Evaluating models of robust word recognition with serial reproduction

Stephan C. Meylan [a,*], Sathvik Nair [b], Thomas L. Griffiths [c]

[a] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[b] Department of Psychology, University of California, Berkeley, Berkeley, CA, United States of America
[c] Department of Psychology, Princeton University, Princeton, NJ, United States of America

ARTICLE INFO

ABSTRACT

Spoken communication occurs in a "noisy channel" characterized by high levels of environmental noise, variability within and between speakers, and lexical and syntactic ambiguity. Given these properties of the received linguistic input, robust spoken word recognition—and language processing more generally—relies heavily on listeners' prior knowledge to evaluate whether candidate interpretations of that input are more or less likely. Here we compare several broad-coverage probabilistic generative language models in their ability to capture human linguistic expectations. Serial reproduction, an experimental paradigm where spoken utterances are reproduced by successive participants similar to the children's game of "Telephone," is used to elicit a sample that reflects the linguistic expectations of English-speaking adults. When we evaluate a suite of probabilistic generative language models against the yielded chains of utterances, we find that those models that make use of abstract representations of preceding linguistic context (i.e., phrase structure) best predict the changes made by people in the course of serial reproduction. A logistic regression model predicting *which* words in an utterance are most likely to be lost or changed in the course of spoken transmission corroborates this result. We interpret these findings in light of research highlighting the interaction of memory-based constraints and representations in language processing.

Spoken communication occurs in a "noisy channel" (Shannon, 1948, 1951): To overcome high levels of environmental noise, variation within and between speakers, and lexical and syntactic ambiguity, listeners must make extensive use of *linguistic expectations*, or knowledge of what speakers are likely to say (Gibson, Bergen, & Piantadosi, 2013). This requires the integration of many sources of information, including knowledge of word frequencies (Howes, 1957; Hudson & Bergman, 1985; Yap, Balota, Sibley, & Ratcliff, 2012), probable word sequences (Miller, Heise, & Lichten, 1951; Smith & Levy, 2013), sub-word phonotactic regularities (Storkel, Armbrüster, & Hogan, 2006; Vitevitch & Luce, 1999), the plausibility of syntactic relationships (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003), cues from the visual scene (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), and pragmatic expectations (Rohde & Ettlinger, 2012). More broadly, people are able to adjudicate between various candidate interpretations of auditory input in light of the specific discourse context, and bring considerable "general world knowledge" (e.g., knowledge of intuitive physics, properties of people and objects) to the task of natural language understanding (Levesque, Davis, & Morgenstern, 2011).

One particularly promising avenue of inquiry has been the development of statistical models that encode regularities in language structure in terms of probability distributions over sequences of words, or *probabilistic generative language models* (PGLMs). These models are distinguished in their ability to provide precise quantitative predictions for a variety of behavioral observables such as reading time and word naming latencies (Goodkind & Bicknell, 2018; Roark, Bachrach, Cardenas, & Pallier, 2009; Smith & Levy, 2013). These models are further distinguished in their ability to reflect human-relevant psycholinguistic hypotheses in their architecture, for example whether syntactic structure is symbolically represented or implicit (Fossum & Levy, 2012; Frank & Bod, 2011). However, as we argue below, there are a number of substantive challenges and limitations in the way these models are currently evaluated, especially in the auditory modality. This leaves important open questions regarding which of these models best capture human expectations in spoken word recognition in sentential contexts.

The behavioral experiment presented here specifically targets a gap in methods for estimating people's linguistic expectations in the audio modality, i.e., what people expect other people to say. In the current work, we develop a method to sample from the linguistic expectations used by people in a naturalistic language task using the technique of
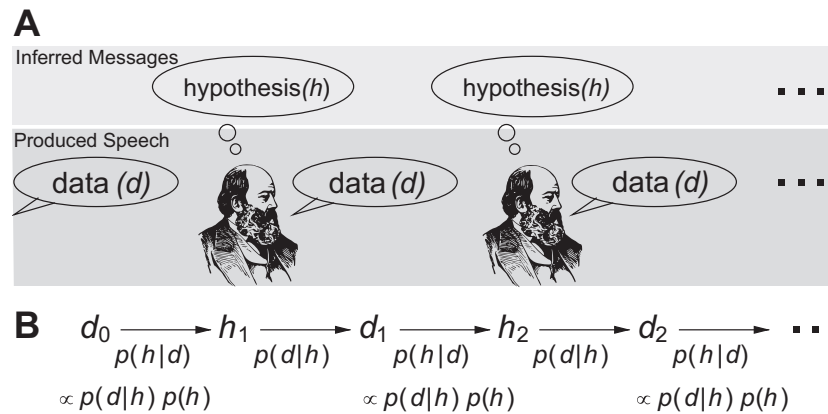
Fig. 1. A: Serial reproduction. Each participant chooses a hypothesis regarding what they heard, and on the basis of that hypothesis produces data for the next listener. B: Under the Bayesian model of the telephone task presented here, participants assign posterior probability $p(h|d)$ to hypotheses regarding what they heard proportional to the product of the likelihood, $p(d|h)$, which depends on the current auditory input, and the prior, $p(h)$, which does not.

"I bought a bear/pear at the farmers' market"

Candidate	Prior	Likelihood	Posterior
bear	Highly improbable (.1)	Fully consistent (1)	.1 /(.1 + .72), p = .12
pear	Probable (.9)	Slightly inconsistent (.8)	.72 /(.1 +.72), p = .88

Fig. 2. Example of the contributions of the likelihood and prior in spoken word recognition in a sentential context. The phonemes /b/ and /p/ are largely distinguished by a single dimension of difference, their voice onset time (VOT), and are relatively easily confused. Using broader language knowledge, a listener can recognize a speaker's likely intended meaning when data is ambiguous or noisy. In this toy example, the prior probability of *pear* in this context overwhelms the evidence for *bear* provided by the received auditory signal.

serial reproduction (Bartlett, 1932; Xu & Griffiths, 2010), in this case a large-scale web-based game of "Telephone". A succession of participants reproduces each initial sentence, such that the utterance produced by each participant is used as input to the next (Fig. 1). Similar to the related technique of *iterated learning* (Kirby, 2001), this technique of serial reproduction can be shown to converge to participants' prior expectations given a sufficient number of iterations (repetitions, with changes introduced by participants), and as long as certain conditions are met (Griffiths & Kalish, 2007). Even with fewer iterations than required for convergence to the stationary distribution, the yielded samples are still useful in that they reflect an approach to the distribution of interest.

To preview our results: the changes people make to sentences in the Telephone game are best explained by probabilistic generative language models that make use of abstract symbolic representations of preceding context. A token-level analysis of which individual words are successfully transmitted from one participant to another provides converging evidence for this result. More generally, we introduce a method to elicit samples from people's expectations for receptive language tasks in the audio modality, and show how that method can be used to evaluate language models in their ability to explain key human linguistic behaviors.

1. Theoretical background

In a landmark study, Luce and Pisoni (1998) showed that the recognition of isolated words embedded in noise can be modeled as a competitive process, combining evidence from the received auditory signal with each word's probability (i.e., relative frequency) in a corpus. Subsequent work has formalized this process of competition among words within an explicitly Bayesian framework, and extended it to the recognition of words in sentential contexts (Norris & McQueen, 2008). It is now widely accepted that linguistic expectations—probabilistic knowledge regarding what is more or less likely to be said—make the process of word recognition substantially more robust than it would be

as a purely data-driven, bottom-up process (Norris, McQueen, & Cutler, 2016). Fig. 2 provides a toy example of how a listener might overcome perceptual noise to arrive at a speaker's intended message. While a listener may receive auditory input more consistent with *bear* than *pear* after the preceding context *I bought a*, the probabilities of the two candidates as continuations given the preceding context would suggest that the latter is a much more likely interpretation. In this way, a listener can overcome noise and ambiguity and successfully recover a speaker's message using their knowledge of language.

