The Telephone Game: Exploring Inductive Biases In Naturalistic Language Use

Stephan C. Meylan(smeylan@berkeley.edu)
Brett Goldstein (brettg@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA
Anna N. Rafferty (rafferty@cs.berkeley.edu)
Computer Science Division, University of California, Berkeley, CA 94720 USA
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

In the classic telephone game, the content of a spoken message evolves as it passes from player to player. Beyond its entertainment value, the telephone game may have considerable scientific utility: Here we investigate the nature of the linguistic knowledge people use to comprehend language by tracking the evolution of a set of visually-presented sentences in a web-based version of the telephone game. Initial sentences are selected from a range of probabilities according to n-gram language models. Both unigram and trigram probabilities of responses increase over the course of iterated transmission for sentences with the lowest initial probabilities, suggesting that edits are conditioned on participants' implicit probabilistic knowledge of their native language. Further investigations of word-level changes reveal not all words and sequences are subject to edits; rather, participants are more likely to change lower probability words and sequences, and replace them with higher probability content.

Keywords: inductive biases, iterated learning, noisy channel, language change

Introduction

In the childhood game of telephone, each player whispers a word or phrase to a neighbor, who in turn whispers it to another neighbor, and so on. After many retellings, the word or phrase yielded by this transmission process is compared to the starting phrase—often with comical effect. The telephone game is a classic example of serial reproduction, in that each player hears a message (data) from the preceding player, and infers the most likely message (hypothesis) given the input they receive, in combination with their expectations and previous knowledge of language (Figure 1A). Thus beyond amusement, the telephone is an ideal context for investigating the ways listeners deal with noise, and how these processing strategies exert communicative pressures on the content of messages.

Information transmission by serial reproduction was first studied by Sir Frederic Bartlett, who tracked the evolution of stories and pictures recreated from memory after rapid presentation (Bartlett, 1932). More recent work has occurred within the conceptual framework of iterated learning models of language evolution (ILM; Kirby, 2001). Griffiths and Kalish (2007) provided a Bayesian analysis of iterated learning in which hypothesis selection is determined by both the likelihood and the prior probability of alternative hypotheses (Figure 1B). In this framework, the likelihood and prior have intuitive interpretations as knowledge provided by the data and knowledge provided by the expectations of the agent, respectively. As the number of iterations increases in what can be interpreted as a Markov process, the probability that a learner selects a hypothesis h converges to the prior distribution $p(h)$ for that learner. In this sense, serial reproduction can be used to reveal learners' *inductive biases*, or factors other than data that lead learners to favor a hypothesis (Mitchell, 1997).

In the domain of language, these expectations take the form of listener's knowledge of what people are more or less likely to say (Gibson et al., 2013). Listeners integrate a variety of information sources to reconstruct likely utterances from imperfect input, including probabilities of words and multiword sequences, general world knowledge, prosodic regularities, relative frequencies of syntactic parses, and cues from the discursive contexts of utterances. In the current study we examine the contribution of the first of these information sources: knowledge of likely words and sequences.

Psycholinguists have long been aware that listeners exhibit an exquisite sensitivity to the probability of words and sequences, and that such linguistic knowledge facilitates inference across a variety of aspects of language processing (for an overview see Jurafsky, 2003). As a simple example, consider that the letters "b" and "p" may be hard to distinguish in rapid visual presentation, and that a reader might be left with the competing hypotheses that a particular message was either "the man ate a pear" or "the man ate a bear." This reader could use two sources of frequency information to infer the more likely message. She might recognize that "bear" is a much more common word than "pear," and erroneously conclude that "bear" was the final word. Alternatively, she might recognize that "pear" is a more likely completion of the sequence "ate a" than "bear." Given recent work suggesting that lexicons may be optimized for communication on the ba-

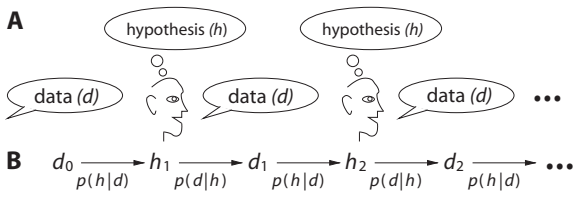

Figure 1: A: In iterated learning, each participant formulates a hypothesis about the observed utterances, and on the basis of that hypothesis produces data for the next generation. B: A Bayesian model of this process; note $p(h|d) \propto p(d|h)p(h)$.

