Modeling human performance in statistical word segmentation

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. Cognition, 117(2), 107-125doi: 10.1016/j.cognition.2010.07.005

Modeling human performance in statistical word segmentation

Michael C. Frank
Department of Psychology, Stanford University

Sharon Goldwater
School of Informatics, University of Edinburgh

Thomas L. Griffiths
Department of Psychology, University of California, Berkeley

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

The ability to discover groupings in continuous stimuli on the basis of distributional information is present across species and across perceptual modalities. We investigate the nature of the computations underlying this ability using statistical word segmentation experiments in which we vary the length of sentences, the amount of exposure, and the number of words in the languages being learned. Although the results are intuitive from the perspective of a language learner (longer sentences, less training, and a larger language all make learning more difficult), standard computational proposals fail to capture several of these results. We describe how probabilistic models of segmentation can be modified to take into account some notion of memory or resource limitations in order to provide a closer match to human performance.

Human adults and infants, non-human primates, and even rodents all show a surprising ability: presented with a stream of syllables with no pauses between them, individuals from each group are able to discriminate statistically coherent sequences from sequences with lower coherence (Aslin, Saffran, & Newport, 1998; Hauser, Newport, & Aslin, 2001; Saffran, Johnson, Aslin, & Newport, 1999; Saffran, Newport, & Aslin, 1996; Toro & Trobalon, 2005). This ability is not unique to linguistic stimuli (Saffran et al., 1999) or to the auditory domain (Conway & Christiansen, 2005; Kirkham, Slemmer, & Johnson, 2002), and is not constrained to temporal sequences (Fiser & Aslin, 2002) or even to the particulars of perceptual stimuli (Brady & Oliva, 2008). This "statistical learning" ability may be useful for a large variety of tasks but is especially relevant to language learners who must learn to segment words from fluent speech.

Yet despite the scope of the "statistical learning" phenomenon and the large literature surrounding it, the computations underlying statistical learning are at present unknown. Following an initial suggestion by Harris (1951), work on this topic by Saffran and colleagues (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) suggested that learners could succeed in word segmentation by computing transitional probabilities between syllables and using low-probability transitions as one possible indicator of a boundary between words. More recently, a number of investigations have used more sophisticated computational models to attempt to characterize the computations performed by human learners in word segmentation (Giroux & Rey, 2009) and visual statistical learning (Orbán, Fiser, Aslin, & Lengyel, 2008) tasks.

The goal of the current investigation is to extend this previous work by evaluating a larger set of models against new experimental data describing human performance in statistical word segmentation tasks. Our strategy is to investigate the fit of segmentation models to human performance. Because existing experiments show evidence of statistical segmentation but provide only limited quantitative results about segmentation under different conditions, we parametrically manipulate basic factors leading to difficulty for human learners to create a relatively detailed dataset with which to evaluate models.

The plan of the paper is as follows. We first review some previous work on the computations involved in statistical learning. Next, we make use of the adult statistical segmentation paradigm of Saffran, Newport, and Aslin (1996) to measure human segmentation performance as we vary three factors: sentence length, amount of exposure, and number of words in the language. We then evaluate a variety of segmentation models on the same dataset and find that although some of the results are well-modeled by some subset of models, no model captures all three results. We argue that the likely cause of this failure is the lack of memory constraints on current models. We conclude by considering methods for modifying models of segmentation to better reflect the memory constraints on human learning.

There are three contributions of this work: we introduce a variety of new human data about segmentation un-

We gratefully acknowledge Elissa Newport and Richard Aslin for many valuable discussions of this work and thank LouAnn Gerken, Pierre Perruchet, and two anonymous reviewers for comments on the paper. Portions of the data in this paper were reported at the Cognitive Science conference in Frank, Goldwater, Mansinghka, Griffiths, and Tenenbaum (2007). We acknowledge NSF grant #BCS-0631518, and the first author was supported by a Jacob Javits Graduate Fellowship and NSF DDRIG #0746251.

Please address correspondence to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall, Jordan Hall (Building 420), Stanford, CA 94305, tel: (650) 724-4003, email: mcfrank@stanford.edu.

der a range of experimental conditions, we show an important limitation of a number of proposed models, and we describe a broad class of models—memory-limited probabilistic models—which we believe should be the focus of attention in future investigations.

Computations Underlying Statistical Learning

Investigations of the computations underlying statistical learning phenomena have followed two complementary strategies. The first strategy is the strategy of evaluating model *sufficiency*: whether a particular model, given some fixed amount of data, will converge to the correct solution. If a model does not converge to the correct solution within the amount of data available to a human learner, either the model is incorrect or the human learner relies on other sources of information to solve the problem. The second strategy evaluates *fidelity*: the fit between model performance and human performance across a range of different inputs. To the extent that a model correctly matches the pattern of successes and failures exhibited by human learners, it can be said to provide a better theory of human learning.

Investigations of the sufficiency of different computational proposals for segmentation have suggested that transitional probabilities may not be a viable segmentation strategy for learning from corpus data (Brent, 1999b). For example, Brent (1999a) evaluated a number of computational models of statistical segmentation on their ability to learn words from infant-directed speech and found that a range of statistical models were able to outperform a simpler transitional probability-based model. A more recent investigation by Goldwater, Griffiths, and Johnson (2009) built on Brent's modeling work by comparing a *unigram model*, which assumed that each word in a sentence was generated independently of each other word, to a *bigram model* which assumed sequential dependencies between words. The result of this comparison was clear: the bigram model substantially outperformed the unigram model because the unigram model tended to undersegment the input, mis-identifying frequent sequences of words as single units (e.g. "whatsthat" or "inthehouse"). Thus, incorporating additional linguistic structure into models may be necessary to achieve accurate segmentation. In general, however, the model described by Goldwater et al. (2009) and related models (Johnson, 2008; Liang & Klein, 2009) achieve the current state-of-the-art in segmentation performance due to their ability to find coherent units (words) and estimate their relationships within the language.

It may be possible that human learners use a simple, undersegmenting strategy to bootstrap segmentation but then use other strategies or information sources to achieve accurate adult-level performance (Swingley, 2005). For this reason, investigations of the sufficiency of particular models are not alone able to resolve the question of what computations human learners perform either in artificial language segmentation paradigms or in learning to segment human language more generally. Thus, investigations of the fidelity of models to human data are a necessary part of the effort to characterize human learning. Since data from word segmentation tasks with human infants are largely qualitative in nature (Saffran, Aslin, & Newport, 1996; Jusczyk & Aslin, 1995), artificial language learning tasks with adults can provide valuable quantitative data for the purpose of distinguishing models.

Three recent studies have pursued this strategy. All three have investigated the question of the representations that are stored in statistical learning tasks and whether these representations are best described by *chunking* models or by *transition-finding* models. For example, Giroux and Rey (2009) contrasted the PARSER model of Perruchet and Vinter (1998) with a simple recurrent network, or SRN (Elman, 1990). The PARSER model, which extracts and stores frequent sequences in a memory register, was used as an example of a chunking strategy and the SRN, which learns to predict individual elements on the basis of previous elements, was used as an example of a transition-finding model. Giroux and Rey compared the fit of these models to a human experiment testing whether adults were able to recognize the sub-strings of valid sequences in the exposure corpus and found that PARSER fit the human data better, predicting sub-string recognition performance would not increase with greater amounts of exposure. These results suggest that PARSER may capture some aspects of the segmentation task that are not accounted for by the SRN. But because each model in this study represents only one particular instantiation of its class, a success or failure by one or the other does not provide evidence for or against the entire class.

In the domain of visual statistical learning tasks, Orbán et al. (2008) conducted a series of elegant behavioral experiments with adults that were also designed to distinguish chunking and transition-finding strategies. (Orbán et al. referred to this second class of strategies as associative rather than transition-finding, since transitions were not sequential in the visual domain). Their results suggested that the chunking model, which learned a parsimonious set of coherent chunks that could be composed to create the exposure corpus, provided a better fit to human performance across a wide range of conditions. Because of the guarantee of optimality afforded by the ideal learning framework that Orbán et al. (2008) used, this set of results provides slightly stronger evidence in favor of a chunking strategy. While Orbán et al.'s work still does not provide evidence against *all* transition-finding strategies, their results do suggest that it is not an idiosyncrasy of the learning algorithm employed by the transition-finding model that lead to its failure. Because this result was obtained in the visual domain, however, it cannot be considered conclusive for auditory statistical learning tasks, since it is possible that statistical learning tasks make use of different computations across domains (Conway & Christiansen, 2005).

Finally, a recent study by Endress and Mehler (2009) familiarized adults to a language which contained three-syllable words that were each generated via the perturbation of one syllable of a "phantom word" (labeled this way because the word was not ever presented in the experiment). At test, participants were able to distinguish words that actually

appeared in the exposure corpus from distractor sequences with low internal transition probabilities but not from phantom words. These data suggest that participants do not simply store frequent sequences; if they did, they would not have indicated that phantom words were as familiar as sequences they actually heard. However, the data are consistent with at least two other possible interpretations. First, participants may have relied only on syllable-wise transition probabilities (which would lead to phantom words being judged equally probable as the observed sequences). Second, participants might have been chunking sequences from the familiarization corpus and making the implicit inference that many of the observed sequences were related to the same prototype (the phantom word). This inference would in turn lead to a *prototype enhancement* effect (Posner & Keele, 1968), in which participants believe they have observed the prototype even though they have only observed non-prototypical exemplars centered around it. Thus, although these data are not consistent with a naïve chunking model, they may well be consistent with a chunking model that captures other properties of human generalization and memory.

To summarize: although far from conclusive, the current pattern of results is most consistent with the hypothesis that human performance in statistical learning tasks is best modeled by a process of chunking which may be limited by the basic properties of human memory. Rather than focusing on the question of chunking vs. transition-finding, our current work begins where this previous work leaves off, investigating how to incorporate basic features of human performance into models of statistical segmentation. Although some models of statistical learning have made use of ideas about restrictions on human memory (Perruchet & Vinter, 1998, 2002), for the most part, models of segmentation operate with no limits on either memory or computation. Thus, our goal in the current work is to investigate how these limitations can be modeled and how modeling these limitations can improve models' fit to human data.

We begin by describing three experiments which manipulate the difficulty of the learning task. Experiment 1 varies the length of the sentences in the segmentation language. Experiment 2 varies the amount of exposure participants were given to the segmentation language. Experiment 3 varies the number of words in the language. Taken together, participants' mean performance in these three experiments provides a set of data which we can use to investigate the fit of models.

Experiment 1: Sentence Length

When learning to segment a new language, longer sentences should be more difficult to understand than shorter sentences. Certainly this is true in the limit: individually presented words are easy to learn and remember, while those presented in long sentences with no boundaries are more difficult. In order to test the hypothesis that segmentation performance decreases as sentence length increases, we exposed adults to sentences constructed from a simple artificial lexicon. We assigned participants to one of eight sentence-length conditions so that we could estimate the change in their performance as sentence length increased.

