dewi	Jones (Aldridge, Borsley, Clack, Creunant, and Jones, 1998)	1;9-2;6	4358	4551

References

Abbot-Smith, K., & Behrens, H. (2006). How known constructions influence the acquisition of new constructions: The German peri-

Author's personal copy

phrastic passive and future constructions. *Cognitive Science, 30*(6), 995–1026.

Aldridge, M., Borsley, R. D., Clack, S., Creunant, G., & Jones, B. M. (1998). The acquisition of noun phrases in Welsh. In *Language acquisition: Knowledge representation and processing. Proceedings of GALA'97*. Edinburgh: University of Edinburgh Press.

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C., & Wilcock, S. (2000). A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal, 24*, 7–23.

Berman, R. A. (1990). Acquiring an (S)VO language: Subjectless sentences in children's Hebrew. *Linguistics, 28*, 1135–1166.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review, 89*(1), 1–47.

Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(4), 575–586.

Bock, J. K., & Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Learning & Verbal Behavior, 19*(4), 467–484.

Bock, K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review, 99*(1), 150–171.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition, 21*(1), 47–67.

Brandt, S., Diessel, H., & Tomasello, M. (in press). The acquisition of German relative clauses: A case study. *Journal of Child Language*.

Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263–311.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science, 26*(5), 609–651.

Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272.

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (in press). Categorizing words using "Frequent Frames: What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*.

Chomsky, N. (1980). *Rules and representations*. Oxford: Basil Blackwell.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Church, K. W. (1989). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of ICASSP-89, Glasgow, Scotland*.

Comrie, B. (Ed.). (1987). *The world's major languages*. Oxford, UK: Oxford University Press.

Corbett, G. (1987). Serbo-Croat. In B. Comrie (Ed.), *The world's major languages*. Oxford, UK: Oxford University Press.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford, UK: Oxford University Press.

Demuth, K. (1992). Acquisition of Sesotho. In D. Slobin (Ed.). *The cross-linguistic study of language acquisition* (Vol. 3, pp. 557–638). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dermatas, E., & Kokkinakis, G. (1995). Automatic stochastic tagging of natural language texts. *Computational Linguistics, 21*(2), 137–163.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language, 35*(5), 724–755.

Ferreira, V. S., & Yoshita, H. (2003). Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research, 32*, 669–692.

Francis, W. N., & Kucera, H. (1979). Brown corpus manual [Electronic Version] from http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM.

Germann, U., Jahr, M., Knight, K., Marcu, D., & Yamada, K. (2004). Fast decoding and optimal decoding for machine translation. *Artificial Intelligence, 154*(1–2), 127–143.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Hajič, J., & Vidová-Hladká, B. (1997). Probabilistic and rule-based tagger of an inflective language – A comparison. In *Proceedings of the fifth conference on applied natural language processing, Washington DC, USA*.

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

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge, UK: Cambridge University Press.

Hirsh-Pasek, K., & Golinkoff, R. M. (1996). *The origins of grammar: Evidence from early language comprehension*. Cambridge, MA: MIT Press.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Klein, D., & Manning, C. D. (2004). Corpus-based induction of syntactic structure: Models of depenency and constituency. In *Proceedings of the 42nd annual meeting of the ACL*.

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics, 25*(4), 607–615.

Kovacevic, M. (2003). *Acquisition of Croatian in crosslinguistic perspective*. Zagreb.

Langkilde-Geary, I. (2002). An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the international natural language generation conference, New York City, NY*.

Lee, T. H. T., Wong, C. H., Leung, S., Man, P., Cheung, A., Szeto, K., et al. (1996). *The development of grammatical competence in Cantonese-speaking children*. Hong Kong: Department of English, Chinese University of Hong Kong (Report of a project funded by RGC earmarked grant, 1991–1994).

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language, 30*(2), 333–367.

Li, C. N., & Thompson, S. A. (1990). Chinese. In B. Comrie (Ed.), *The world's major languages* (pp. 811–833). Oxford, UK: Oxford University Press.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676–703.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Maslen, R., Theakston, A., Lieven, E., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language and Hearing Research, 47*, 1319–1333.

Miikkulainen, R., & Dyer, M. G. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science, 15*(3), 343–399.

Miller, M. (1976). *Zur Logik der frühkindlichen Sprachentwicklung: Empirische Untersuchungen und Theoriediskussion*. Stuttgart: Klett.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91–117.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*(4), 393–424.

Miyata, S. (2000). The TAI corpus: Longitudinal speech data of a Japanese boy aged 1;5.20–3;1.1. *Bulletin of Shukutoku Junior College, 39*, 77–85.

Miyata, S., & Naka, N. (1998). *Wakachigaki Guideline for Japanese: WAKACHI98 v.1.1*. The Japanese Society for Educational Psychology Forum Report No. FR-98-003, The Japanese Association of Educational Psychology.

Narasimhan, R. (1981). *Modeling language behavior*. Berlin: Springer.

Author's personal copy

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). *Bleu: A method for automatic evaluation of machine translation* (No. RC22176 (W0109-022)). Yorktown Heights, NY: IBM Research Division, Thomas J. Watson Research Center.

Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics, 18*(2), 123–138.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language, 42*(2), 168–182.

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19*, 9–50.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science, 29*, 1007–1028.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425–469.

Redington, M., Chater, N., Huang, C., Chang, L.-P., Finch, S., & Chen, K. (1995). The universality of simple distributional methods: Identifying syntactic categories in Chinese. In *Proceedings of the cognitive science of natural language processing, Dublin*.

Réger, Z. (1986). The functions of imitation in child language. *Applied Psycholinguistics, 7*(4), 323–352.

Smith, M., & Wheeldon, L. R. (1999). High level processing scope in spoken sentence production. *Cognition, 73*, 205–246.

Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language, 39*(3), 392–417.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*(1–2), 217–257.

Suppes, P., Smith, R., & Leveillé, M. (1973). The French syntax of a child's noun phrases. *Archives de Psychologie, 42*, 207–269.

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*(1), 127–152.

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. (1992). *First verbs: A case study of early grammatical development*. Cambridge: Cambridge University Press.

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

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language, 31*, 101–121.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73*(2), 89–134.

Tseng, H., Jurafsky, D., & Manning, C. (2005). Morphological features help POS tagging of unknown words across language varieties. In *Proceedings of the fourth SIGHAN workshop on Chinese language processing*.

Tsujimura, N. (1996). *An introduction to Japanese linguistics*. Cambridge, MA: Blackwell Publishers Inc..

Vihman, M. M., & Vija, M. (2006). The acquisition of verbal inflection in Estonian: Two case studies. In N. Gagarina & I. Gluzow (Eds.), *The acquisition of verbs and their grammar: The effect of particular languages* (pp. 263–295). Dordrecht: Springer.

Yamashita, H., & Chang, F. (2001). Long before short preference in the production of a head-final language. *Cognition, 81*(2), B45–B55.