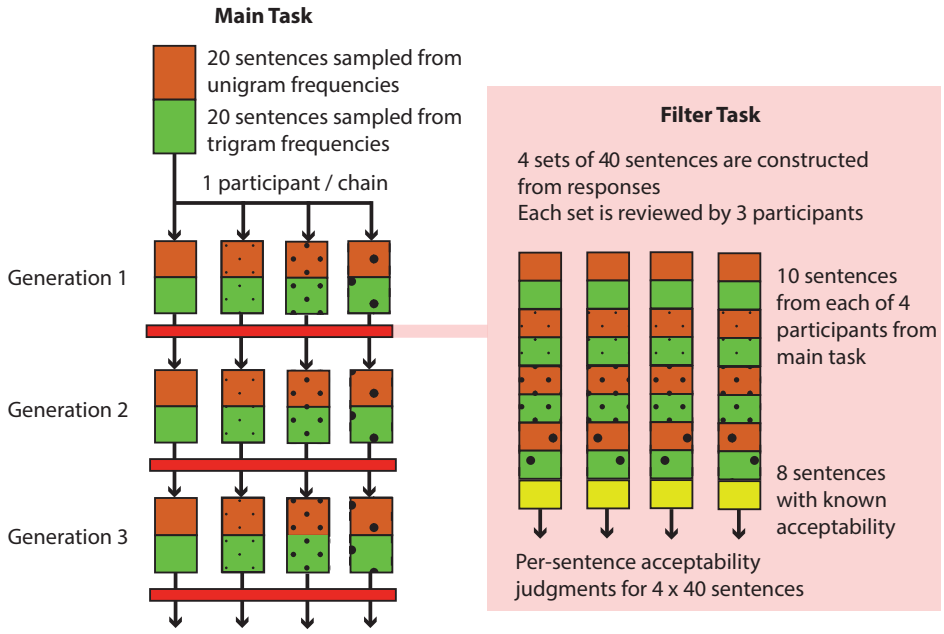

Figure 2: Structure of the serial reproduction task. Participants in the main task played the telephone game, while participants in the filter task assessed whether the responses of each particiant were well-formed English sentences.

sis of predictability—probability given a particular linguistic context—rather than pure word frequency (Piantadosi et al., 2011), we assess how both models of probabilistic linguistic knowledge predict participants' edit rates.

As in the traditional version of the telephone game, participants in the current experiment must replicate the input they receive as best they can, and their response is used as the input for the next participant. Unlike the childhood game, initial sentences are carefully selected for their statistical properties, players interact through a web application, and all inputs and responses are stored in a database. The resulting rich dataset allows us to characterize the nature of the changes over time, and examine the predictive utility of these two models of probabilistic linguistic knowledge.

Methods

Experimental Design

The experiment consisted of two tasks: a main experiment in which participants transcribed sentences following fast visual presentation, and a filter experiment in which a separate pool of participants assessed whether the responses from the first group were well-formed English sentences (Figure 2). Acceptable responses served as input to later participants in the main task. The constraint on syntactic and semantic well-formedness was enforced to ensure that the same faculties used in fluent language use were required for processing stimuli in the main experiment. The filter task provides naturalistic judgments of sentence acceptability from a population rather than relying on arbitrary judgments of the experimenter. Both tasks were completed by workers on Amazon Mechanical Turk; additional workers were recruited as new work became available in either task.