A critical question at Marr's computational level of analysis (Marr, 1982) is what sort of information sources might be combined to accomplish the task of word recognition, independent of the precise timecourse or computational tractability. Luce and Pisoni (1998) use word probability—normalized frequency—as an example of basic word-level knowledge that listeners have access to before encountering auditory input. This simple model would, however, often lead to incorrect predictions for the bear/pear example in Fig. 2, because *bear* is more frequent than *pear* in many linguistic corpora. Rather, the plausibility of the two candidates reflects more detailed world knowledge, for example what sort of things might be bought and sold at a farmer's market. In the absence of models that encode this sort of world information directly, increasingly sophisticated models of linguistic structure, or probabilistic generative language models (PGLMs), have been used to approximate people's expectations.

Following on the work of Hale (2001), Levy (2008) tested how the probabilistic linguistic expectations encoded by such a model predicts sentence processing difficulty, as revealed by measures of reading time. This work specifically advanced the concept of a "causal bottleneck:" the expectedness of a word determines its processing difficulty, and this expectedness may reflect a broad range of linguistic and non-linguistic knowledge. The negative log probability of that upcoming word under a listener's expectations, or *surprisal* (see also Itti & Baldi, 2009 for a related formulation in the visual domain), provides a succinct measure of a word's expectedness. Levy (2008) further demonstrated that surprisal estimates derived from a PGLM can be used to approximate

people's rich implicit knowledge of linguistic regularities, and can provide significant explanatory power above and beyond theories of processing difficulty that focus on memory constraints alone.

Though Levy (2008) used a specific probabilistic generative model (the Earley parser of Hale, 2001), the theory of a causal bottleneck posits a broader relationship between surprisal and processing difficulty. Specifically, more sophisticated models—those capable of capturing additional information sources available to people—should be expected to produce surprisal estimates more strongly correlated with observed processing difficulty. Thus while these models entail strong simplifying assumptions regarding the knowledge of language structure available to people (let alone knowledge of language structure available to linguists), they can nonetheless be used to derive expectations from large corpora or other large-scale datasets that are strongly predictive of human linguistic behavior, such as response times in word naming tasks. Subsequent work in psycholinguistics and cognitive science has demonstrated the utility of a variety of PGLMs in understanding human sentence processing, especially for reading (Demberg & Keller, 2008, 2009; Fine, Jaeger, Farmer, & Qian, 2013; Fossum & Levy, 2012; Frank & Bod, 2011; Futrell & Levy, 2017; Goodkind & Bicknell, 2018; Smith & Levy, 2011, 2013). However, the question of *which* models best capture human linguistic expectations—using which information sources and via which representations—remains a central open question.

In addition to the utility of increasingly accurate statistical models of language structure for psychology and cognitive science, these same models are of critical importance for a broad range of speech-related engineering applications, where noise, ambiguity, and speaker variation have been well-known challenges since the 1940s (e.g., Shannon, 1948). Generative models of utterance structure have received extensive treatment in the fields of Computational Linguistics, Natural Language Processing, and Automatic Speech Recognition, where such models are collectively known as "language models" (Jurafsky & Martin, 2009). Their probabilistic form allows them to be combined in principled ways with auditory information using Bayesian techniques, or approximations thereof (e.g., Graves, Fernández, Gomez, & Schmidhuber, 2006). A growing interest in commercial applications in recent years has resulted in a profusion of model architectures, especially ones taking advantage of large-scale deep artificial neural networks (e.g., Devlin, Chang, Lee, & Toutanova, 2018; Dyer, Kuncoro, Ballesteros, & Smith, 2016; Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018; Hannun et al., 2014; Jozefowicz, Zaremba, & Sutskever, 2015; Peters et al., 2018; Zaremba, Sutskever, & Vinyals, 2014). These neural network models have received significant attention in their ability to capture human linguistic phenomena above and beyond simpler structural models, though the exact nature of such remains an active area of investigation (Futrell et al., 2019; Futrell & Levy, 2019; Linzen, Dupoux, & Goldberg, 2016; Linzen & Leonard, 2018).

1.1. Language models

The primary objective of this contribution is to evaluate a variety of broad-coverage probabilistic generative language models (PGLMs) in their ability to capture the linguistic expectations implicit in participants' behavior in a game of Telephone. By "language models," we follow the convention of the natural language processing literature to refer to probabilistic models that encode information regarding the distribution over possible continuations (the next word in the utterance) before encountering the relevant auditory input. We thus exclude many "language models" in the broader sense that are concerned with the online dynamics of word recognition like the logogen model (Morton, 1969), cohort model (Marslen-Wilson, 1987), or TRACE model of word recognition (McClelland & Elman, 1986). In that modeling human performance in the Telephone game requires large vocabularies and a means of handling newly-encountered words, all language models here track expectations over 10,000 or more word types and have a method for handling out-of-vocabulary items.

The language models evaluated here are outlined in Table 2. They include *n*-gram models of varying orders, parsers that use probabilistic context-free grammars, and recurrent neural networks, including LSTMs. Where possible, two versions of each model are evaluated: one fit/trained on the Penn Treebank, and one with a large dataset known to yield competitive prediction results. For additional details, refer to *Technical Appendix 1: Language Models and Training Procedure*. Insofar as each model is the subject of many dissertations, conference papers, and journal articles in its own right, the overviews provided here are neither exhaustive nor formally rigorous; we refer readers to the appropriate publications for more details.

The PGLMs under study here vary in structural complexity from none at all (a unigram model, which may be considered a null model positing no structure at all) to distributions over parse trees, positing hierarchical syntactic relations, such as noun phrases and verb phrases, for all words in a sentence. Language models with higher structural complexity bring the task of in-context spoken word recognition into contact with the task of *sentence processing*: if people think that the utterances they hear will adhere to grammatical rules, their expectations about upcoming words will be quite different. While people certainly use detailed representations of relationships between words in sentence understanding more broadly, it remains an open question whether this same information is used online in the process of identifying words in noisy linguistic input (C. M. Connine & Clifton, 1987; Lash, Rogers, Zoller, & Wingfield, 2013; Levy, Bicknell, Slattery, & Rayner, 2009; McQueen & Huettig, 2012).

One of the central distinctions that we address here is whether there is evidence that people use *categorical, hierarchical* syntactic representations in the service of prediction. Contrary to previous work finding evidence of extensive use of abstract representations in sentence processing, (e.g., Gibson, Desmet, Grodner, Watson, & Ko, 2005), Frank and Bod (2011) found no additional predictive power for models that use explicit hierarchically-structured representations of preceding sentence context for predicting reading times in the Dundee corpus. The models under comparison in their study included *n*-gram models, the Roark parser (Roark, 2001), and echo state networks (Jaeger, 2001), a kind of recurrent neural network architecture. They further expand on this position and describe the prospects and implications of a non-hierarchical theory of language processing in (Frank, Bod, & Christiansen, 2012). By contrast, a replication and extension of Frank and Bod (2011) by Fossum and Levy (2012) found that the use of improved lexical *n*-gram controls derived from a larger model and using a more sophisticated smoothing technique eliminated the performance differences between unlexicalized sequential and hierarchical models. Further, they showed that a state-of-the-art *lexicalized* hierarchical model from Petrov and Klein (2007)—a model that tracks more granular relationships between words and grammatical categories—predicts reading times better yet. We expand on the contrast between lexicalized and unlexicalized models below. Van Schijndel and Schuler (2015) also found evidence that models with hierarchical representations of syntactic context better predict reading times in the same corpus.