Methods

Participants. We tested 101 MIT students and members of the surrounding community, but excluded three participants from the final sample based on performance greater than two standard deviations below the mean for their condition.

Materials. Each participant in the experiment heard a unique and randomly generated sample from a separate, randomly generated artificial language. The lexicon of this language was generated by concatenating 18 syllables (*ba*, *bi*, *da*, *du*, *ti*, *tu*, *ka*, *ki*, *la*, *lu*, *gi*, *gu*, *pa*, *pi*, *va*, *vu*, *zi*, *zu*) into six words, two with two syllables, two with three syllables, and two with four syllables. Sentences in the language were created by randomly concatenating words together without adjacent repetition of words. Each participant heard 600 words, consisting of equal numbers of tokens of each vocabulary item.

Participants were randomly placed in one of eight sentence length conditions (1, 2, 3, 4, 6, 8, 12, or 24 words per sentence). All speech in the experiment was synthesized using the MBROLA speech synthesizer (Dutoit, Pagel, Pierret, Bataille, & Van Der Vrecken, 1996) with the us3 diphone database, in order to produce an American male speaking voice. All consonants and vowels were 25 and 225ms in duration, respectively. The fundamental frequency of the synthesized speech was 100Hz. No breaks were introduced into the sentences: the synthesizer created equal co-articulation between every phone. There was a 500ms break between each sentence in the training sequence.

Test materials consisted of 30 target-distractor pairs. Each pair consisted of a word from the lexicon paired with a "part-word" distractor with the same number of syllables. Part-word distractors were created as in (Saffran, Newport, & Aslin, 1996): they were sequences of syllables of the same lengths as words, composed of the end of one word and the beginning of another (e.g. in a language with words *badutu* and *kagi*, a part word might be *dutuka*, which combines the last two syllables of the first word with the first syllable of the second). In all conditions except the length 1 condition, part-word sequences appeared in the corpus (although with lower frequency than true words). No adjacent test pairs contained the same words or part-word distractors.

Procedure. Participants were told that they were going to listen to a nonsense language for 15 minutes, after which they would be tested on how well they learned the words of the language. All participants listened on headphones in a quiet room. After they had heard the training set, they were instructed to make forced choice decisions between pairs of items from the test materials by indicating which one of the two "sounded more like a word in the language they just heard." No feedback was given during testing.

Results and Discussion

Performance by condition is shown in Figure 1, top left. Participants' individual data were highly variable, but

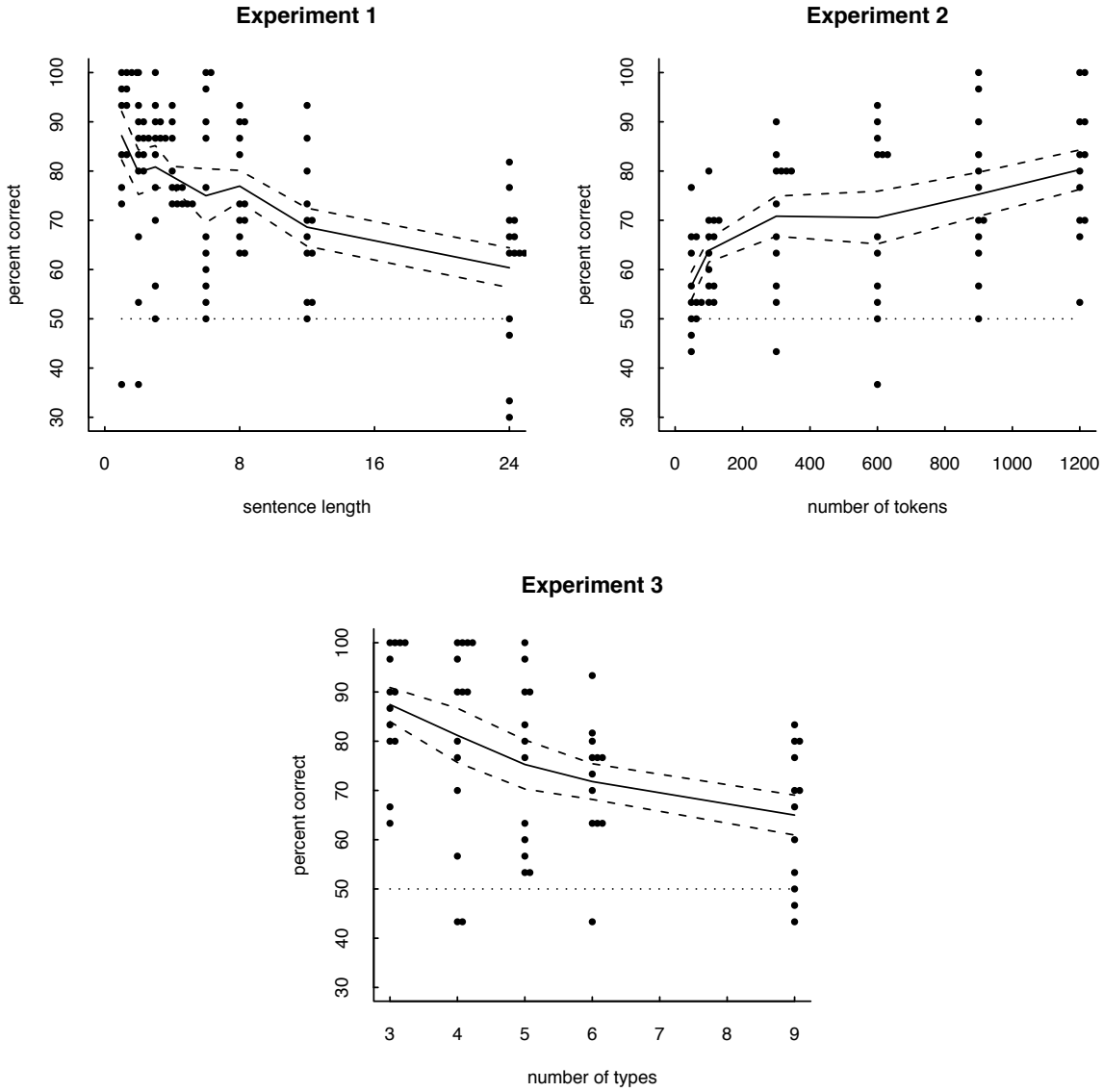

Figure 1. Data from Experiments 1 – 3. The percentage of test trials answered correctly by each participant (black dots) is plotted by sentence length, number of tokens, or number of types, respectively. Overlapping points are slightly offset on the horizontal axis to avoid overplotting. Solid lines show means, dashed lines show standard error of the mean, and dotted lines show chance.

showed a very systematic trend in their mean performance across conditions. Although the spread in participants' performance could have been caused by random variation in the phonetic difficulty of the languages we created, the variability was not greater than that observed in previous studies (Saffran, Newport, & Aslin, 1996). Thus, we focused on modeling and understanding mean performance across groups of participants, rather than individual performance.

We analyzed test data using a multilevel (mixed-effect) logistic regression model (Gelman & Hill, 2006). We included a group-level (fixed) effect of word length and a separate intercept term for each sentence-length condition. We also added a participant-level (random) effect of participant identity. We fit a separate model with an interaction of word length and sentence length condition but found that it did not significantly increase model fit ($\chi^2(7) = 10.67$, $p = .15$) so we pruned it from the final model.

There was no effect of word length ($\beta = -.00022$, $z = .028$, $p = .99$). In contrast, coefficient estimates for sentence lengths 1, 2, 3, 4, 6, and 8 were highly reliable ($\beta = 2.31$, 1.56, 1.64, 1.39, 1.32, and 1.31 respectively, and $z = 7.00$, 5.16, 5.10, 4.47, 4.24, and 4.20, all p-values $< .0001$), while length 12 reached a lower level of significance ($\beta = .86$, $z = 2.82$, and $p = .004$). Length 24 was not significant ($\beta = 0.45$, $z = 1.55$, $p = .12$), indicating that performance in this condition did not differ significantly

from chance. Thus, longer sentences were considerably more difficult to segment.

Experiment 2: Amount of Exposure

The more exposure to a language learners receive, the easier it should be for them to learn the words. To measure this relationship, we conducted an experiment in which we kept the length of sentences constant but varied the number of tokens (instances of words) participants heard.

Methods

Participants. We tested 72 MIT students and members of the surrounding community. No participants qualified as outliers by the criteria used in Experiment 1.

Materials. Materials in Experiment 2 were identical to those in Experiment 1, with one exception. We kept the number of words in a sentence constant at four words per sentence, but we manipulated the total number of words in the language sample that participants heard. Participants were randomly placed in one of six exposure length conditions (48, 100, 300, 600, 900, and 1200 total tokens). Numbers of tokens were chosen to ensure that they were multiples of both 4 (for sentence length to be even) and 6 (for the frequencies of words to be equated). There were a total of 12 participants in each condition.

Procedure. All procedures were identical to those in Experiment 1.

Results and Discussion

The results of Experiement 2 are shown in Figure 1, top right. As in Experiment 1, we analyzed the data via a multilevel logistic regression model. There was again no interaction of condition and word length, so the interaction term was again pruned from the model ($\chi^2(5) = 1.69$, $p = .89$). Coefficient estimates for the 48 and 100 conditions did not differ from chance ($\beta = .082$ and .39, $z = .31$ and 1.47, $p = .76$ and .14). In contrast, coefficients for the other four conditions did reach significance ($\beta = .75$, .76, 1.03, and 1.33, $z = 2.78$, 2.81 3.77, and 4.72, all p-values $< .01$). Performance rose steeply between 48 tokens and 300 tokens, then was largely comparable between 300 and 1200 tokens.

Experiment 3: Number of Word Types

The more words in a language, the harder the vocabulary of that language should be to remember. All things being equal, three words will be easier to remember than nine words. On the other hand, the more words in a language, the more diverse the evidence that you get. For a transition-finding model, this second fact is reflected in the decreased transition probabilities between words in a larger language, causing part-word distractors to have lower probability. For a chunking model, the same fact is reflected in the increase in complexity of viable alternative segmentations. For example, in a three-word language of the type described below, hypothesizing boundaries after the first syllable of each word rather than in the correct locations would result in a segmentation requiring six words rather than three—a relatively small increase in the size of the hypothesized language. In contrast, a comparable alternative segmentation for a nine-word language would contain 72 "words," which is a much larger increase over the true solution, and therefore much easier to rule out. Across models, larger languages result in an increase in the amount and diversity of evidence for the correct segmentation.

In our third experiment, we varied the number of distinct word types in the languages we asked participants to learn. If the added cost of remembering a larger lexicon is larger than the added benefit given by seeing a word in a greater diversity of contexts, participants should do better in segmenting smaller languages. If the opposite is true, we should expect participants to perform better in larger languages.