Initial Sentences

For the main task, unigram and trigram log probabilities for all 10-word U.S. English sentences from the TASA (Zeno et al., 1995) and Brown Corpora (Kucera & Francis, 1967) were calculated with the SRI Language Modeling toolkit (Stolcke, 2002). *N*-gram probabilities were estimated using counts from a separate set of texts, the British National Corpus (2007), to prevent overfitting. Unigram probability of words were calculated using a discount factor that allocated half of the probability mass to words not present in the corpus. Sentence probability is defined as the product of probabilities of the constituent words. To test sentences across a range of unigram frequencies, twenty sentences were chosen by selecting one from each 5% quantile bin.

To test sentences from across the range of contextual predictability values, another twenty sentences were selected from 5% quantile bins of trigram probabilities. To find the probability of a trigram (three word sequence), we find the conditional probability of the three word sequence given the initial two-word sequence. Sparsity poses an even greater problem in the trigram model than the unigram model, as many possible three-word sequences may not be observed in the corpus. We thus used modified Kneser-Ney interpolation, as described in Chen and Goodman (1998), to smooth trigram probabilities. Again, half of the probability mass was allocated to sequences not observed in the corpus. Sentence trigram probability was calculated as the product of probabilities of the constituent trigrams.

To control for the length of sentences as a property influencing transmission, all initial sentences consisted of 10 words and were 51 characters long (42 content characters with 9 spaces). Input sentences were manually filtered to exclude inputs with contractions, proper nouns, acronyms, numbers, hyphenated words, profanity or other offensive content, as well as phrasing in the passive voice. Each time a sentence failed to meet requirements, a new sentence was drawn from the same probability bracket and checked against these criteria. Four example initial sentences are shown in Table 1.

Main Task: Telephone

The same set of 40 sentences, presented in random order, served as the stimuli for the first participant in each of four independent "chains," equivalent to four separate games of telephone. For each participant, each sentence was presented word-by-word for 200ms (2s per sentence), followed by a 500ms blank screen before an input screen where they were instructed to type the sentence they had just seen. Participants were not allowed to continue if a submission was not 10 words long, contained punctuation, or contained profanity, and were provided with a brief error message identifying which requirements their submission failed to meet. Two practice sentences were used to familiarize the participants with the task. Upon completion, responses from the main task were sent to the filter task for acceptability judgments. Sentences deemed acceptable replaced their predecessors in the stimuli presented to the next participant; for sentences deemed unacceptable, the last accepted instance of that sentence was presented to the next participant.

Given the difficult nature of the task, most participants provided a few (< 5) syntactically and/or semantically ill-formed responses that could not be used as input to the next participant in the chain (Figure 3A). By accepting and rejecting submissions on a sentence-by-sentence basis rather than accepting and rejecting all work from a participant, the vast majority of responses could be accepted while maintaing the integrity of the input for future participants. All four chains ran participants in parallel, but the increment of each sentence was independent of all other sentences in the chain, as well as all other chains.

Participants were naive to the serial nature of the task. Compensation for the main task was $.75 for approximately 12 minutes, with no conditional incentives offered. A participant who participated in the main task could not repeat the main task, nor participate in the filter task.

Filter Task: Acceptability Judgments

A second parallel task on Mechanical Turk was used to assess whether the responses of each participant in the main experiment constituted acceptable—namely semantically and syntactically well-formed, with no filter words—input to later participants. Three distinct participants in the filter task provided acceptability judgments for each set, ensuring that sentences from a single participant in the main task were reviewed by 12 participants in the filter task. After one participant from each chain submitted a response in the main task, these responses were recombined to create four new sets of 40 sentences, each containing ten sentences from each participant Each participant in the filter task received 40 randomly-ordered sentences, presented sequentially, and were asked to accept a sentence "if it sounds like something someone would say," and to reject it otherwise. Further specifications were provided to reject any sentence with consecutive repeated words, sentences saying something like "I forgot" or "I wasn't paying attention," and any sentences with filler material, profanity, suggestive language, or any misspelled words. Participants were also asked to provide a short reason if they chose to reject a sentence. Two practice questions were used to familiarize participants with the task.