In the current work, we investigate the utility of abstract representations in the auditory domain, using a larger sample of language models. We adopt a different typology of models than that used in the above works: we group recurrent neural network models—of which we include two more recent architectures—with models that infer parse trees. While Fossum and Levy (2012) and Frank and Bod (2011) separate those models that make full hierarchical representations of the phrase structure ("latent context in the form of syntactic trees") vs. others, we note that recurrent neural networks, such as the echo state network used by Frank and Bod (2011), capture significant higher-order linguistic regularities of a qualitatively similar type to the nonterminals in a phrase structure grammar. An examination of the word embeddings from Elman (1990), a simple recurrent neural network, show that the average activation pattern for a word reflects a combination of *both* semantic and syntactic similarity, in that both of these are reflected in the lexical distributions in the surrounding context. While this is
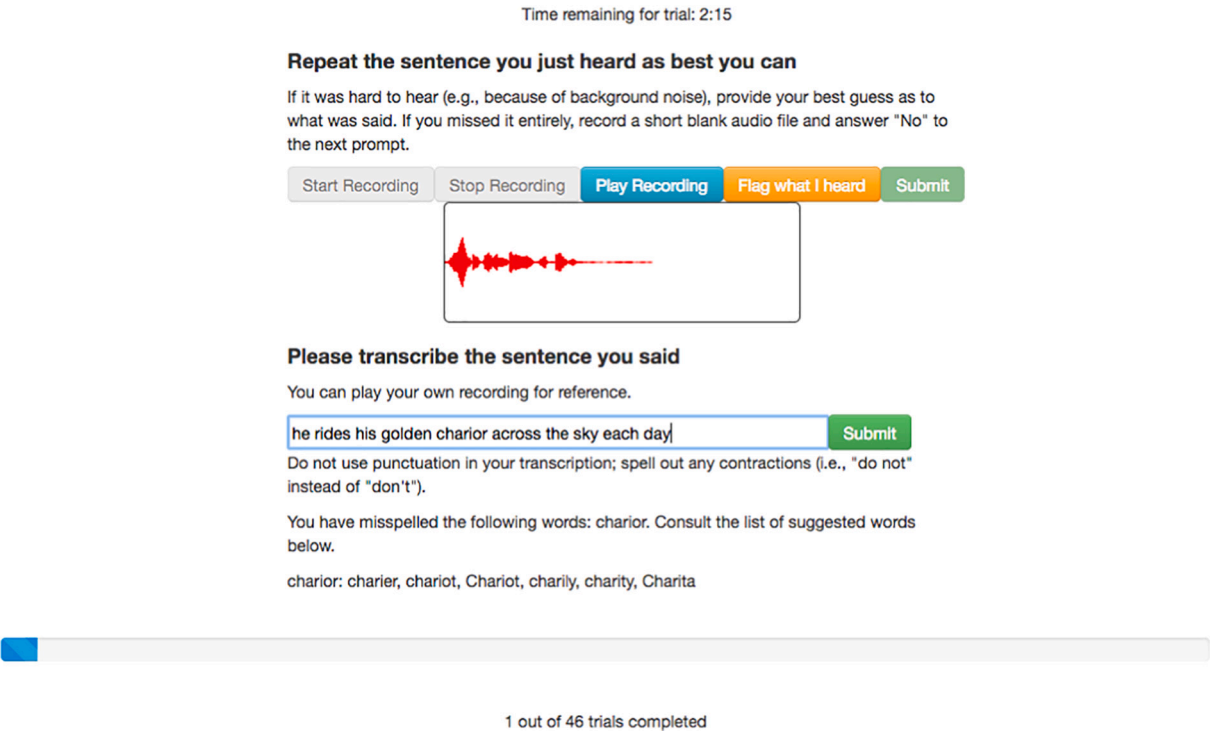
Time remaining for trial: 2:15

Repeat the sentence you just heard as best you can

If it was hard to hear (e.g., because of background noise), provide your best guess as to what was said. If you missed it entirely, record a short blank audio file and answer "No" to the next prompt.

| Start Recording | Stop Recording | Play Recording | Flag what I heard | Submit |

Please transcribe the sentence you said

You can play your own recording for reference.

he rides his golden charior across the sky each day   Submit

Do not use punctuation in your transcription; spell out any contractions (i.e., "do not" instead of "don't").

You have misspelled the following words: charior. Consult the list of suggested words below.

charior: charier, chariot, Chariot, charily, charity, Charita

1 out of 46 trials completed

Fig. 3. The browser-based audio recording and transcription interface used by participants. This screenshot depicts the state of the application after the participant has listened to a recording by a previous participant, recorded a response, and is in the process of submitting a transcription for their contributed recording. In this case the participant has been flagged for misspelling a word.

admittedly far from a full hierarchical parse tree, the use of *any* sort of abstract representation of context may have a stronger effect on the sort of expectations that are encoded in a model, rather than whether the representation is represented symbolically as hierarchical. We evaluate support for this partition of models, as well as the performance of the language models in predicting the pattern of changes, using our collected serial reproduction data. Following Frank and Christiansen (2018), we do not entertain the hypothesis that language processing and especially sentence processing is devoid of hierarchical phenomena (as tested by e.g., Altmann & Kamide, 1999). Rather, we focus our investigation on whether hierarchical or non-hierarchical models better account for the data collected from our task, specifically investigating the role of prediction in in-context word recognition.

As noted in the introduction, PGLMs are typically evaluated according to the probability that they assign to a test sample of language (in natural language processing) or in their ability to predict experimental observables pertaining to processing difficulty (in psycholinguistics). Each of these evaluation methods has notable limitations.

Engineering-centric research in natural language processing typically evaluates generative language models in terms of the probability that they assign to a held-out dataset: the model that assigns a higher probability to a held-out dataset (or *perplexity*, in which probability is normalized by its length in words) is the better model in the absence of confounding factors (Jurafsky & Martin, 2009). This metric assigns the highest scores to those models that best capture the statistics of the corpora they are tested on. Fitting language models to large naturalistic corpora produces excellent first-approximation, broad-coverage estimates of people's general-purpose language expectations. However, people's expectations may deviate significantly from these corpus statistics: people do not expect newly encountered speakers to produce utterances following the statistical properties of corpora derived from books or news (e.g*.,* the Penn Treebank). Even the use of naturalistic corpora like Switchboard (Godfrey, Holliman, & McDaniel, 1992) or Santa Barbara (Du Bois, Chafe, Meyer, Thompson, & Martey, 2000) fails to address this problem, in that participants' expectations for the

purposes of comprehension may deviate significantly from that of any known corpora, a point on which we elaborate below. Likewise, eliciting a sample of unconditioned speech may not reflect a person's expectations of what *others* are likely to say.

Regarding maximizing test sample probability, there is no guarantee that a corpus-derived test set has a high probability under human linguistic expectations. As rational agents, people should be expected to develop expectations that match the task of interpreting linguistic material under noisy conditions, but in the case of language it is very hard to know what will constitute future "typical" linguistic experience. Composition in terms of written vs. auditory sources, topics, and speech registers (i.e., levels of formality) is undoubtedly subject to significant individual variation. Furthermore, even with a perfect estimate of a

Table 1
Example transcriptions of initial stimuli, their unigram and trigram probabilities, percentile ranks, and their character length.

Sentence	Unigram		Trigram		
	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

person's experience with language thus far, there is no guarantee that linguistic material encountered in the future will have the same composition as that encountered in the past. In such a case, a rational agent might be expected to allocate probability mass to as yet unobserved linguistic events to avoid overfitting on the basis of previous experience. As such, we argue that *no* corpus collected from natural sources should be expected to reflect people's linguistic expectations for the task of in-context word recognition.[1]

Psycholinguistic studies, by contrast, have typically evaluated how well surprisal estimates derived from probabilistic language models predict measures of processing difficulty, such as reaction times in word naming, looking time or regressions in reading as measured by eye tracking, and neural measures such as the magnitude of event-related potentials. We note a challenge in interpreting the relationship between surprisal and processing difficulty. Unpredictable linguistic events should cause increased processing difficulty *if and only if* they are correctly interpreted: they may also simply be missed by listeners/readers and result in transmission errors (i.e., a listener could incorrectly interpret an unexpected word as a more probable alternative, without immediately incurring any increase in processing difficulty). As such, behavioral observables constitute an imperfect record of human linguistic expectations. Models should thus be evaluated with respect to how surprisal estimates predict measures of processing difficulty, as well as the kinds of transmission errors they predict.