Methods

Participants. We tested 63 MIT students and members of the surrounding community. We excluded two participants from the final sample due to performance lower than two standard deviations below the mean performance for their condition.

Materials. Materials in Experiment 3 were identical to those in Experiments 1 and 2, with one exception. We fixed the number of words in each sentence at four and fixed the number of tokens of exposure at 600, but we varied the number of word types in the language, with 3, 4, 5, 6, and 9 types in the languages heard by participants in each of the five conditions. Numbers of types were chosen to provide even divisors of the number of tokens. Note that token frequency increases as the number of types decreases; thus in the 3 type condition, there were 200 tokens of each word, while in the 9 type condition, there were approximately 66 tokens of each word.

Procedure. All procedures were identical to those in Experiments 1 and 2.

Results and Discussion

Results of the experiment are shown in Figure 1, bottom. As predicted, performance decreased as the number of types increased (and correspondingly as the token frequency of each type decreased as well). As in Experiments 1 and 2, we analyzed the data via multi-level logistic regression. We again pruned the interaction of word length and condition from the model ($\chi^2(4) = 7.06$, $p = .13$). We found no significant effect of word length ($\beta = .031$, $z = .91$, $p = .36$), but we found significant coefficients for all but the 9 types condition ($\beta = 2.12$, 1.77, 1.17, .84, and .39, $z = 5.46$, 4.92, 3.17, 2.25, and 1.32, $p < .0001$ for 3 and 4 types and $p = .0015$, .025, and .19, for 5, 6, and 9 types, respectively). Thus, performance decreased as the number of types increased.

One potential concern about this experiment is that while we varied the number of types in the languages participants learned, this manipulation co-varied with the number of tokens in the language. To analyze whether the results of experiment were due to the number of tokens of each type rather than the number of types *per se*, we conducted an additional analysis. We consolidated the data from Experiments 2 and 3 and fit a multi-level model with two main predictors: number of types and number of tokens per type, as well as a binary term for which experiment a participant was in. Because of the relatively large number of levels and because the trend in Experiment 3 was roughly linear, we treated types and tokens per type as linear predictors, rather than as factors as in the previous analyses. (We experimented with adding an interaction term but found that it did not significantly increase model fit). We found that there was still a negative effect of number of types ($\beta = -.11$, $z = -2.81$, $p = .004$), even with a separate factor included in the model for the number of tokens per type ($\beta = 0.0049$, $z = 5.31$, $p < .0001$). Thus, although the type-token ratio does contribute to the effect we observed in Experiment 3, there is still an independent effect of number of types when we control for this factor.

Model Comparison

In this section, we compare the fit of a number of recent computational proposals for word segmentation to the human experimental results reported above. We do not attempt a comprehensive survey of models of segmentation.[1] Instead we sample broadly from the space of available models, focusing on those models whose fit or lack of fit to our results may prove theoretically interesting. We first present our materials and comparison scheme; we next give the details of our implementation of each model. Finally, we give results in modeling each of our experiments.

Because all of the models we evaluated were able to segment all experimental corpora correctly—that is, find the correct lexical items and prefer them to the distractor items—absolute performance was not useful in comparing models. In the terminology introduced above, all models passed the criterion of sufficiency for these simple languages. Instead, we compared models' fidelity: their fit to human performance.

Details of simulations

Materials. We compiled a corpus of twelve randomly generated training sets in each of the conditions for each experiment. These training sets were generated identically to those seen by participants and were meant to mimic the slight language-to-language variations found in our training corpora. In addition, because some of the models we evaluated rely on stochastic decisions, we wanted to ensure that models were evaluated on a range of different training corpora. Each training set was accompanied by 30 pairs of test items, the same number of test items as our participants received. Test items were (as in the human experiments) words in the generating lexicon of the training set or part-word distractors.

We chose syllables as the primary level of analysis for our models. Although other literature has dealt with issues of the appropriate grain of analysis for segmentation models (Newport & Aslin, 2004), in our experiments, all syllables had the same structure (consonant-vowel), so there was no difficulty in segmenting words into syllables. In addition, because we randomized the structure of the lexicon for each language, we chose to neglect syllable-level similarity (e.g., the greater similarity of *ka* to *ku* than to *go*). Thus, training materials consisted of strings of unique syllable-level identifiers that did not reflect either the CV structure of the syllable or any syllable-syllable phonetic similarities.

Evaluation. Our metric of evaluation was simple: each model was required to generate a score of some kind for each of the two forced-choice test items. We transformed these scores into probabilities by applying the Luce choice rule (Luce, 1963):

$$P(a) = \frac{S(a)}{S(a) + S(b)} \tag{1}$$

where a and b are test items and $S(a)$ denotes the score of a under the model. Having produced a choice probability for each test trial, we then averaged these probabilities across test trials to produce an average probability of choosing the correct item at test (which would be equivalent over repeated testing to the corresponding percent correct). We then averaged these model runs across training corpora to produce a set of average probabilities for each condition in each experiment.

Overall model performance was extremely high. Therefore, rather than comparing models to human data via their absolute deviation (via a measure like mean squared error), we chose to use a simple Pearson correlation coefficient[2] to evaluate similarities and differences in the shape of the curves produced when run on experimental data. By evaluating model performance in this way, our approach focuses exclusively on the relative differences between conditions rather than models' absolute fit to human performance, as in Orbán et al. (2008). Note that the use of a correlation rather than mean squared error is conservative: a model which fails to fit even the shape of human performance will fail even more dramatically when it is evaluated against the absolute level of human performance.

The Luce choice rule is not the only way to combine two scores into a single probability of success on a two-alternative forced-choice trial. In the following discussion of simulation results we will return to the issue of why we chose this particular evaluation metric.

[1] See e.g. Brent (1999b) and Brent (1999a) for a systematic explanation and comparison of models and Goldwater et al. (2009) for more results on recent probabilistic models.

[2] Pearson (parametric) correlations allow us to fit the shape of curves. We also ran Spearman (rank-order) correlations; the results are comparable, so we have neglected these values for simplicity of comparison.

Models

Transitional probability/Mutual information. As noted in the Introduction, one common approach to segmentation employs simple bigram statistics to measure the relationship between units. To model this approach, we began with the suggestion of Saffran, Newport, and Aslin (1996) to use transitional probability as a cue for finding word boundaries. We calculated transitional probability (TP) by creating unigram and bigram syllable counts over the training sentences in our corpus with a symbol appended to the beginning and end of each sentence to indicate a boundary. TP was defined with respect to these counts as

$$TP(s_{t-1}, s_t) = \frac{C(s_{t-1}, s_t)}{C(s_{t-1})} \tag{2}$$

where $C(s_{t-1})$ and $C(s_{t-1}, s_t)$ denote the count (frequency) of the syllable s_{t-1} and the string $s_{t-1}s_t$, respectively. We additionally investigated point-wise mutual information, a bidirectional statistic that captures the amount that an observer knows about one event given the observation of another:

$$MI(s_{t-1}, s_t) = \log_2 \frac{C(s_{t-1}, s_t)}{C(s_{t-1})C(s_t)} \tag{3}$$

Having computed transitional probability or mutual information across a corpus, however, there are many ways of converting this statistic into a score for an individual test item. We consider several of these proposals:

1. Local minimum: the lexicon of the language is created by segmenting the corpus at all local minima in the relevant statistic. Those test items appearing in the lexicon are assigned a score relative to their frequency in the corpus. Words not appearing in the lexicon are assigned a constant score (0 in our simulations).

2. Within-word minimum: words are assigned scores by the minimum value of the statistic for the syllable pairs in that word.

3. Within-word mean: words are assigned scores based on the mean of the relevant statistic.

4. Within-word product: words are assigned scores based on the product of the relevant statistic.

Only some of these options are viable methods for modeling variability in our corpus. For instance, the local minimum method predicts no differences in any of our experiments, since frequencies are always equated across items and all conditions have statistical minima between words. Therefore, we evaluated the within-word models: minimum, mean, and product. We found that within-word minimum and within-word product produced identical results (because we always compared words of the same length at test). Within-word mean produced slightly worse results. Therefore we report within-word product results in the simulations that follow.[3]

Both the TP and MI models explicitly took into account the boundaries between sentences. We implemented this feature by assuming that all sentences were bounded by a start/end symbol, #, and that this symbol was taken into account in the computation of transition counts. Thus, in computing counts for the sentence #*golabu*#, a count would be added for #*go* as well as for *gola*. This decision was crucial in allowing these models to have defined counts for transitions in Experiment 1's sentence length 1 condition (otherwise, no word-final syllable would ever have been observed transitioning to any other syllable).

Bayesian Lexical Model. We evaluated the Lexical Model described in Goldwater, Griffiths, and Johnson (2006a); Goldwater et al. (2009). We made use of the simple unigram version (without word-level contextual dependencies) since the experimental corpora we studied were designed to incorporate no contextual dependencies. The model uses Bayesian inference to identify hypotheses (sequences of word tokens) that have high probability given the observed data d. For any particular hypothesis h, the posterior probability $P(h|d)$ can be decomposed using Bayes' Rule, which states that $P(h|d) \propto P(d|h)P(h)$. Therefore, optimizing $P(h|d)$ is equivalent to optimizing the product of the likelihood $P(d|h)$ (which tells us how well the data is explained by the hypothesis in question) and the prior $P(h)$ (which tells us how reasonable the hypothesis is, regardless of any data). In this model, the likelihood term is always either 0 or 1 because every sequence of word tokens is either entirely consistent or entirely inconsistent with the input data. For example, if the input is *golabupadoti*, then hypotheses *golabu padoti* and *go la bu pa doti* are consistent with the input, but *lookat that* is not. Consequently, the model need only consider consistent segmentations of the input, and of these, the one with the highest prior probability is the optimal hypothesis.

The prior probability of each hypothesis is computed by assuming that words in the sequence are generated from a distribution known as a Dirichlet process. A Dirichlet process is a probabilistic process which generates Dirichlet distributions—discrete distributions over sets of counts—and these Dirichlet distributions are used in the lexical model to parameterize the distribution of word frequencies. The Dirichlet process gives higher probabilities to more concentrated Dirichlet distributions, corresponding to small lexicons with high frequency words. Individual words in the lexicon are then generated according to an exponential distribution, which gives higher probabilities to shorter words.