In addition to the sentences requiring review from the main task, the filter task stimuli for each participant also contained 8 randomly interspersed sentences drawn from a set of 40 semantically and/or syntactically ill-formed variants of the initial sentences. Participants in the filter task were told in the instructions that a subset of sentences were of known acceptability and would be used to gauge the quality of their submission; whenever they erroneously accepted a known unacceptable sentence they were shown an error message instructing them to proceed more slowly and to read the sentences more carefully. If a participant in the filter task accepted more than two of the gold sentences incorrectly (corresponding to a binomial probability of .965), their judgments were not used; if they passed, their acceptability judgments were retained for use in combination with the judgments of other participants in the filter task. Once twelve filter tasks were completed, all sentences from four participants in the main task had been rated three times. Sentences with at least two positive judgements were accepted for use in the next iteration of the main task. Four more participants were allowed to enter the main task; the next participant assigned to each chain was then presented with the latest accepted submission within

Table 1: Example input sentences, their unigram and trigram probabilities, percentile ranks, and their character length.

	Unigram		Trigram		
Sentence	Log Prob.	Percentile	Log Prob.	Percentile	Characters
"they found that they had many of the same interests"	-37.28	94	-27.09	92	51
"the molecules that make up the matter do not change"	-38.71	79	-28.51	85	51
"they went on a short hike one warm autumn afternoon"	-42.72	18	-35.89	25	51
"each nonfiction book has a call number on its spine"	-43.56	11	-41.01	5	51

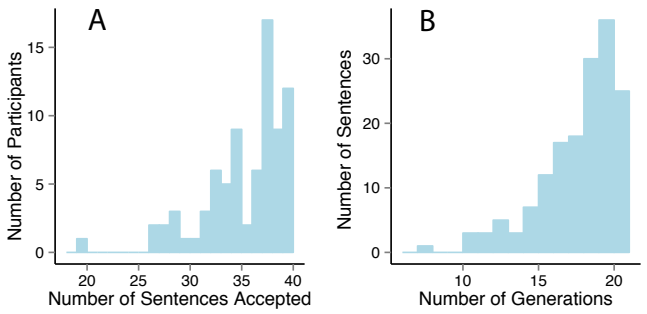

Figure 3: A: Number of sentences accepted from 80 participants in the main task. B: Number of generations of responses for each of 160 sentences (40 initial sentences × 4 chains).

their chain for each of the 40 sentences. The by-sentence approval procedure meant that while most of the sentences came from the immediately preceding participant, a small number came from earlier participants within the chain. Payment for the filter task was $.50 for approximately 8 minutes, with no conditional incentives offered. A participant who participated in the filter task could do up to 10 repetitions, but was not permitted to participate in the main task. For both experiments, English speakers from the U.S. were recruited through Amazon Mechanical Turk, though no test was administered to assess language skills because the filter task identifies participants with insufficient knowledge of English to produce suitable responses.

Results

Twenty sets of four participants were run in the main task; 2749 of their 3200 (85.9%) responses were accepted by participants in the filter task. The number of sentences accepted per participant is displayed in Figure 3A. Between 7 and 20 responses were accepted for each sentence in each chain, the distribution of which is show in Figure 3B. In the filter task, judgments from 240 of 400 participants were accepted; the remainder failed to catch planted unacceptable sentences. The filter task was very challenging precisely because of the phenomenon under study: participants' language models influence their interpretation of the input. As such, many participants in the filter task had trouble detecting singular-plural agreement issues, verbs with improper argument structures, and other errors. Despite low yields, the filter task fulfilled its objective of ensuring that the content of sentences remained syntactically and semantically well-formed.

Table 2: Example edit string from input sentence to output sentence. M, D, I, and S indicate Match, Deletion, Insertion, and Substitution respectively.