1.2. Characterizing transmission errors

A substantial body of work in signal processing has evaluated the performance of automatic speech recognition systems (ASR) on conversational speech (Lippmann, 1997). While most of this work has focused on engineering applications, some work has treated these systems as computational-level cognitive models of word recognition. Goldwater, Jurafsky, and Manning (2010), for example, evaluated two contemporary ASR systems in their ability to recognize words from a standard dataset. Consistent with the human results reviewed below, they found higher error rates for low probability words, with strong independent effects of unigram and trigram probabilities. They also found that longer words have lower error rates, that words are more likely to be mis-recognized at the beginnings of utterances, and that the model was particularly prone to transmission failures among "doubly confusable pairs:" pairs of words with similar acoustic profiles that appear in similar lexical contexts. More recently, Xiong et al. (2018) found that neural network-based ASR systems make similar errors to people, though many of the remaining differences can be traced to how these models handle backchanneling (e.g. "mm-hmm") and disfluencies ("uh","hmmm"). This work evaluating ASR systems has, however, systematically evaluated how transmission errors vary as a function of language model.

Word recognition errors have also been thoroughly explored in psycholinguistics and audiology. Early work focused on establishing psychometric functions to relate signal-to-noise ratios of word presentation in noise to recognition accuracy (Egan, 1948; French & Steinberg, 1947). Subsequent work tested the effects of word frequency (Broadbent, 1967; Owens, 1961; Savin, 1963), the interaction of word frequency and acoustic similarity (Luce & Pisoni, 1998), and the importance of lexical neighborhoods (the number of phonetically similar words to a target word, Treisman, 1978; Luce, Pisoni, & Goldinger, 1990; J. Ziegler, Muneaux, & Grainger, 2003) as determinants of word recognition accuracy. While much of this work focuses on the

recognition of individual words in isolation, a major thread explores the facilitatory effects of utterance context (Benichov, Cox, Tun, & Wingfield, 2012; C. Connine, Blasko, & Hall, 1991; Kalikow, Stevens, & Elliott, 1977; Miller et al., 1951). We note however that these studies have largely focused on recognition accuracy in specific sentence positions, usually with small sets of carefully experimentally-controlled stimuli. Further, the characterization of predictability has relied on small corpora, and predictability estimates have often been coded as categorical rather than continuous variables. As noted by Theunissen, Swanepoel, and Hanekom (2009), experiments investigating the transmission of entire utterances have rarely characterized the word-level changes. In the current work, we adopt a methodology that allows us to track transmission errors among *all* words in a large collection of utterances, and uses predictability estimates from sophisticated language models trained on large datasets.

One challenge for word recognition research is that there is a rich correlational structure among word features: shorter words are more frequent, have more phonologically similar competitors, are learned earlier, and are more central in lexical networks (Vincent-Lamarre et al., 2016). In the absence of controlling for these variables, this makes it difficult to interpret many of the above results. More recent work addresses this challenge by including a larger set of predictors, explicitly characterizing the correlational structure among those predictors, and using linear and logistic mixed-effects models to account for participant-level and item-level variability. Looking at this expanded set of predictors and using appropriate regression models, Gahl and Strand (2016) reproduced the longstanding effect of lexical frequency (more frequent words are easier to recognize) and found that words with many phonological neighbors are harder to recognize.

More broadly, there are several reasons to question whether the effects observed in isolated word recognition—which has received the overwhelming balance attention in the literature thus far—will extend to the recognition of words in utterance contexts. First, many candidate interpretations may be much less competitive when contextual cues are available to guide processing. In isolated word recognition, participants may adapt to the circumstances of the task (e.g., expect that low-frequency and high-frequency words are equally likely in an experimental context), and thus diverge from expectations for in-context

Table 2
Overview of probabilistic language models used in the current work.

Model	Type	Training data	Citation
OBWB/BIG LM	LSTM	One Billion Word Benchmark	Jozefowicz et al. (2015)
BNC/Kneser-Ney Trigram	Smoothed *n*-gram	British National Corpus	Chen and Goodman (1998)
BNC/Unigram	*n*-gram	British National Corpus	
DS/Kneser-Ney 5-gram	Smoothed *n*-gram	Librivox, Fisher, Switchboard	Chen and Goodman (1998)
PTB + GTB / BLLIP	PCFG	Penn + Google Treebanks	Charniak and Johnson (2005)
PTB/BLLIP	PCFG	Penn Treebank	Charniak and Johnson (2005)
PTB/Roark Parser	PCFG	Penn Treebank	Roark et al. (2009)
PTB/RNNLM	RNN	Penn Treebank	Mikolov, Karafiát, Burget, Černockỳ, and Khudanpur (2010)
PTB/Unigram	*n*-gram	Penn Treebank	
PTB/Good-Turing Trigram	Smoothed *n*-gram	Penn Treebank	Gale and Sampson (1995)
PTB/Good-Turing 5-gram	Smoothed *n*-gram	Penn Treebank	Gale and Sampson (1995)

---

[1] We note, however, that language models fit on large corpora of naturalistic text are an excellent first approximation to people's linguistic expectations, and that they provide broad coverage for relatively little effort by comparison to the iterated learning method introduced here. We return to this point, as well as prospects for combining the two approaches, in the Discussion.
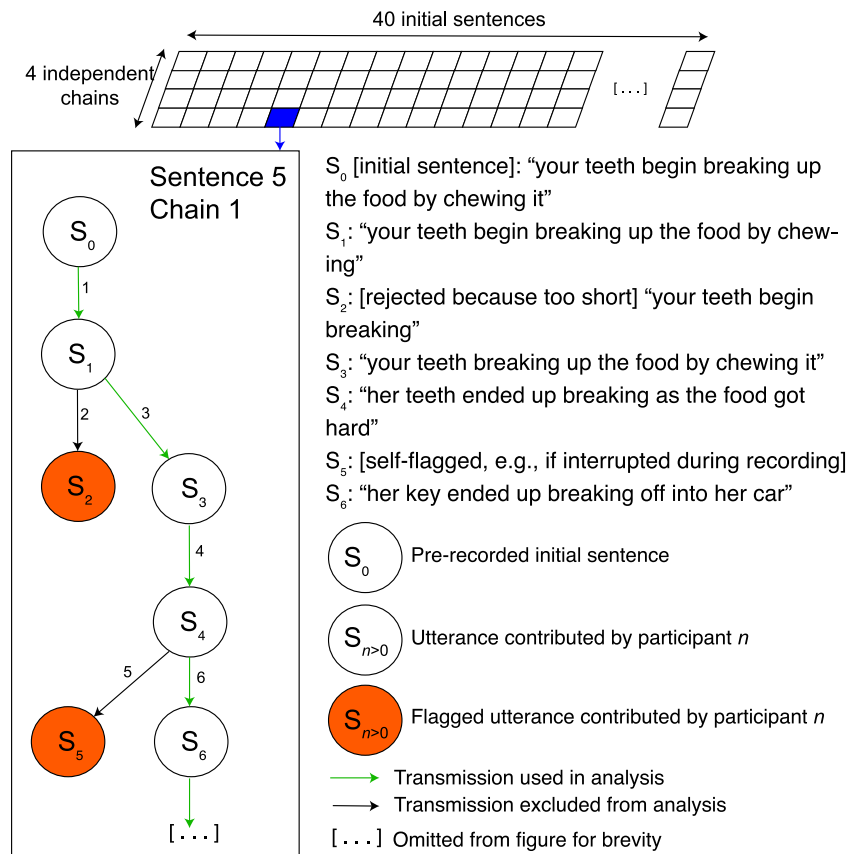
Fig. 4. Schematic of the transmission structure for a single sentence in the experiment. Participants contribute recordings according to the current state of each utterance, unlike generation-based designs. The web app maintains a record of which recordings are flagged by participants or automated methods as too short, bad recordings etc., and offers the last non-flagged recording as input to a new participant.

conversational speech. Second, many of these studies have investigated only a limited set of words, for example only monosyllabic words or words without complex consonant clusters. The current work thus provides an opportunity to test whether patterns observed in isolated word recognition experiments extend to the task of in-context speech recognition.

1.3. Serial reproduction

In the present study, we use the technique of *serial reproduction* to investigate human in-context word recognition. Intuitively, we start with a small test corpus and use serial reproduction to gradually change the properties of that corpus so that it better reflects people's linguistic expectations. This transition in corpus properties emerges in the course of serial reproduction because participants' expectations are reflected in the changes they make between the utterance they hear and the utterance they produce at each iteration.