The probability of the entire sequence of words in the sequence can be found by multiplying together the probability of each word given the previous words in the sequence. The probability of the ith word is given by

$$P(w_i = w|w_1 \ldots w_{i-1}) = \frac{n_{i-1}(w) + \alpha P_0(w)}{i - 1 + \alpha} \tag{4}$$

where $n_{i-1}(w)$ is the number of times w has occurred in the previous $i - 1$ words, α is a parameter of the model, and P_0

[3] Within-word product also has the advantage of being equivalent to word production probability in a Markov model, a fact which makes it an appropriate choice of measure for comparison with the Lexical model in later simulations.

is a distribution specifying the probability that a novel word will consist of the phonemes $x_1 \ldots x_m$:

$$P_0(w = x_1 \ldots x_m) = \prod_{j=1}^{m} P(x_j) \quad (5)$$

According to these definitions, a word will be more probable if it has occurred many times already ($n_{i-1}(w)$ is high), but there is always a small chance of generating a novel word. The relative probability of a novel word decreases as the total number of word tokens ($i - 1$) increases, and novel words are more probable if they are shorter (contain fewer phonemes). The overall effect is a set of soft constraints that can be viewed as constraints on the lexicon: the model prefers segmentations that result in lexicons containing a small number of items, each of which is relatively short, and many of which occur very frequently.

The definition given above provides a way to evaluate the probability of any given hypothesis; in order to actually find high-probability segmentations, Goldwater, Griffiths, and Johnson (2009) use a *Gibbs sampler*, a type of algorithm that produces samples from the posterior distribution. The sampler works by randomly initializing all potential word boundary locations (i.e., all syllable boundaries) to either contain a word boundary or not. It then iteratively modifies the segmentation by choosing whether or not to change its previous decision about each potential word boundary. Each choice is random, but is influenced by the underlying probability model, so that choices that increase the probability of the resulting segmentation are more likely. After many iterations through the data set, this algorithm is guaranteed to converge to producing samples from the posterior distribution. Note that, since the goal of Goldwater, Griffiths, and Johnson (2009) was an ideal observer analysis of word segmentation, their algorithm is designed to correctly identify high-probability segmentations, but does not necessarily reflect the processes that might allow this in humans. In particular, it is a batch algorithm, requiring all input data to be stored in memory at once. We return to this point later. For more information on Gibbs sampling, see Gelman, Carlin, Stern, and Rubin (2004) and MacKay (2003).

Using the Bayesian lexical model described above, we defined the score for a particular word at test to be the posterior probability of the word, estimated by summing over a large number of samples from the posterior. Because the posterior probability of the correct solution was normally so high (indicating a high degree of confidence in the solution the model found), we ran the Gibbs sampler using a range of temperatures to encourage the model to consider alternate solutions.[4] Although this manipulation was necessary for us to be able to evaluate the posterior probability of distractor items, it had relatively little effect on the results across a large range of temperatures (2 – 20). We therefore report results from temperature = 2. The model had one further parameter: the α parameter of the Dirichlet process, which we kept constant at the value used in Goldwater et al. (2009).

In the language of the introduction, the Lexical model is a chunking model, rather than a transition-finding model: though its inference algorithm works over boundary positions, its hypotheses are sequences of chunks (words) and the measures with which it evaluates hypotheses are stated in terms of these chunks (how many unique chunks there are, their frequencies, and their lengths).

MI Clustering. We evaluated the mutual information-based clustering model described in Swingley (2005). This model is a clustering model which calculates n-gram statistics and pointwise mutual information over a corpus, then takes as words those strings which exceed a certain threshold value both in their frequency and in the mutual information of their constituent bisyllables. Unlike our versions of the TP and MI models, the MI Clustering model looks for coherent chunks which satisfy its criteria of frequency and mutual information (and it evaluates these criteria for every possible chunk in the exposure corpus). Thus, like the Lexical model, it is a chunking model rather than a transition-finding model.

In order to run the model on the language of our experiment, we added support for four-syllable words by analogy to three-syllable words. We then defined the score of a string under the model (given some input corpus) as the maximum threshold value at which that string appeared in the lexicon found by the model. In other words, the highest-scoring strings were those that had the highest percentile rank both in mutual information and in frequency.

PARSER. We implemented the PARSER model described in Perruchet and Vinter (1998) and Perruchet and Vinter (2002).[5] PARSER is organized around a lexicon, a set of words and scores for each word. The model receives input sentences, parses them according to the current lexicon, and then adds sequences to the lexicon at random from the parsed input. Each lexical item decays at a constant rate and similar items interfere with each other. The model as described has six parameters: the maximum length of an added sequence, the weight threshold for a word being used to parse new sequences, the forgetting and interference rates, the gain in weight for reactivation, and the initial weight of new words. Because of the large number of parameters in this model, it was not possible to complete an exhaustive search of the parameter space; however, we experimented with a variety of different combinations of interference and forgetting rates and maximum sequence lengths without finding any major differences in forced-choice performance. We therefore report results using the same parameter settings used in the initial paper.

We made one minor modification to the model to allow it to run on our data: our implementation of the model iterated through each sentence until reaching the end and then began

[4] Temperature is a parameter which controls the degree to which the Gibbs sampler prefers more probable lexicons, with higher temperature indicating greater willingness to consider lower-probability lexicons. See Kirkpatrick, Gelatt, and Vecchi (1983) for more details.

[5] We also thank Pierre Perruchet for providing a reference implementation of PARSER, whose results were identical to those we report here.

anew at the beginning of the next sentence—thus, it could not add sentence-spanning chunks to its lexicon.

Comparison results

Figure 2 shows the performance of each model plotted side-by-side with the human performance curves shown in Figure 1. For convenience we have adjusted all the data from each model via a single transformation to match their scale and intercept to the mean human performance. This adjustment does not affect the Pearson *r* values given in each subplot, which are our primary method of comparison. We discuss the model results for each experiment in turn, then end by considering effects of word length on model performance.

Experiment 1. Most models replicated the basic pattern of performance in Experiment 1, making fewer errors (or assigning less probability to distractors) when words were presented alone or in shorter sentences. However, there were differences in how well the models fit the quantitative trend.

Both the TP and MI models showed a relatively large decrease in performance from single-word sentences to two-word sentences. In the TP model, this transition is caused by the large difference between distractors which have a score of 0 (in the single-word sentences) and distractors which have a score of $1/10$ (in the two word sentences—since every other word is followed by one of 5 different words). However, the relative difference in probabilities between sentences with 12 words and sentences with 24 words was very small, as reflected in the flatness of both the TP and MI curves. This trend did not exactly match the pattern in the human data, where performance continued to decrease from 12 to 24.

Compared to the TP and MI models, the Lexical Model showed a slightly better fit to the human data, with performance decreasing more gradually as sentence length increased. The Lexical Model's increased difficulty with longer sentence lengths can be explained as follows. The model is designed to assign a probability to every possible segmentation of a given input string. Although the probabilities assigned to incorrect segmentations are extremely low, longer sentences have a much larger number of possible segmentations. Thus, the total probability of all incorrect segmentations tends to increase as sentence length increases, and the probability of the correct segmentation must drop as a result. Essentially, the effects of sentence length are modeled by assuming competition between correct and incorrect segmentations.

The MI Clustering model in this (and both other) experiments produced an extremely jagged curve, leading to a very low correlation value. The reason for this jaggedness was simple: since the model relies on percentile rankings of frequency rather than raw frequencies, its performance varied widely with very small changes in frequency (e.g., those introduced by the slight differences between input corpora in our studies). Because the model is deterministic, this noise could not be averaged out by multiple runs through the input corpus.

PARSER failed to produce the human pattern of results. In any given run, PARSER assigned the target a non-zero (often high) score and the distractor a score of 0 or very close to zero, producing a flat response curve across conditions. In order to verify that this was not an artifact of the relatively small number of simulations we performed, we ran a second set of simulations with 100 independent PARSER runs for each training set. The results of these simulations were qualitatively very similar across all three experiments despite the integration of 1200 datapoints for each point on each curve. We return to the issue of PARSER's performance in the discussion on target and distractor probabilities.

Experiment 2. The TP and MI models, which performed relatively well in Experiment 1, failed to produce the pattern of gradual learning in Experiment 2. This result stems from the fact that both models produce *point estimates* of descriptive statistics. These point estimates require very little data to converge to their eventual values. Regardless of whether the models observe 300 or 1200 sentences, the transitional probability they find between any two words will be very close to $1/5$ (since any word can be followed by any other except itself). This same fact also explains the relatively flat performance of the MI Clustering model, though this model's high performance with very little input reflects the likelihood of not having observed the distractor items even once at the beginning of exposure.

In contrast, both the Lexical Model and PARSER succeeded in fitting the basic pattern of performance in this experiment. Although PARSER is an online model which walks through the data in a single pass and the Lexical Model is a batch model which processes all the available data at once, both models incorporate some notion that more data provides more support for a particular hypothesis. In the Lexical Model, which evaluates hypothesized lexicons by their parsimony relative to the length of the corpus that is generated from them, the more data the model observes, the more peaked its posterior probability distribution is. In PARSER, the more examples of a particular word the model sees in the exposure corpus, the larger its score in the lexicon (and the longer it will take to forget). Another way of stating this result: more examples of any concept make that concept faster to process and easier to remember.

Experiment 3. The results of the model comparison on Experiment 3 were striking. No model succeeded in capturing the human pattern of performance in this experiment. MI Clustering and PARSER were uncorrelated with human performance. All three of the other models showed a clear trend in the opposite direction of the pattern shown by the human participants. These models performed better on the languages with larger numbers of word types for the same reason: languages with a larger number of word types had distractors that were less statistically coherent. For example, a language with only three types has a between-word TP of $1/2$ while a language with nine types has a far lower between-word TP of $1/8$, leading to part-words with very low within-word TP scores.