M	M	M	M	D	I	I	M	D	M	S	M
you	may	not	notice	yourself			grow	from	day	to	day
you	may	not	notice		as	you	grow		day	by	day

Sentence-Level Analyses

Unigram and trigram probabilities for response sentences were calculated according to the procedure outlined in Methods. The probability of sentences from the lowest quartile of both initial unigram and trigram probability increased more over time than sentences from the other quartiles (Figure 4).

To track edits over the course of the experiment, Levenshtein edit distance from each input sentence to the corresponding response sentence was computed over token strings in which unique words were mapped to characters (Table 2). The cost of substitution was set to twice the cost of deletion and insertion in order to maximize the number of matches between input and output strings. A conservative set of substitutions were found thereafter by finding sequential

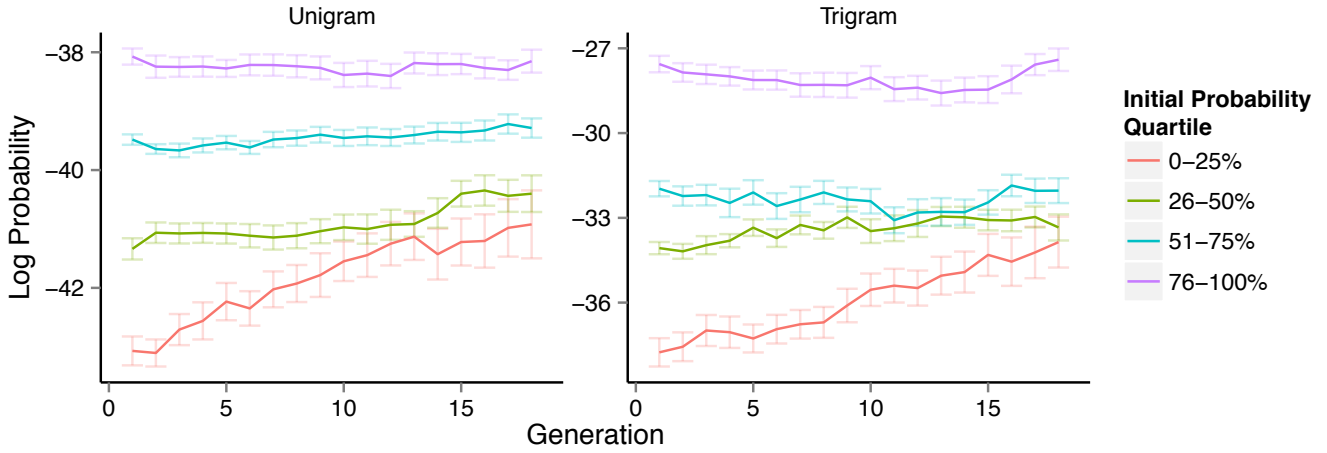

Figure 4: For each sentence, the change in probability of responses over the course of the experiment depends on the initial sentence probability. Participants change sentences with the lowest initial probabilities to produce higher probability sentences, while more probable sentences remain the same. Error bars show standard error.