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 identified that the technique, like iterated learning (Kirby, 2001), can be used to experimentally reveal inductive biases, or reasons that people would favor one hypothesis over another independent of observed data in inferential tasks (Mitchell, 1997). If participants (reproducers for serial reproduction, learners for iterated learning) use Bayesian inference to infer the posterior distribution over hypotheses, and then draw from that distribution in production, then their output at each sequential generation reflects a combination of observed data and participants' inductive biases. Over time, the distribution implicit in the output data comes to reflect participants' inductive biases. For a trial in

the Telephone game, participants' hypotheses are the set of possible interpretations that they might provide for an utterance, in that the message they infer reflects both the auditory input and their expectations regarding language.

Under certain assumptions, both kinds of transmission chains (iterated learning and serial reproduction) can be interpreted as a form of Gibbs sampling, a common technique in Markov chain Monte Carlo based parameter estimation (Geman & Geman, 1984). The transmission of a linguistic utterance across a succession of participants can be interpreted as a Markov process; given certain assumptions, the resulting Markov chain is guaranteed to converge to its stationary distribution if run for a sufficient number of iterations. A participant estimates the probability of the interpretation of an utterance h given some observed data (i.e. an audio recording) d following Bayesian inference (Fig. 1). In particular, following a process of sampling a hypothesis h from the posterior, $p(h|d) \propto p(d|h)p(h)$ and sampling data from $p(d|h)$ means that the probability that a participant selects a hypothesis h converges to their prior distribution $p(h)$ (Griffiths & Kalish, 2007). This approach has been profitably used to reveal inductive biases in memory (Xu & Griffiths, 2010), category structure (Griffiths, Christian, & Kalish, 2008; Sanborn, Griffiths, & Shiffrin, 2010), and function learning (Griffiths, Kalish, & Lewandowsky, 2008). In the domain of language specifically, iterated learning has been used to reveal biases relevant to language evolution (Griffiths & Kalish, 2007).

We motivate the use of serial reproduction with language in the auditory modality by analogy to the function learning experiments of Griffiths, Kalish, and Lewandowsky (2008), which can be thought of as a game of "Telephone" in the space of mathematical functions. In these experiments, participants saw a small number of points in a two-dimensional space drawn from an unknown—and not necessarily

your teeth begin breaking up the food by chewing it
your teeth begin breaking up the food by chewing it
your teeth end up breaking up the food by chewing
Your teeth end up breaking up the food by chewing
your teeth end up breaking up the food by chewing
your teeth end up breaking up the food by chewing it
her teeth ended up breaking to the food back to you
her teeth ended up breaking as the food got hard
her key ended up breaking off into her car
her key ended up breaking in to her car
our key ended up breaking into the car
Our key ended up breaking into the door
Berkie ended up breaking in to the door
Our key ended up breaking into the door
Our key ended up breaking in to the door
Her key ended up breaking into the door
her key ended up breaking into the door
Her key ended up breaking in the door
her key ended up breaking in the lock
The key ended up breaking in the lock
the key ended up breaking in the lock
The key ended up opening the lock
the key ended up opening the lock
the key ended up opening the lock
The key ended up opening the lock
the key to it is upholding the law

Fig. 5. A representative recording chain yielded by the serial reproduction experiment. The first transcription is the initial stimulus. Each subsequent transcription is that of a participant who heard the preceding sentence presented in naturalistic background noise. All responses are collected as audio recordings first, then participants are prompted to provide a written transcription.

linear—function. They were then asked to provide their best guess regarding the function which generated the observed points. This function was then used to generate points for the next participant in the chain. While different transmission chains started with radically different functions, including some of considerable complexity, all are reduced to simple, mostly positive linear functions in the course of reproduction. In the face of limited, noisy data people tend towards their expectation for a simple linear function. The case of language implicates a more complex and nuanced set of expectations—what people expect other people to say—but the basic experimental logic remains the same.[2]

The process of serial reproduction can consequently be thought of as gradually changing a corpus so that it better reflects what people expect others to say. To the degree that the initial set of utterances in that corpus is not representative of what people expect to hear, then the process of serial reproduction will introduce edits that change them to increasingly reflect the broader linguistic expectations of participants. If, for example, the initial corpus contained an excessive number of low-frequency words pertaining to finance, participants should be expected to misinterpret these and replace them with words prototypical of normal conversational registers. To our knowledge, no previous work has used this technique of serial reproduction to investigate language processing. However, we note that the task of serial reproduction of isolated speech sounds was previously used to investigate whether

repeated imitation of environmental noises can give rise to word-like forms (Edmiston, Perlman, & Lupyan, 2018).

2. Methods

We use a web experiment to gather chains of audio recordings and transcriptions appropriate for answering our primary research questions. This web experiment lets participants listen to and record audio, coordinates the flow of stimuli such that recordings from one participant can be used as stimuli for later participants, and ensures that the succession of recordings remains interpretable linguistic material, while avoiding explicit judgments of appropriateness by the experimenters. For simplicity of exposition, we first describe how a participant would experience the task; then we describe the flow of data in the experiment, focusing on how the successive contributions of participants are used as stimuli for future participants.

2.1. Experimental interface

Upon accepting the experiment on Amazon Mechanical Turk, a participant is provided with a link to a web application that allows them to listen to and record responses to audio recordings of utterances (Fig. 3). Participants proceed through two practice trials that test their speakers and microphone. A participant is then introduced to the full trial format, where 1) they listen to a recording 2) they choose whether or not to flag the recording they heard as appropriate and interpretable 3) they record a response (i.e., their best guess of the content of the utterance they heard) 4) they decide whether to flag their *own* recording (e.g., in case of speech errors or excessive ambient noise) 5) they provide a written transcription of the utterance they recorded. After the participant submits a trial, the new recording and transcription are programmatically evaluated with a number of filters confirming that the audio recording is not blank, that it is of similar length to the previous utterance, and that it passes basic tests of consistency between recording and transcription using an automated speech recognition system. The participant does two practice trials in this full trial format to confirm their understanding of the directions of the task; if they fail to pass the basic checks, they continue through additional practice trials. For additional details regarding the web interface and the automated checks, refer to *Technical Appendix 2: Experiment Interface*.

2.2. Stimuli and experimental design

We select a set of 40 initial target sentences and 9 filler sentences from the TASA corpus (Zeno, Ivens, Millard, & Duvvuri, 1995) and the Brown Corpus (Kucera & Francis, 1967). These sentences are chosen to provide maximal variation in probability under different language models for sentences of the same length (Table 1). First we determined which sentence length (in terms of words and characters) was the most common in the corpus (10 words, 42 nonspace characters +9 spaces). All were evaluated as grammatical by the experimenters. For this cohort of length-matched sentences, we then obtain their probabilities under unigram and trigram language models trained on the British National Corpus. For the trigram model, we used modified Kneser-Ney smoothing (Chen & Goodman, 1998) on transitions of order three (for further details regarding this smoothing scheme, see Language Models below). For both unigram and trigram probabilities, we used the empirical distribution of probabilities for the yielded sentences to generate 20 5-percentile tranches. Then for each tranch, we iterated through sentences, rejecting all sentences phrased as questions, interpretable as inappropriate, or containing numerical characters, hyphenated words, or contractions. This yielded a single sentence for each of the 20 unigram and 20 trigram tranches. Two additional sentences were chosen from the TIMIT corpus (Garofolo, Lamel, Fisher, Fiscus, & Pallett, 1993). Initial audio stimuli were read by a male speaker at a normal conversational pace in a soundproof environment.