Put another way, the advantage that human participants gained from the clearer statistics of the larger language did

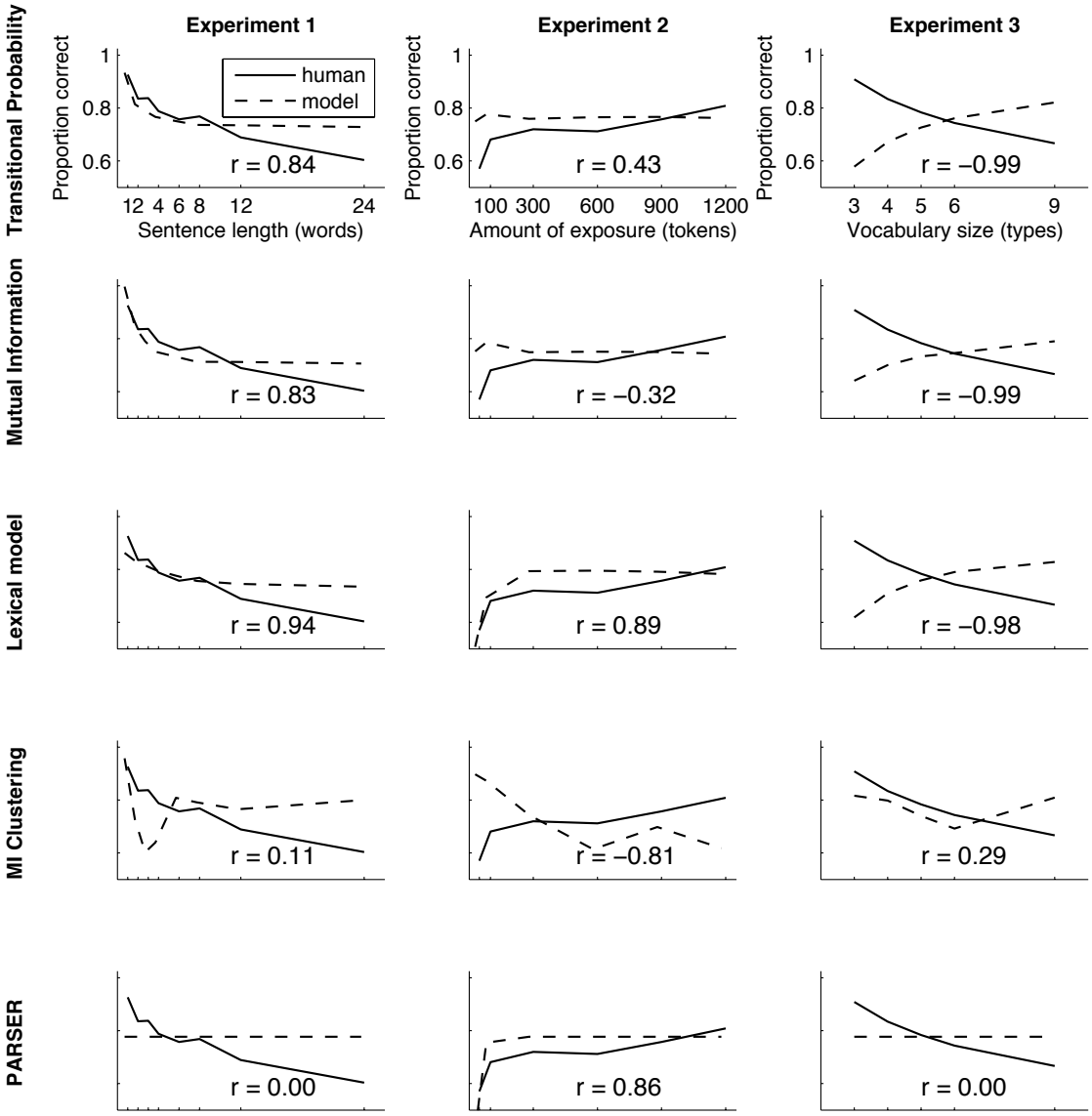

Figure 2. Comparison of model results to human data for Experiments 1 – 3. All results from each model are adjusted to the same scale and intercept as human data to facilitate comparison (this manipulation does not affect the Pearson correlation coefficient given at the bottom of each subplot). Models are slightly offset in the horizontal axis to avoid overplotting.

not outweigh the disadvantage of having to remember more words and having to learn them from fewer exposures. The TP, MI, and Lexical Models, in contrast, all weighted the clearer statistics of the language far more heavily than the decrease in the number of tokens for each word type.

Probabilities of targets and distractors

Models differed widely in whether conditions which produced lower performance did so because targets were assigned lower scores, or because distractors were assigned higher scores. This section discusses these differences in the context of previous work and our decision to use a choice rule which includes both target and distractor probabilities.

Figure 3 shows the relative probabilities of targets and distractors for each experiment and model. We have chosen not to give vertical axis units since the scale of scores for each model varies so widely; instead, this analysis illustrates the qualitative differences between models.

The transitional probability model produces different patterns of performance only because the probabilities of distractors vary from condition to condition. As noted previously, in the languages we considered, transitions within words are always 1; therefore overall performance depends on the TP at word boundaries. In contrast, because mutual

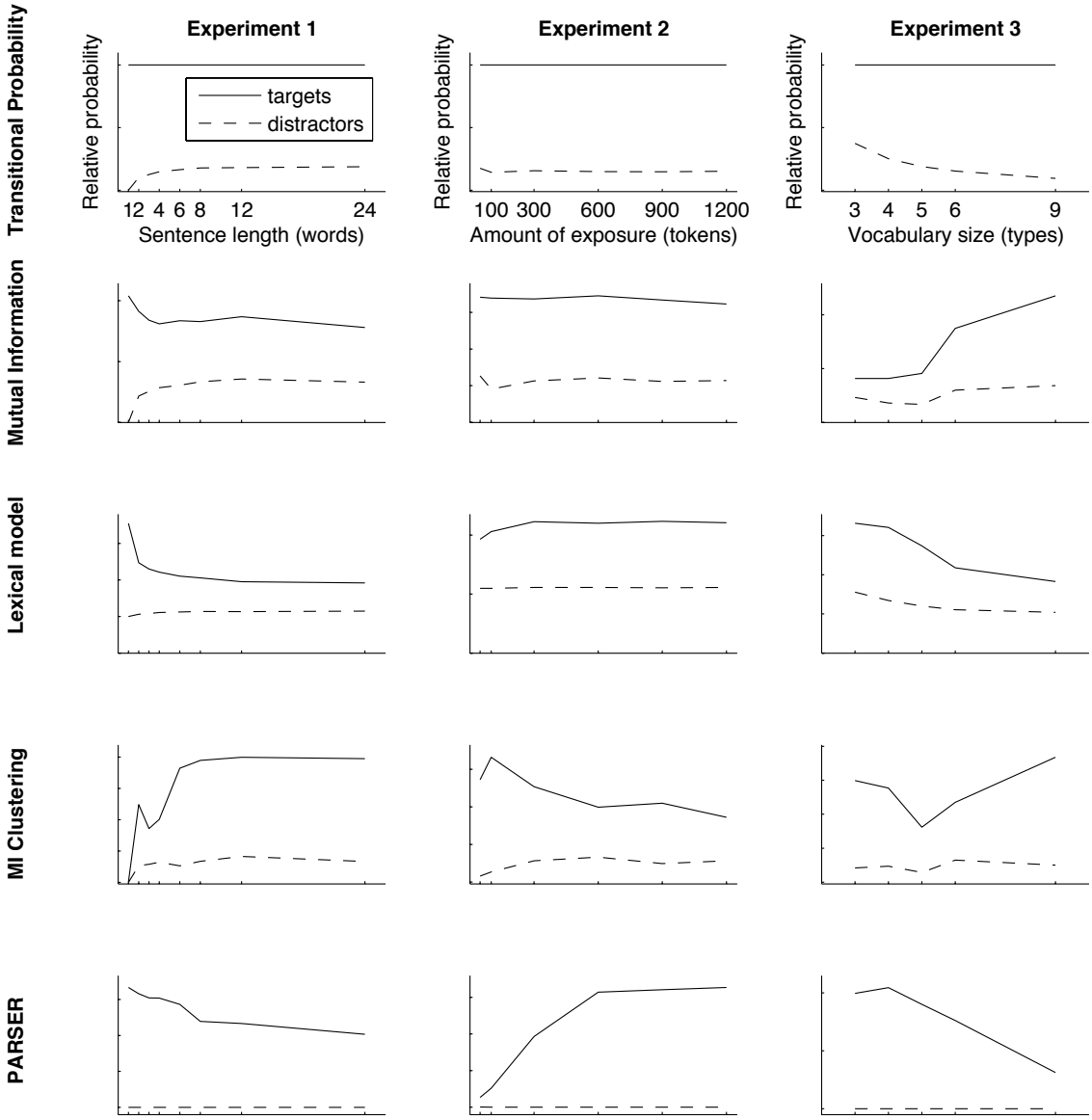

Figure 3. Relative probabilities of target and distractor items for the models we evaluated. Vertical axis is scaled separately for each row. To illustrate trends in the Lexical model, results are plotted from temperature 20 (though results were qualitatively similar across all temperatures).

information is normalized bidirectionally (and hence takes into account more global properties of the language), target MI changed as well as distractor MI. Nonetheless, the Luce choice probabilties for MI and TP were quite similar, suggesting that the differences in the MI model were not substantive.

Considering target and distractor probabilities for the Lexical model we see several interesting features. First, effects in Experiments 1 and 2 seem largely (though not entirely) driven by target probability. Although there is some increase in distractor probability in Experiment 1, targets become considerably less probable as sentence lengths increase. Likewise in Experiment 2, distractor probabilities remain essentially constant, but the greater amount of exposure to target words increases target probabilities. Experiment 3 shows a different pattern, however: target probabilities match human performance, decreasing as word lengths increase. But distractor probabilities decrease more slowly than target probabilities, canceling this effect and leading to the overall reversal seen in Figure 3. Put another way: it is not that targets increase in probability as languages grow in vocabulary size. Instead, the effects we observed in the Lexical model in Experiment 3 are due to the probabilities of distractors relative to targets. (Correlations between Lexi-

cal Model target probabilities and human performance were r = .82, .89, and .96, respectively).

The MI Clustering model showed patterns of target scores that were the reverse of human performance for all three experiments, though in Experiment 1 changes in distractor score were large enough to reverse this trend. For the other two experiments we saw only limited effects of distractor probability.

Target and distractor probabilities for PARSER were revealing. Target probabilities for all three experiments followed the same qualitative pattern as human performance: decreasing in Experiments 1 and 3 and increasing in Experiment 2. Thus, it was purely the fact that PARSER assigned no score to distractors that prevented it from capturing general trends in human performance. Were correlations assessed with target scores alone, PARSER would correlate at r = .92, .84, and .89 with human performance across the three experiments, respectively (comparable to the level of the Lexical Model target probabilities and to the resource-limited probabilistic models discussed in the next section).

Overall, this analysis suggests that the Luce choice rule we used (and that it weighted target and distractor probabilities equivalently) led to the patterns we observed in the comparison with human performance. This result begs the question of why we chose the Luce choice rule in particular.

The key argument for considering distractor probability in evaluating model performance is given by the experimental data reported in Saffran, Newport, and Aslin (1996). In Experiment 1, Saffran et al. trained two groups of human participants on the same exposure corpus but tested them on two different sets of materials. The targets in each set were the legal words of the corpus that the participants had been exposed to, but one group heard non-word distractors—concatenations of syllables from the language that had not been seen together in the exposure corpus—while the other group heard part-word distractors that (comparable to our distractors) included a two-syllable substring from a legal word in the corpus. Performance differed significantly between these two conditions, with participants rejecting non-words more than part-words. These results strongly suggest that human learners are able to assign probabilities to distractors and that distractor probabilities matter to the accuracy of human judgments at test.

Distractor probabilities could be represented in a number of ways that would be consistent with the current empirical data. For example, human learners could keep track of a discrete lexicon like the one learned by PARSER, but simply include more possible strings in it than those represented by PARSER in our simulations. On this kind of account, the difficulty learners had in the part-word condition of Saffran et al.'s experiment would be caused by confusion over the part words (since non-words would not be represented at all). This kind of story would still have to account for the difficulty of the non-word condition, though, since most participants were still not at ceiling. On the other hand, a TP-style proposal (much like the probabilistic DM-TP model proposed in the next section) would suggest that participants could evaluate the relative probabilities of any string in the language they heard. Current empirical data do not distinguish between these alternatives but they do strongly suggest that human learners represent distractor probabilities in some form.