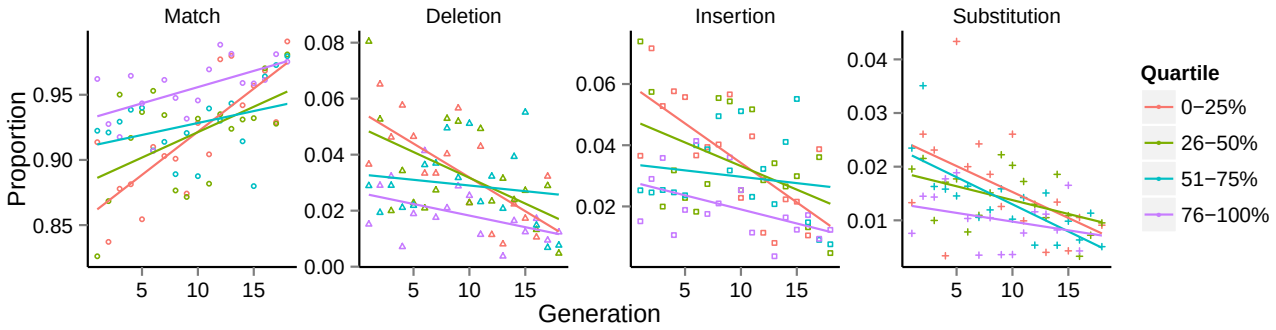

Figure 5: The proportion of matches increases and the proportion of deletions, insertions, and substitutions decreases over the course of the telephone game. Near the beginning of the game, participants change sentences with lower initial probabilities more than those with high initial probabilities, but edit rate converges to similar levels after many generations.

deletion-insertion sequences bracketed by matches. While we make no assertion regarding the cognitive reality of these edit types,[1] this method provides a concise measurement of how frequently each sentence changes over time (Figure 5). Edit rates were calculated by dividing the number of instances of each edit type by the total number of edits of all types. Sentences in the lower probability quartiles (0-25 and 26-50%) started with lower match rates and higher rates of deletion, insertion, and substitution. By later (15+) generations, edit rates converged with higher probability quartiles (51-75 and 76-100%). Though the corresponding edit rates converged, log probability of responses of lower quartiles did not converge with those of higher quartiles (see Figure 4, left). This discrepancy suggests some low probability sentences may be more robustly transmitted than predicted by their unigram and trigram probabilities alone.

We constructed a linear mixed effects regression model to predict edit rate (edits of all types / (edits + matches)) as a function of generation, chain, initial trigram probability quantile, initial unigram probability quantile, and the interaction of each initial sentence probability term with generation as fixed effects. Participant identity was treated as a random intercept. The utility of each predictor was assessed by comparing Bayesian Information Criteria (BIC) for variants of the above model with and without each predictor in turn. This model pruning procedure revealed either—but not both—initial unigram and trigram probability, as well as the corresponding interaction terms with generation, could be removed without appreciably sacrificing model fit (Table 3). In that trigram probability has minimal predictive utility beyond unigram probability, we believe that unigram frequencies are in large part responsible for the observed trigram effect.

[1] It is not possible to determine whether a participant considered a change a substitution or a deletion and an insertion, nor would we assert that these primitives have psychological salience

Word-Level Analyses

Inferring likely edit strings additionally allowed us to track the statistical properties of the contexts in which words remained the same, those in which new material was inserted, and those in which old material was removed (Figure 6). Lower unigram probability words and trigram probability sequences were statistically significantly more likely to contain an edit. Inserted material is higher unigram and trigram probability than deleted material.

Discussion

The current experiment demonstrates that unigram and trigram models are useful for characterizing the inductive biases of participants in a serial reproduction task using natural language stimuli. While unigram and trigram statistics may not be psychologically salient for language users, we nonetheless believe that they reflect some important aspects of people's implicit probabilistic knowledge of language. Additional investigation (across modalities, noisy channels with different

Table 3: Linear mixed effects regression model for edit rate. Degrees of freedom and significance are calculated according to Sattherwaite's approximation (Satterthwaite, 1946).

	Coef β	SE(β)	Approx. *df*	*t*	$Pr(>\|t\|)$
Intercept	0.0805	0.01660	362.0964	4.8490	<**.001**
Initial Unigram Prob. Quantile	0.0046	0.00107	2455.2246	4.3510	<**.001**
Generation	−0.0032	0.00154	554.8863	−2.0671	<**.0001**
Init. Un. Prob. Quant. × Gen.	−0.0003	0.00011	2449.1514	−2.4003	<**.0001**