---

[2] An intuitive interpretation of this process is that repetition entails reversion to the mean. In the case of linguistic expectations, this "mean" is of unknown character and the subject of broad theoretical interest.
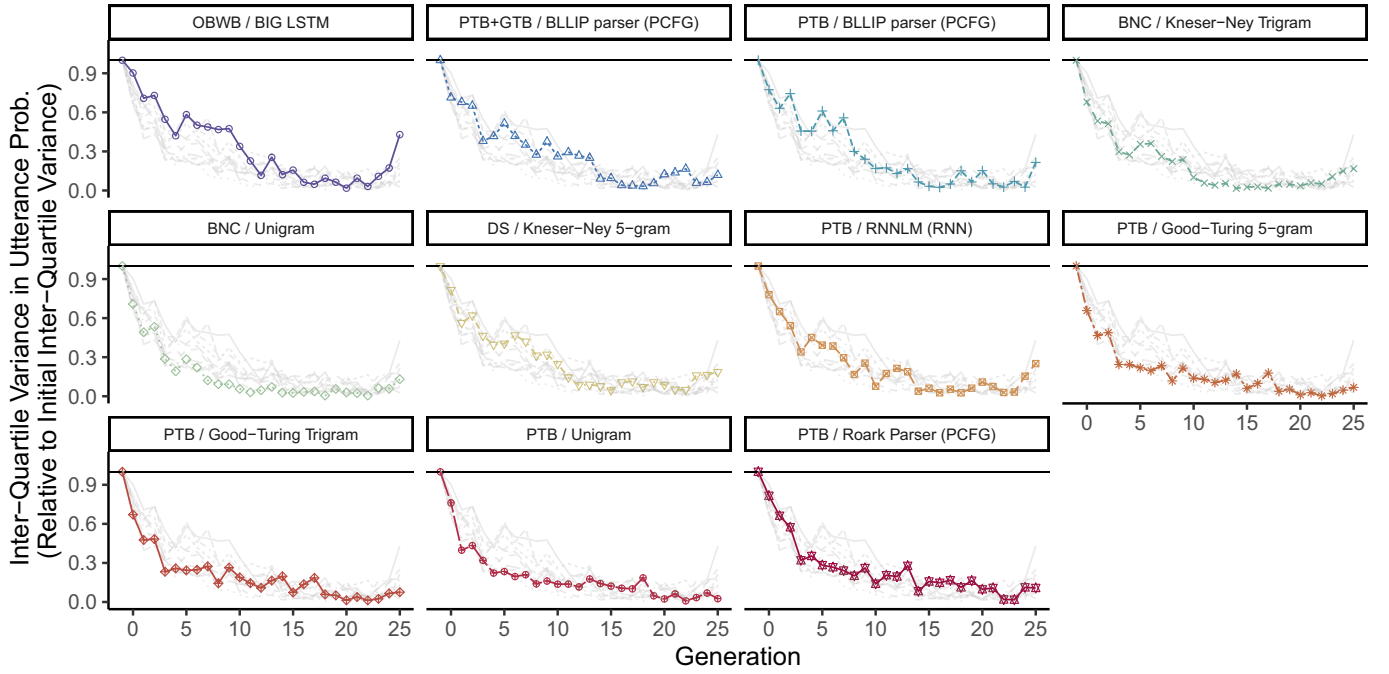
Fig. 6. Variance in language probability estimates at each generation of the serial reproduction experiment relative to initial variance. Variance estimates are computed with respect to means for four groups of chains, defined by initial sentence probability quartile. The reduction in variance suggests that the utterances in chains starting with high probability sentences and low probability sentences becomes much more similar over the course of the experiment.

All chains are initialized with a grammatical, semantically interpretable sentence, but we make no explicit effort to maintain either property over the course of serial reproduction besides the requirement for intelligibility by downstream participants. Because later recordings may not be grammatical sentences, we refer to them as utterances, though they are sentences in a high proportion of cases (see Discussion).

2.2.1. Serial transmission structure

The web application included logic for routing newly-contributed recordings to other, later participants in the experiment. Iterated learning and serial transmission experiments typically use a strict generation-based design where each participant contributes data in response to all relevant experiment stimuli, after which they are given to another participant. We implemented a system that tracked the succession of inputs for each sentence separately for two reasons. First, this
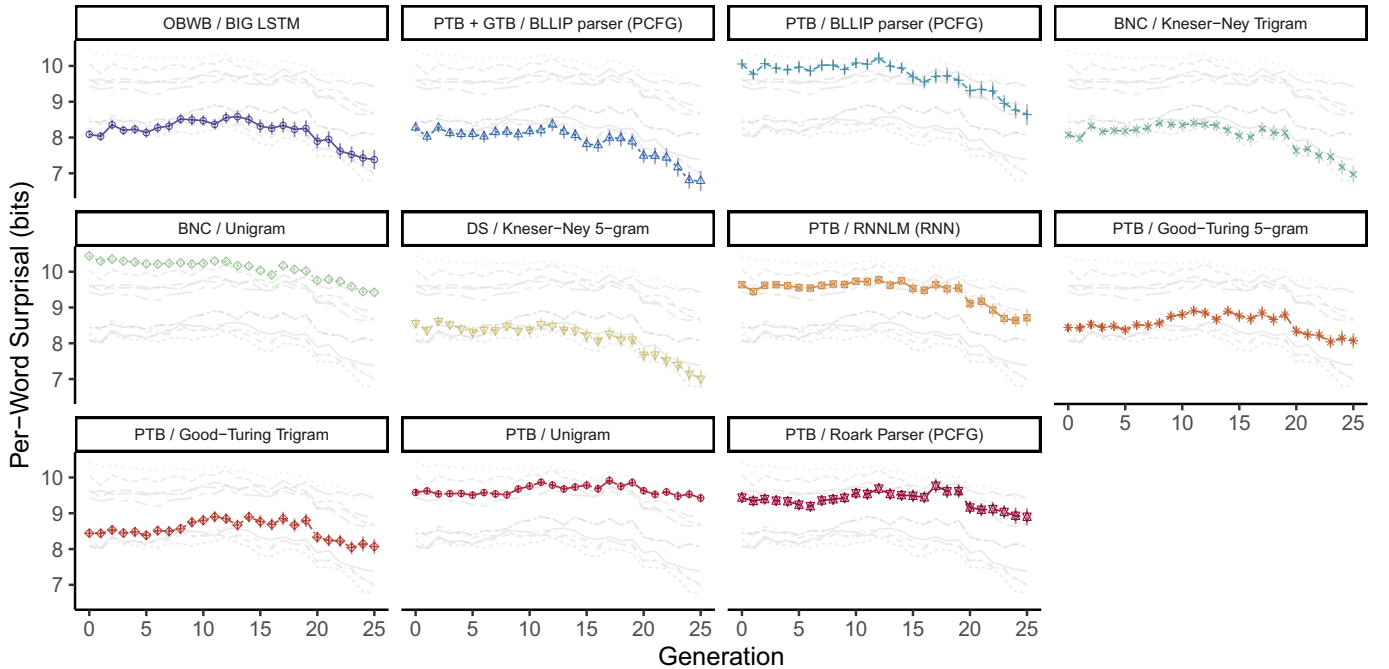