Discussion

Under the evaluation scheme we used, no model was able to fit even the relative pattern of results in all three experiments. In particular, no model produced a similar trend to human data in Experiment 3, and many failed to in Experiment 2 as well. Although some models assigned relative probabilities to target items that matched human performance, when distractor probabilities were considered, model performance diverged sharply from humans.

We speculate that the match and mismatch between models and data is due to the failure of this first set of models to incorporate any notion of resource limitations. Human learners are limited in their memory for the data they are exposed to—they cannot store hundreds of training sentences—and for what they have learned—a large lexicon is more difficult to remember than a smaller one. This simple idea accounts for the results of both Experiment 2 and Experiment 3. In Experiment 2, if participants are forgetting much of what they hear, hearing more examples will lead to increased performance. In Experiment 3, although larger languages had clearer transition statistics, they also had more words to remember. The next section considers modifications to two probabilistic models (the Lexical Model and a modified version of the TP model) to address human memory limitations.

Adding Resource Constraints to Probabilistic Models

Memory limitations provide a possible explanation for the failure of many models to fit human data. To test this hypothesis, the last section of the paper investigates the issue of adding memory limitations to models of segmentation. We explore two methods. The first, *evidence limitation*, implements memory limitations as a reduction in the amount of the evidence available to learners. The second, *capacity limitation*, implements memory limitations explicitly via imposing limits on models' internal states.

For this next set of simulations, we narrow the field of models we consider, looking only at *probabilistic generative models*. These models are "Bayesian models" because they are often stated in terms of a decomposition into a prior over some hypothesis space and a likelihood over data given hypotheses, allowing the use of a family of Bayesian inference techniques for finding the posterior distribution of hypotheses given some observed data. We choose this modeling framework because it provides a common vocabulary and set of tools for stating models with different representations and performing inference in these models; hence modeling insights from one model can easily be applied to another model. (The results of Orbán et al., 2008, provide one example of the value of comparing models posed in the same framework).

The only probabilistic generative model in our initial comparison was the Lexical Model. However, standard transitional probability models are closely related to probabilistic generative models. Therefore, before beginning this investigation we rewrite the standard transitional probability model, modifying it so that it is a standard Bayesian model which can be decomposed into a prior probability distribution and a likelihood of the data given the model. We refer to this new model as the DM-TP (Dirichlet-multinomial TP) model because it uses a Dirichlet prior distribution and multinomial likelihood function.

Modifications to the TP model

As we defined it above, the transitional probability model includes no notion of strength of evidence. It will make the same estimate of TP whether it has observed a given set of transitions once or 100 times. This property comes from the fact that Equation 2 is a maximum-likelihood estimator, a formula which gives the highest probability estimate for a particular quantity regardless of the confidence of that estimate. In contrast, a Bayesian estimate of a particular quantity interpolates between the likelihood of a particular value given the data and the prior probability of that value. With very little data, the Bayesian estimate is very close to the prior. In the presence of more data, the Bayesian estimate asymptotically approaches the maximum likelihood estimate. In this section we describe a simple Bayesian version of the TP model which allows it to incorporate some notion of evidence strength.

In order to motivate the Bayesian TP model we propose, we briefly describe the equivalence of the TP model to simple Markov models that are commonly used in computational linguistics (Manning & Schütze, 1999; Jurafsky & Martin, 2008). A Markov model is simply a model which makes a Markov, or independence, assumption: the current observation depends only on the n previous observations and is independent of all others. In this way of viewing the model, each syllable is a state in a finite-state machine: a machine composed of set of states which emit characters, combined with transitions between these states. In such a model, the probabilities of a transition from each state to each other state must be learned from the data. The maximum-likelihood estimate of the transition probabilities from this model is the same as we gave previously in Equation 2: for each state we count the number of transitions to other states and then normalize.

One weakness of Markov models is their large number of parameters, which must be estimated from data. The number of parameters in a standard Markov model is exponential, growing as m^n, where m is the number of states (distinct syllables) and n is the "depth" of the model (the number of states that are taken into account when computing transition probabilities—here, $n = 2$ because we consider pairs of syllables). Because of this problem, a large amount of the literature on Markov models has dealt with the issue of how to estimate these parameters effectively in the absence of the necessary quantity of data (the "data sparsity" problem). This process is often referred to as "smoothing" (Chen & Goodman, 1999).

After observing only a single transition (say in the sequence *gola*) a standard maximum-likelihood TP model calculates that the transition from *go* to *la* happens with probability 1 and that there is no other syllable in future exposure that will ever follow *go*. This tendency to jump to conclusions has the consequence of severely limiting the ability of unsmoothed Markov models to generalize from limited amounts of data. Instead, they tend to overfit whatever data they are presented with, implicitly assuming that the evidence that has been presented is perfectly representative of future data. (This is precisely the phenomenon that we saw in the failure of the TP model in Experiment 2: even with a small amount of evidence, the model overfit the data, learning the transition probabilities perfectly.)

One simple way to deal with this issue of unseen future data is to estimate transitions using Bayesian inference. Making Bayesian inferences consists of assuming some prior beliefs about the transition structure of the data—in this case, the conservative belief that transitions are uniform—that are gradually modified with respect to observed data. To do this, we assume that transitions are samples from a multinomial distribution. Rather than estimating the maximum-likelihood value for this distribution from counts (as in Equation 2), we add a prior distribution over possible values of $P(s_{t-1}, s_t)$. The form of this prior is a Dirichlet distribution with parameter α. Because of the conjugacy of the Dirichlet and multinomial distributions (Gelman et al., 2004), we can express the new estimate of transition probability simply by adding a set of "pseudo-counts" of magnitude α to each estimate:

$$P(s_{t-1}, s_t) = \frac{C(s_{t-1}, s_t) + \alpha}{\sum_{s' \in S}(C(s_{t-1}, s') + \alpha)}. \quad (6)$$

Under this formulation, even when a particular syllable s has not been observed, its transitional probability is not zero. Instead, it has some smoothed base value that is determined by the value of α.[6]

Note that as α becomes small, the DM-TP model reduces to the TP model that we already evaluated above. The Appendix gives results on the DM-TP model's fit to all three experiments over a wide range of values of α. In the next sections, however, we investigate proposals for imposing memory limitations on the DM-TP and Lexical models.

Modeling memory effects by evidence limitation

One crude way of limiting models' memory is never to provide data to be remembered in the first place. Thus, the first and most basic modification we introduced to the Lexical and TP models was simply to limit the amount of evidence available to the models.

[6] Adding this prior distribution to a TP model creates what is known in computational linguistics as a smoothed model, equivalent to the simple and common "add-delta" smoothing described in Chen and Goodman (1999)'s study of smoothing techniques. For more detail on the relationship between Dirichlet distributions and smoothing techniques see MacKay and Peto (1994), Goldwater, Griffiths, and Johnson (2006b), and Teh (2006).

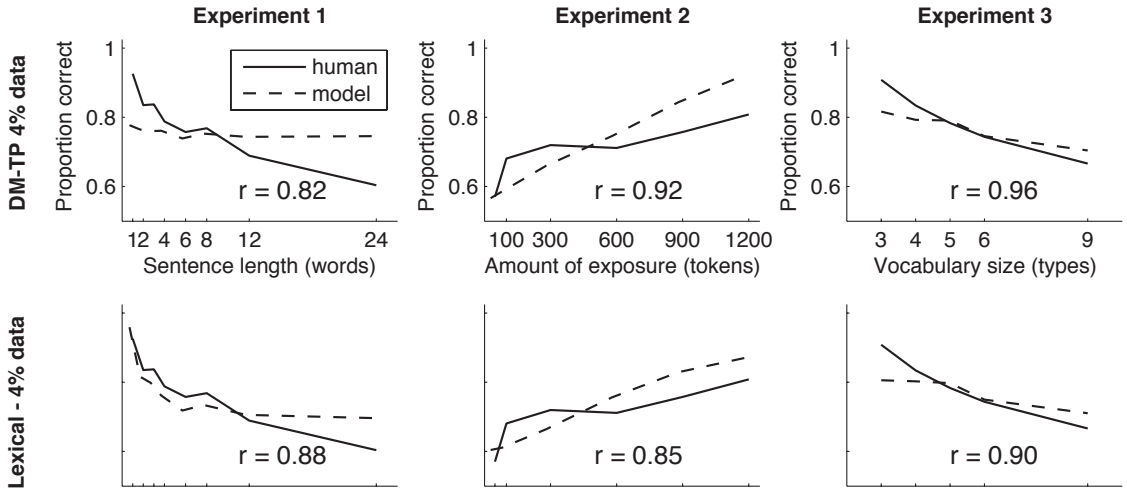

Figure 4. Comparison of Bayesian transitional probability model and Lexical Model, both trained with 4% of the original dataset, to human data. Model results are offset, scaled, and adjusted to the same intercept as human data to facilitate comparison, as in Figure 2.

Methods. We conducted these simulations by running both the DM-TP and the Lexical Model on a new set of experimental materials. These materials were generated identically to the experimental materials in the previous section except that we presented the models with only 4% of the original quantity of data. (We chose 4% to be the smallest amount of data that would minimize rounding errors in the number of sentences in each exposure corpus). For example, in Experiment 1, models were presented with 24 (rather than 600) word tokens, distributed in 24 1-word sentences, 12 2-word sentences, 8 3-word sentences, etc. The number of tokens was reduced similarly in Experiments 2 and 3 while holding all other details of the simulations (including the number of types) constant from the original model comparison.

Results. We evaluated performance across a range of values of α for the DM-TP model and across a range of temperatures for the Lexical Model. As before, although there was some variability between simulations, all temperatures of 2 and above produced substantially similar results. In contrast, results on Experiment 3 (but not the other two experiments) varied considerably with different values of α. We return to this issue below.

Results for $\alpha = 32$ and temperature 3 are plotted in Figure 4. For these parameter settings, performance in the two models was very similar across all three experiments ($r = .90$), and both models showed a relatively good fit to the data from all three conditions. In the Appendix, we further explore the α parameter and its role in fitting human data. In Experiment 1, both models captured the general trend, although the DM-TP model showed this pattern only within an extremely compressed range. In Experiment 2, both models captured the basic direction of the trend but not its asymptotic, decelerating shape. In Experiment 3, in contrast to the results of our first set of simulations, both models captured the trend of decreasing performance with increasing vocabulary size (although they did not exactly match the decrease in performance between languages with 3 and 4 words).