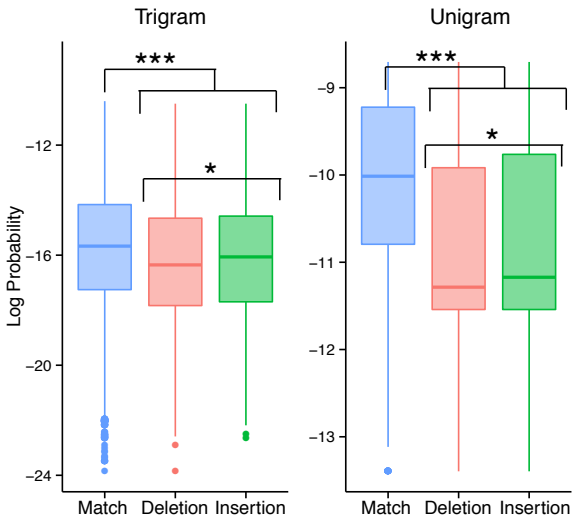

Figure 6: Measures of unigram and trigram log probability of words that were transmitted without change in the output (matches), deleted, or inserted. Substitutions are treated as a deletion and an insertion. *** indicates $p < .001$, * $p < .05$.

characteristics, different communicative contexts, and across languages) is necessary to determine the generality of these observations, as well as to identify how knowledge of probable and improbable sequences interacts with other inductive biases.

A first necessary dimension of future investigation is to explore the effects of frequency and predictability across modalities. Stimuli were presented visually in the current experiment, and participants typed the sentences they saw. While much is shared, the nature of linguistic knowledge that participants employ may vary across modalities. As a trivial example, confusability in sounds (\k\and \g\) is different than confusability in glyphs "k" and "g." Individuals may use different prior linguistic knowledge as a consequence.

A remarkably simple way to produce a sentence of higher probability than the input is to remove a word. In order to investigate trends other than shortening, we enforced the constraint of fixed word length of response sentences. In the absence of such a constraint, we would expect that the probability of all responses would increase as participants elide words. This elision, far from noise, may well constitute a communicative strategy for dealing with uncertainty under noise, may be highly systematic in its application, and as such deserves further explication in its own right.

A final necessary dimension of future investigation is a cross-linguistic evaluation of the extensibility of these results. Inductive biases for English speakers may be qualitatively different than those of other languages. For example, speakers of a relatively free word-order "non-configurational" language like Warlpiri (Austin & Bresnan, 1996) may exploit trigram probabilities to a lesser degree in a serial reproduction task than do English speakers.

Conclusion

More than an idle childhood pastime, the game of telephone can be conceptualized as a close variant of the iterated learning model (Kirby, 2001) wherein listeners infer speakers' likely messages by combining particular hypotheses about the input with their their inductive biases, or general knowledge of language. Trigram and unigram probability models both capture some component of these implicated inductive biases. An analysis of word-by-word changes reveals that edits are more likely for words with lower unigram and trigram probabilities, and that the unigram and trigram probability of the output of such edits is higher than the corresponding input. But these observations are only the first few: this experiment yields a rich dataset suited to testing a variety of hypotheses regarding communicative pressures on naturalistic speech. Complementing experimental paradigms in iterated learning of artificial languages, investigations of language change in serial reproduction provide a valuable new method for characterizing the inductive biases of speakers in their native language.

References

Austin, P., & Bresnan, J. (1996). Non-Configurationality in Australian Aboriginal Languages. *Natural Language and Linguistic Theory*, *14*, pp. 215-268.

Bartlett, F. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

The British National Corpus. (2007). (Version 3, XML Edition. Distributed by Oxford University Computing Services on behalf of the BNC Consortium)

Chen, S., & Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, *13*, 359-393.

Gibson, E., Bergen, L., & Piantadosi, S. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441-480.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic Linguistics.* MIT Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*, 102-110.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Mitchell, T. M. (1997). *Machine Learning*. New York, NY, USA: McGraw-Hill, Inc.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110-114.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing* (Vol. 2, p. 901-904).

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates (TASA), Inc.