Fig. 7. Average per-word surprisal in bits ($-\log_2(p(w|c))$) under each language model over the course of serial reproduction. Error bars indicate standard error of the mean.
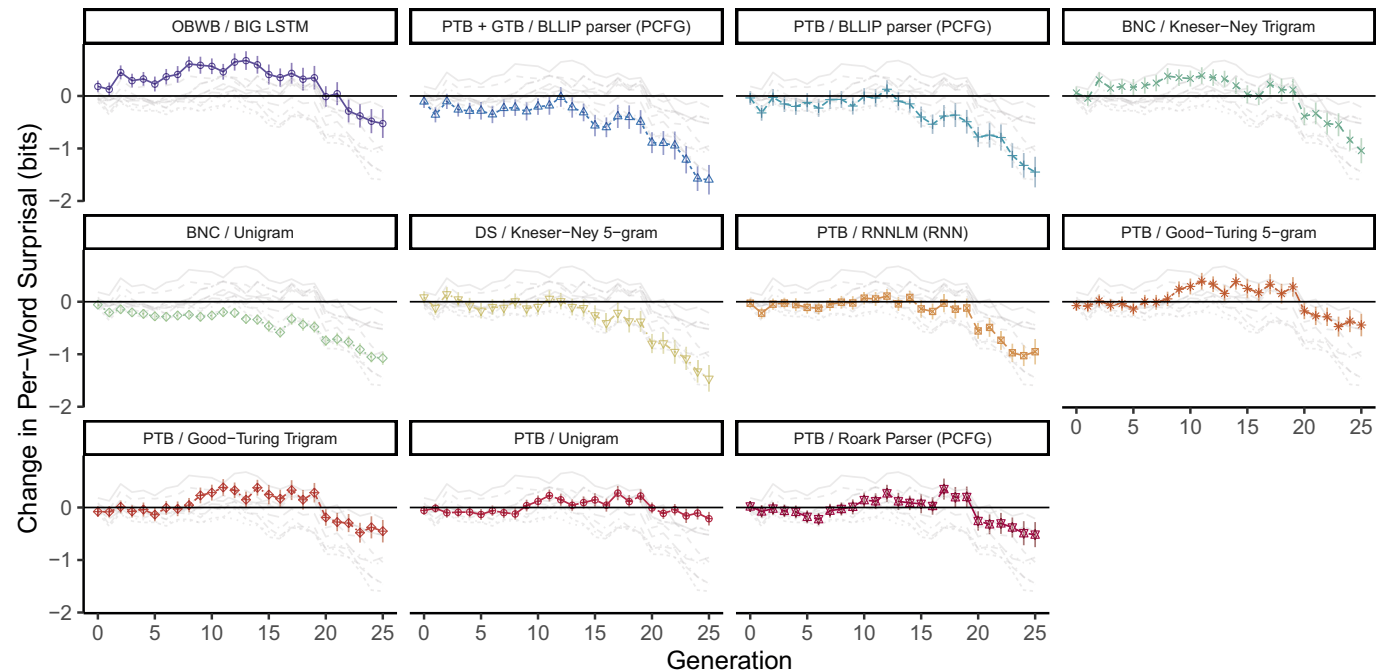
Fig. 8. Change in the average per-word surprisal for utterances over the course of the telephone game under 11 language models. A decrease in average per-word surprisal of one bit means that words in the utterances yielded from serial reproduction are on average twice as probable under that model.

increases the throughput of the experiment: under a strict generation-based design, a new participant cannot start until the previous participant has finished. Second, it allows the experiment to balance the contributions of participants across chains in response to real time flagging of audio files (e.g., for a blank audio files). Sequence of utterances in a chain satisfy the Markov condition because participants sequentially contribute one utterance each. However, unlike a strict generation-based designs, two participants can provide inputs (i.e., be the immediately preceding participant) for one another on different sentences. An example history of a sentence in the experiment, including contributions of new recordings and flagging, is presented in Fig. 4. For more details on the transmission structure, see *Technical Appendix 3: Experiment Transmission Structure.*

2.2.2. Participants and protection of human subjects

A total of 266 participants were recruited using Amazon Mechanical Turk. Participants were limited to those living in the United States with a fast internet connection, a microphone, and speakers. Each IP address and worker identifier (temporarily stored locally in hashed form) was allowed to participate only once. The quality of each participant's internet connection was screened at the beginning of the experiment to avoid possible issues with downloading and uploading relatively bandwidth-intensive audio. Data collection methods, including the audio data retention and distribution policy, were reviewed and approved by the U.C. Berkeley Committee for Protection of Human Subjects. In addition to providing informed consent, participants also completed a media release allowing their submitted audio recordings
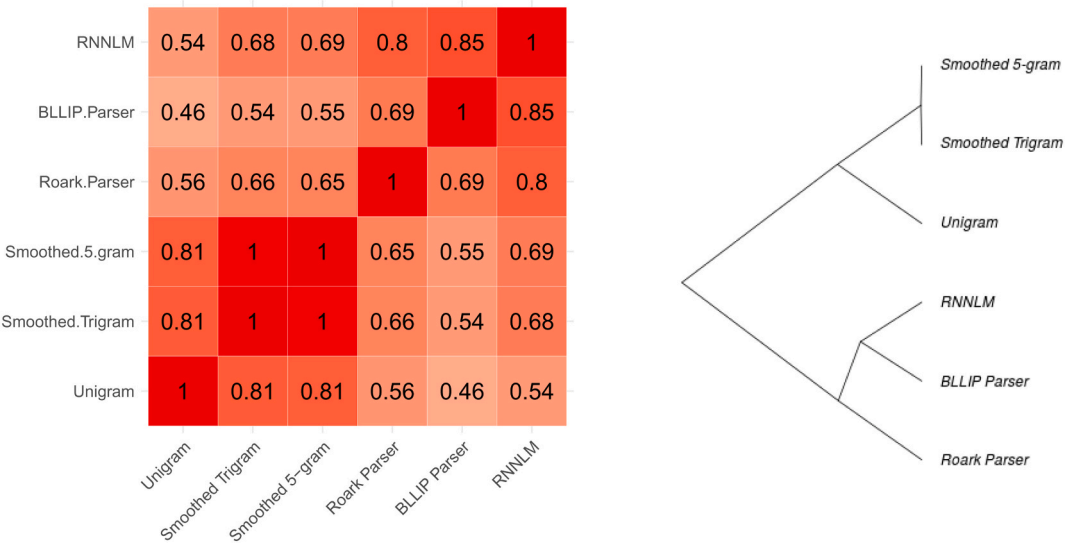

Fig. 9. *A.* Spearman's rank correlation for pairwise combinations of sentence probabilities across the 3193 sentences from the experiment. *B.* The resulting dendrogram, derived from the rank correlations using Ward's method, shows that the model-based probability estimates reflect the architectural difference of interest—whether models represent hierarchical syntactic relations.

your teeth begin breaking up the food by chewing it
your teeth begin breaking up the food by chewing it
your teeth end up breaking up the food by chewing
your teeth end up breaking up the food by chewing
your teeth end up breaking up the food by chewing
your teeth end up breaking up the food by chewing it
her teeth ended up breaking to the food back to you
her teeth ended up breaking as the food got hard
her key ended up breaking off into her car
her key ended up breaking in to her car
our key ended up breaking into the car
our key ended up breaking into the door
berkie ended up breaking in to the door
our key ended up breaking into the door
our key ended up breaking in to the door
her key ended up breaking into the door
her key ended up breaking into the door
her key ended up breaking in the door
her key ended up breaking in the lock
the key ended up breaking in the lock
the key ended up breaking in the lock
the key ended up opening the lock
the key ended up opening the lock
the key ended up opening the lock
the key ended up opening the lock
the key to it is upholding the law

Fig. 10. Probability that each word from an example recording chain is changed in transmission as predicted by a mixed effects logistic regression model. Red indicates that a word is likely to be deleted or replaced with a substitution, while green suggests that a listener is likely to reproduce the word successfully. Colors are mapped from high to low for each sentence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(which constitute personally-identifiable information) to be used in publicly-available corpora. With the exception of the audio recordings, Mechanical Turk worker identifiers, and IP addresses (stored in a hashed format and discarded after the completion of the experiment), no other personally-identifiable information was collected. Participants were not explicitly told that the recordings that they heard might come from other participants, though the media release states that their recordings *could* be used as experimental stimuli. The introduction to the experiment stated that the task was designed to gather data on how people recognize words in conditions with high levels of background noise.

3. Results

The final recording chains consist of 2,999 utterances; a representative recording chain is shown in Fig. 5. An analysis of participants' flagging behavior is presented in *Technical Appendix 4: Analysis of Flagged Recordings*. In this section, we evaluate a suite of probabilistic generative language models in their ability to predict the changes made by participants in a large web-based game of "Telephone." We begin by confirming that the parallel sequences of responses elicited in the telephone game demonstrate movement towards a consistent set of linguistic expectations. We then evaluate predictive language models with two different techniques. First, we evaluate which language models demonstrate the highest magnitude increase in probability over the course of the experiment. Second, we test whether surprisal under each model is predictive of which words are transmitted successfully from speaker to listener.