Discussion. Why did limiting the evidence available to the models lead to success in fitting the decreasing trend in human performance in Experiment 3? Performance in Experiment 3 for both models comes from a tradeoff between two factors. The first factor is the decreasing statistical coherence of distractors as the number of types in the language increases. With 3 types, distractors (part-words) contain at most one transition with probability 1/2; with 9 types in contrast, distractors contain a transition with probability 1/8. The second factor at work is the decreasing number of tokens of each type in conditions with more types. A smaller number of tokens means less evidence about any particular token.

In the original simulations, the first factor (statistical coherence) dominated the second (type/token ratio) for both the Lexical Model and the TP model. With its perfect memory, the Lexical Model had more than enough evidence to learn any particular type; thus the coherence of the distractors was a more important factor. Likewise, the point estimates of transition probability in the TP model were highly accurate given the large number of tokens. In contrast, in the reduced-data simulations, the balance between these two factors was different. For instance, in the Lexical Model, while a larger vocabulary still lead to lower coherence within the distractors, this factor was counterbalanced by the greater uncertainty about the targets (because of the small number of exposures to any given target). In the DM-TP model, the same principle was at work. Because of the prior expectation of uniform transitions, conditions with more types had a small enough number of tokens to add uncertainty to the estimates of target probability. Thus, for both Bayesian models, less

data led to greater uncertainty about the targets relative to the distractors.

For both models, learning with limited evidence in Experiment 2 was much more gradual than for the human participants with the full amount of data. (In contrast, the Lexical Model with the full human dataset fit human performance in this experiment quite well). This mismatch points to a limitation of the evidence limitation method. While only a severe reduction in the amount of data led to a good fit in Experiment 3, a much smaller reduction in data led to a better fit to Experiment 2 (see Table 1 in the Appendix). The intuition behind this mismatch is that human participants are limited in their memory for what they extract from the data, not for the data themselves. That is, they learn rapidly and asymptotically from a relatively small amount of data, but they are sharply limited in the size of the vocabulary that they can learn. This mismatch suggests that future proposals should put limits on the model's internal storage capacity for what is learned, not the data that the models are presented with in the learning situation. We explore this approach in the next section.

Modeling memory effects by capacity limitation

In our final section, we make use of recent innovations in probabilistic inference to impose limits on the internal memory capacity of the Lexical Model. In other words, rather than limiting the amount of data it has access to, we limit its ability to remember what it learns from that data.

We perform this set of simulations with the Lexical Model only. In the DM-TP model, limiting or forgetting model-internal state (transition counts) is equivalent to exposing the model to less data, which is also equivalent to increasing the strength of the smoothing parameter (see Appendix for more details), regardless of whether forgetting is implemented randomly over the entire dataset or sequentially, since the distribution of transitions is uniform across the corpus. Thus, the results of randomly forgetting particular transition counts would be the same (in the limit) as those presented in Figure 4 and Table 1.

Materials and Methods. In order to create a memory-limited version of the lexical model, we modified the original inference scheme for the Lexical Model. Rather than making use of a Gibbs sampler—a form of batch inference which operates over the entire dataset at once—we instead implemented a *particle filter* (Doucet, De Freitas, & Gordon, 2001). A particle filter is a sequential Monte-Carlo technique which represents multiple hypotheses (particles) but updates them by moving sequentially through the data, one sentence at a time. Like Gibbs sampling and other Markov-chain monte-carlo (MCMC) techniques, particle filters are guaranteed in the limit to converge to the posterior distribution over hypotheses. While standard MCMC schemes store the entire dataset but consider hypotheses one by one, particle filters store many different hypotheses but consider individual data points. Thus, particle filters represent a promising possibility for a more realistic style of inference in probabilistic models (Sanborn, Griffiths, & Navarro, 2006; Daw & Courville, 2008; Brown & Steyvers, 2009; Levy, Reali, & Griffiths, 2009; Vul, Frank, Alvarez, & Tenenbaum, 2009), in which learners are not assumed to have all training data accessible at any given moment.

Nevertheless, a standard particle filter still assumes that there is no capacity limitation on generalizations learned from data. While the particle filter is an "online" rather than "batch" form of inference, it is still expected to produce an estimate of the same posterior distribution as the Gibbs sample that Goldwater et al. originally used (and hence provide the same fit to human data as the simulations with the Gibbs sampler). Thus, to match human performance better we implemented a number of forgetting schemes on top of this particle filter. Each approximated the standard assumption about human learners: that they operate with memory restrictions on both their memory for the input data *and* for the generalizations they draw from that input data. Note that these modifications eliminate the guarantee of optimality that comes with using the Bayesian framework. While the Gibbs sampler and particle filter are both expected to converge to the highest probability lexicon under the model (as either the number of iterations of the sampler or the number of particles go to infinity), these memory-limited versions of the particle filter are not necessarily even asymptotically optimal.

We conducted simulations using the full dataset as input to three variations on the Lexical Model. Each variation implemented a simple forgetting rule on the lexical hypothesis maintained by the particle filter. The three models we evaluated were:

1. *Uniform forgetting of word types*: If the number of types in the lexicon exceeds n delete a type uniformly at random.

2. *Frequency-proportional forgetting of word types*: If the number of types in the lexicon exceeds n delete type w with probability proportional to $1/P(w)$ where $P(w)$ is the predictive probability of w under the current lexicon.

3. *Uniform forgetting of word tokens*: If the number of tokens in the lexicon exceeds n delete a token uniformly at random.

Each simulation was run with only a single particle, meaning that each run of each model produced exactly one guess about the contents of the lexicon (Sanborn et al., 2006). Because of the large amount of noise present in each simulation, we ran 10 simulated "participants" for each different set of input materials (meaning that performance for each condition is estimated from 120 independent simulations). As before we systematically varied temperature, and we also varied the value of n for each model.

Results and Discussion. Results are plotted in Figure 5. For each model, we have plotted the temperature and value of n that best fit the human data (note that there is no reason that the values of n should be comparable across different models). All three models produced a good fit to human performance across all three experiments for some parameter regime. In particular, all three models produced the human pattern of results in Experiment 3: limiting the capacity of the models made languages with larger numbers of types

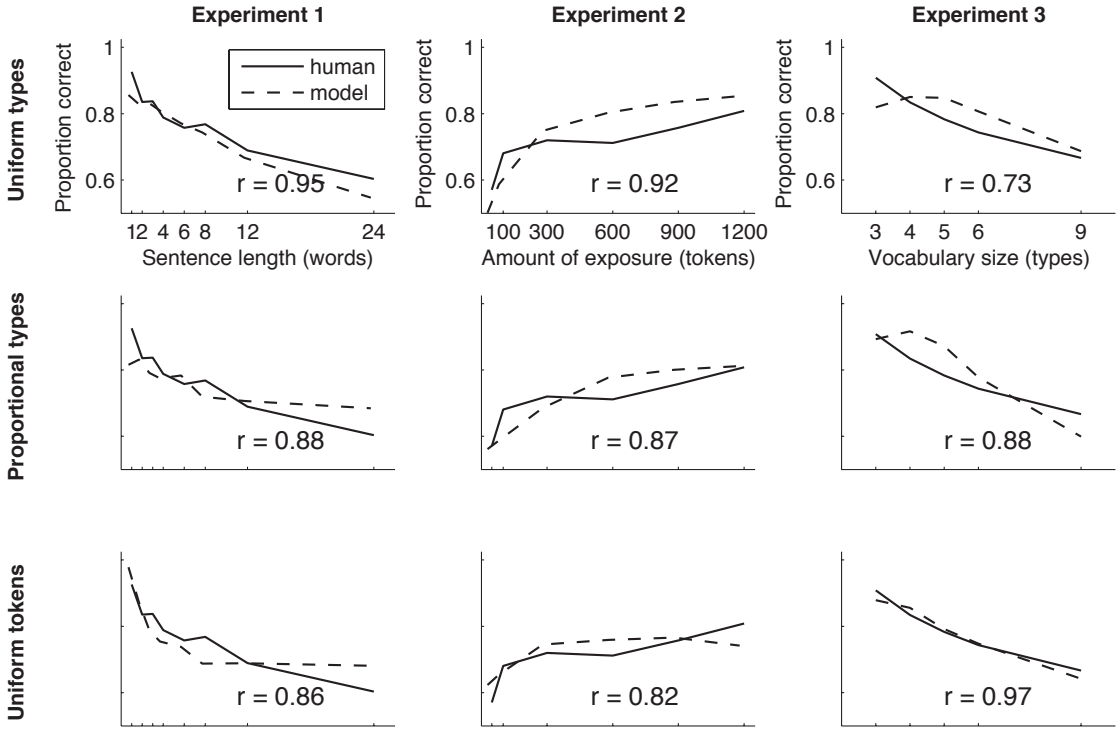

Figure 5. Comparison of three capacity-limited variations on the Lexical Model (see text) to human data. Model results are offset, scaled, and adjusted to the same intercept as human data to facilitate comparison.

more difficult to learn than languages with smaller numbers of types.

Several features of the results from these capacity-limited models stand out as preferable to the results from the evidence-limitation method and DM-TP model. First, all three models produced curves with roughly the same asymptotic shape as the human data in Experiment 2. In contrast, both evidence-limited models showed a more linear trend. Second, it is suggestive that the DM-TP model did not appear to show the same scale of effect across the three experiments. For example, the scale of the effect in Experiment 1 for the less-data simulations was clearly too small relative to the scale of the other effects, while the scale in Experiment 2 was too large. In contrast, the scale of the models in the current set of simulations matches more closely across the three experiments.

General Discussion

We presented results from three adult artificial language segmentation experiments. In each of these experiments we varied one basic aspect of the composition of the language that participants learned while holding all others constant, producing a set of three average performance curves across a wide variety of input conditions. The results in all three conditions were intuitive: longer sentences, less data, and a larger number of words all made languages harder to segment. However, a variety of models sampled from the computational literature on segmentation all failed to fit the basic qualitative trends in one or several experiments. In particular, all models were unable to account for the result that human learners found languages with more words more difficult to learn. The intuition behind this failure was simple: no models took into account the greater difficulty involved in remembering a larger number of words.

In the second part of our study, we used two probabilistic models, the Lexical Model of Goldwater et al. (2009) and a Bayesian transitional probability model (DM-TP), to investigate how models could be adapted to take into account memory limitations. We evaluated two proposals for modeling memory effects: evidence limitations and capacity limitations. While both of these proposals were successful in allowing the models to fit data from Experiment 3 (where we varied the number of word types in the language), the capacity-limited Lexical Model provided a slightly better fit on several dimensions. Because the probabilistic modeling framework provides a toolkit for limiting some aspects of storage and computation during inference, we used the DM-TP model and the Lexical model as case studies of memory limitations.

Although our investigation of memory and resource constraints focused on imposing resource limitations on probabilistic models, PARSER incorporates these restrictions as

part of its design (Perruchet & Vinter, 1998, 2002). These design features make PARSER more difficult to evaluate on its sufficiency for naturalistic corpus data, but they considerably simplify evaluation for fidelity to experimental data. In addition, the versions of the Lexical model that are explored in the final section bear a striking resemblance to PARSER and—modulo the important issue of un-scored distractor probabilities—perform comparably. Thus, our results overall provide strong support for incremental models with limits on their ability to retain what they learn. We hope that our work provides a template for future modeling work to incorporate limitations on memory resources.

The issues we have explored here are not unique to statistical segmentation or even to statistical learning. Instead we view this particular investigation as an instance of a larger problem: how to differentiate between models of learning. While much work has focused on the types of linking assumptions that can be used to model different response measures, here we focused on a different aspect of how to link models to data: the imperfections in human leaners' abilities to maintain complex representations in memory. We found that a variety of simple manipulations which capture the basics of human resource limitations can be effective in capturing some aspects of human performance. An exciting possibility is that these memory limitations may also lead to a better fit to human inductive biases by describing the ways that forgetting data can lead to effective generalizations (Newport, 1990; Gómez, 2002).

How broadly applicable are our conclusions? One potential concern about our results is that we have made a number of choices in model design in our simulations in the final part of the paper. Some of these decisions were made at least partially in response to the data we collected, thus we may have unwittingly introduced additional effective degrees of freedom in our models (Hastie, Tibshirani, Friedman, & Franklin, 2005). Crucially, however, all of the models we tested are not simply models capturing the shape of the human data, they are also models of the task of statistical segmentation. It is only after a model succeeds in the task that we compare its success across conditions. All models are fit to only 21 data points across the three experiments, and in many cases we experimented with a range of parameter values (and more informally, a range of modeling decisions). Nevertheless, this restriction of sufficiency (that a model actually accomplish the task) severely restricts the space of possible models to evaluate for their fidelity to human performance. Without this restriction it would almost certainly be possible to describe the results of our three experiments with a much smaller number of parameters and modeling decisions. Thus given the sufficiency constraint, we believe the lack of fidelity to the data of the set of models evaluated in the first section of the paper is informative about the necessity of incorporating memory limits into models of human performance.

Given that the best-fitting memory-limited models in our final comparison were variations on the Lexical model, our results are consistent with previous literature that supports a chunking view of statistical learning (Giroux & Rey, 2009; Orbán et al., 2008). Although the memory models that we used in our simulations did not include the kind of inferential processes that would be necessary to capture the prototype enhancement effect shown by Endress and Mehler (2009), the addition of a noise process such as that used by Orbán et al. (2008) or (in a slightly different domain) Goodman, Tenenbaum, Feldman, and Griffiths (2008) would likely account for this phenomenon. However, more work will be necessary to decide this issue conclusively. It may even be the case that there are probabilistic transition-finding or associative accounts which can make the distinction between evidence and capacity limitations and can enforce the same kind of parsimony biases as chunking models do (for an example of a parsimony bias in an associative model, see e.g. Dayan & Kakade, 2000).

Finally, we note that our Experiment 3 (which varied the number of word types in the segmentation language while keeping the total number of tokens constant) takes a single trajectory through the space defined by the number of types and the number of tokens. A more complex experiment could in principle vary these independently, fully mapping the effects of type/token ratio on learning. While this kind of experiment is beyond the scope of the current investigation (and potentially quite difficult using the between-subjects laboratory methods we used) it would likely be a valuable contribution, since the relationship between types and tokens has been quite important in recent discussions of human and machine learning (Goldwater et al., 2006b; Gerken & Bollt, 2008; Richtsmeier, Gerken, & Ohala, submitted).

In this work, we have only scratched the surface of modeling human memory, and have avoided one of its most puzzling aspects: how it is both so effective and so limited. By the time they are eighteen, English-speakers are estimated to know more than 60,000 words (Aitchison, 2003) yet they can remember only a handful of particular words at any one time (Miller, 1965; Cowan, 2001). Though these facts are well-appreciated in the literature on human memory they are largely neglected in work on statistical learning. We believe that integrating computational accounts of learning with the strengths and limitations on human memory is one of the most important challenges for future work in this area.

References

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon*. Oxford, UK: Blackwell.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324.

Brady, T., & Oliva, A. (2008). Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. *Psychological Science*, *7*, 678–685.

Brent, M. R. (1999a). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1), 71-105.

Brent, M. R. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, *3*(8), 294–301.

Brown, S., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*(1), 49–67.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, *13*(4), 359-394.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology Learning Memory and Cognition*, *31*(1), 24-3916.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Daw, N., & Courville, A. (2008). The pigeon as particle filter. *Advances in Neural Information Processing Systems*, *20*, 369–376.

Dayan, P., & Kakade, S. (2000). Explaining away in weight space. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14*. Cambridge, MA: MIT Press.

Doucet, A., De Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. New York, NY: Springer Verlag.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van Der Vrecken, O. (1996). The MBROLA project: towards a set of high quality speechsynthesizers free of use for non commercial purposes. In *Proceedings of the fourth international conference on spoken language* (Vol. 3, pp. 1393–1396). Philadelphia, PA.

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

Endress, A., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351–367.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822-15826.

Frank, M., Goldwater, S., Mansinghka, V. K., Griffiths, T., & Tenenbaum, J. (2007). Modeling human performance on statistical word segmentation tasks. In *Proceedings of the 28th Annual Conference of Cognitive Science Society.* Mahwah, NJ: Lawrence Earlbaum, Inc.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, UK: Chapman and Hall.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gerken, L., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, *3*, 228-248.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*, 260-272.

Goldwater, S., Griffiths, T., & Johnson, M. (2006a). Contextual dependencies in unsupervised word segmentation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (Vol. 44, p. 673).

Goldwater, S., Griffiths, T., & Johnson, M. (2006b). Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 459–466). Cambridge, MA: MIT Press.

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21-54.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108-154.

Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago, IL: University of Chicago Press.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). *The elements of statistical learning: data mining, inference and prediction*. New York, NY: Springer.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a human primate: statistical learning in cotton-top tamarins. *Cognition*, *78*, B53-B64.

Johnson, M. H. (2008). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Human Language Technology and Association for Computational Linguistics* (pp. 398—406).

Jurafsky, D., & Martin, J. (2008). *Speech and language processing*. Prentice Hall.

Jusczyk, P., & Aslin, R. N. (1995). Infants detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1-23.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*, B35-B42.

Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Levy, R. P., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 937–944).

Liang, P., & Klein, D. (2009). Online EM for Unsupervised Models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 611–619).

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.

MacKay, D. J. C., & Peto, L. C. B. (1994). A hierarchical dirichlet language model. *Natural Language Engineering*, *1*, 1–19.

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

Miller, G. (1965). The magic number seven, plus or minus two. *Psychological Review*, *63*, 81–97.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.

Newport, E., & Aslin, R. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*, 2745–2750.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(246-263).

Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness as an alternative model of the mind. *Behavioral and Brain Sciences*, *25*(03), 360–380.

Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353.

Richtsmeier, P. T., Gerken, L., & Ohala, D. K. (submitted). Contributions of phonetic token variability and word-type frequency to phonological representations.

Table 1
Correlation between the DM-TP model (using the full training dataset) and human data across a range of values of the Dirichlet prior parameter α.

α-value	Expt. 1	Expt. 2	Expt. 3
0	0.84	0.43	-0.99
1	0.84	0.92	-0.99
2	0.84	0.93	-0.99
4	0.84	**0.93**	-0.98
8	0.84	0.93	-0.96
16	0.84	0.92	-0.90
32	0.84	0.91	-0.64
64	0.85	0.90	0.16
128	0.85	0.89	0.75
256	0.85	0.88	0.92
512	**0.85**	0.87	**0.96**

Saffran, J. R., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606-621.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 726–731).

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86-132.

Teh, Y. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 985–992).

Toro, J. M., & Trobalon, J. B. (2005). Statistical computations over a speech stream in a rodent. *Perception and Psychophysics*, *67*(5), 867-875.

Vul, E., Frank, M., Alvarez, G., & Tenenbaum, J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 1955–1963).

Appendix: Tradeoffs between prior and evidence in the DM-TP model

In our limited evidence simulations, we found that the addition of a Dirichlet prior to the simple TP model enabled a relatively good quantitative fit to the pattern of human data. Why did adding this prior help? The simplest interpretation of this kind of prior is as a belief held by the learner before coming into the experiment that there is a very uniform transition matrix between the syllables *go* and *la*. Taken literally this interpretation seems odd or inappropriate.

On closer inspection, however, the smoothing scheme which we describe above is mathmatically equivalent to any number of other formulations, many of which have much more reasonable psychological interpretations. The observation that motivates these equivalences is that a strong prior (a large value of α) combined with a standard likelihood (normalized transition count) produces the same posterior value as a normal prior and a small likelihood. Thus, we can see that Equation 7 could also be written as

$$P(s_{t-1}, s_t) = \frac{\alpha^{-1} \cdot C(s_{t-1}, s_t) + 1}{\sum_{s' \in S} (\alpha^{-1} \cdot C(s_{t-1}, s') + 1)} \tag{7}$$

A variable prior can thus be thought of as equivalent to a variable amount of weight on the likelihood. This variable weight on the likelihood can in turn be interpreted as a failure to store or encode some portion of the training examples.

There is a simple tradeoff between reducing the amount of data available to the DM-TP model and increasing the smoothing prior, therefore. Although we used both manipulations to equate the DM-TP and lexical models in the simulations we reported, we could also have presented the model with the full amount of evidence but varied α across a wider range.

To investigate this issue, we varied α across the range 2^0 ... 2^9 and did not limit the input corpus. Correlation coefficients are shown in Table 1. Changing the smoothing coefficient did not change the correlation with human performance on Experiment 1 (though as in Figure 4 it did change the absolute range of performance). In Experiment 2, a small amount of smoothing produced the kind of gradual increase seen in human performance, resulting in a high level of correlation for values of α between 1 and 16. Even a very small α value produced some difference in the accuracy of TP estimates between 48 and 96 exposures and captured the basic shape of the human trend. However, as we increased α further the trend became increasingly more gradual, less accurately reflecting the asymptotic shape of the huma higher n curve. This trend mirrors the same result we saw with limited evidence and smaller values of α, further reinforcing the point that within the DM-TP model, a larger prior is equivalent to a smaller amount of data.

In Experiment 3, in contrast, small values of α did little to change the model's performance. It was only when α was increased to a level greater than the total number of tokens that participants saw for any particular pairing that the model performance began to reflect human performance. What caused this change? The TP model succeeds based in the different numbers of counts between within-word and between-word transitions. With a low value for α, this difference is much greater in the 9-type condition. However, because the number of tokens for any particular transition is smaller in the 9-type condition, some large values of alpha "wash out" the size of this difference even though a meaningful difference remains in the 3-type condition.