3.1. Evaluating movement towards convergence

We first evaluate whether the recording chains change in a way that suggests that they are headed towards convergence. Convergence among sampling chains in Markov chain Monte Carlo is often evaluated in terms of a *potential scale reduction factor*, or PSRF (Gelman & Rubin, 1992). The PSRF measures the degree to which between-chain variance in parameter estimates reduces with respect to within-chain variance over the course of sampling. In this case, we cannot directly access quantitative estimates of the parameters of the "true" latent model — the linguistic expectations of human participants. Instead we use a proxy measure: whether the probability estimates for independent chains become more similar within a model over the course of the serial reproduction experiment. While the probability measures could potentially become progressively more similar for other reasons, this pattern is a necessary condition that the chains are approaching the distribution of interest. Because the model-based proxy measures above are noisy, we aggregate utterance chains into groups. Chains are stratified into four groups on the basis of the probability quartile of the initial sentence as computed by each language model. Initial conditions of the recording chains are known to vary significantly on this dimension, in that initial stimuli reflect a stratified sample based on unigram and trigram probability measures. We take the mean probability within each quartile at each generation (sequential position within the recording chain), and from that compute the inter-quartile variance.

The asymptotic decrease in inter-quartile variance seen in Fig. 6 suggests that the recording chains are increasingly similar over the course of the experiment under all language models. Fig. 6 shows that variance drops to less than a quarter of the initial value for all models (with one exception for Big LM at the 25th generation), and around a tenth of the initial value for several of the models trained on large datasets (*n*-gram models trained on the BNC and the DeepSpeech datasets). However, we caution against the stronger assertion that the chains are sampling *directly* from people's expectations because utterances continue to increase in probability at the end of the experiment (next section), suggesting that the utterances need to undergo further changes to converge to human expectations. We leave the possibility of sampling directly from the participants' revealed expectations after convergence to future research, and for now use samples generated as participants *approach* the distribution of interest.

3.2. Evaluating language models

Ideally, the language models here could be evaluated using the probability each assigns to a large set of utterances sampled from participants after strong evidence that sampling chains have fully converged to participants' expectations—both that inter-quartile variance approaches zero and that the probability of utterances is no longer increasing. Because utterance probabilities continue to increase, these chains are likely not directly sampling from participants' expectations by the end of the experiment.

Instead, we use the fact that serial reproduction yields sentences that are increasingly representative of people's linguistic expectations in the task, even if they have not yet converged. As such, the pattern of changes towards the target distribution can be used to evaluate models: more representative models should have a greater magnitude increase in probability (manifested as a decrease in average per-word surprisal) over the course of the experiment. A model reflecting the true expectations of participants should exhibit the largest possible decrease in surprisal over the course of serial transmission. Though we lack direct access to people's expectations, the magnitude of the increase in utterance probability for each language model—as an approximation of those human expectations– is sufficient to rank it with respect to others.

To eliminate the effect of shorter utterances on probability, which would trivially be assigned higher probabilities under all language models, we divide each utterance's negative log probability by the

number of words in that utterance. The resulting measure, "average per-word surprisal" has an intuitive interpretation as the average surprisal, measurable in bits, for each utterance under each model. The average of this measure across chains over the course of the experiment is shown in Fig. 7.

Models vary in their surprisal estimates for reasons outside of the scope of the current analysis, especially the choice of smoothing scheme. For example, consider two unigram models that withhold different amounts of probability mass for word tokens not encountered during training, e.g., 0.05 and 0.01. If these two models were used to produce probability estimates for a set of sentences comprised of exclusively known tokens, the first model would assign a lower probability estimate compared to the second model, even though the probability estimates would be perfectly correlated. We thus focus not on the intercept for each model but the slope of the change in surprisal, taking advantage of the fact that the change in the number of bits ($-\log_2 p(w)$) corresponds to a constant multiplicative factor in probabilities (an increase of one bit, whether it is the third or the seventh, means that the words in the set of utterances are, on average, half as probable).

While average per-word surprisal under each model for recording chains through 25 sequential transmissions is plotted in Fig. 7, the change in average per-word surprisal (in bits) is plotted in Fig. 8. This latter graph reveals that higher-order n-gram models trained on large datasets and the BLLIP PCFG parser trained on the Penn Treebank best capture the changes observed in the course of serial reproduction. The observed pattern of increases in probability corresponds with the absolute model probabilities, in that these models are the same ones that assign the highest probability to the final generations of the sampling chains. n-gram models and the Roark parser (trained on the same Treebank dataset) show statistically significant — though more modest — increases in probability for later sentences. Big LM exhibits low initial average per-word surprisal, but exhibits a relatively minor increase in probability for later sentences.

We next turn to the question of whether there are differences in performance between broad classes of models as a result of theoretically-relevant architectural distinctions, in particular whether models that represent hierarchical syntactic relationships more accurately reflect the changes made in the course of the serial reproduction experiment.

To confirm that the identified theoretical contrast (usage of abstract higher-order representations) is indeed reflected in the probability estimates produced by the models, we first measure the similarity between language models on the basis of the probability estimates they provide for all $n = 3193$ utterance transcriptions yielded by the experiment. If model performance is sensitive to this distinction, correlations should be strongest within the two partitions (models that make use of hierarchical representations vs. those that do not). Because the form of the relationship may not be linear, we use Spearman's rank correlation coefficient to evaluate the pairwise similarity. We limit this analysis to variants of models trained on the Penn Treebank to eliminate the effect

of training dataset on model performance (see *Technical Appendix 5: Similarity of Language Models* for a comparison of all models).

The results of this analysis yield two major clusters, the n-gram models, which do not use higher-order representations of lexical context for prediction, and the RNN-LM and the two PCFG models which do (Fig. 12). Among the n-gram models, the unigram model is distinguished from the higher-order n-gram models; these longer-context n-gram models are more similar than the unigram model to the models with higher-order abstract structure. The yielded similarities show that the predictions derived from these models reflect this key architectural distinctions of theoretical interest regarding the representation of linguistic context. Further, these results substantiate our claim that recurrent neural network language models are better grouped with the PCFG models than the n-gram models given their representational capacities, in contrast to the classification used in Frank and Bod (2011) and Fossum and Levy (2012).

The central architectural question we address is whether models that represent higher-order regularities, such as phrase structure or abstract semantic representations of the preceding context, better reflect the revealed linguistic expectations of participants than models that track only specific preceding words. The n-gram models (where $n > 1$) use purely lexicalized representations of preceding context, and do not posit any sort of higher-order abstractions such as verb phrases or noun phrases. By contrast, the PCFGs explicitly represent phrase structure, and recurrent neural network language models are sensitive to higher-order syntactic and semantic regularities insofar as these are captured in the lower-dimensional embedding of preceding context.

We construct a mixed effects linear regression model to predict average per-word surprisal with the following predictors: dataset (Penn Treebank vs. large representative dataset; treatment coded with the latter as the reference level), abstract representation of context (used vs. not used; treatment coded with the latter as the reference level), and the interaction of both preceding terms with generation. Unique recording chains (independent sequences of utterances from serial reproduction) are treated as random intercepts, and we account for a `recording chain × generation` random slope. The regression model is fit with 11 model estimates of average per-word surprisal for each of 2864 utterances produced between the 1st and 26th participant (thereby omitting the initial sentences).

We first test whether accounting for each language model's type of representation (abstract hierarchical vs. lexical) improves the overall fit of the mixed-effects model by comparing the above full model specification with a nested model lacking type of context representation as a predictor. This reveals that the model that includes the type of context representation as well as its interaction with generation exhibits significantly better fit ($\chi^2 = 1401.1$, $p < 0.0001$). For this full model (Table 3), there is a statistically significant negative coefficient for the interaction for `generation * abstract structure: used` ($\beta = -0.0044$, t value $= -6$, $p < 0.0001$). This greater magnitude decrease in

Table 3
A mixed-effects linear regression examining average per-word surprisal estimates (negative log probability under a model) across 9 language models as a function of 1) whether the model uses an abstract representation of the preceding context to inform expectations and 2) the dataset used to fit the model.

Fixed effects	Coef β	SE(β)	t value	p
(Intercept)	2.3362	0.0272	85.81	**<0.0001**
Generation	0.001	0.0025	0.38	0.7
Abstract structure: used	0.2382	0.0092	25.8	**<0.0001**
Dataset: PTB	0.2923	0.0088	33.38	**<0.0001**
Generation x abstract structure: used	−0.0044	7e-04	−6	**<0.0001**
Generation x dataset: PTB	0.0042	7e-04	6.03	**<0.0001**
Random effects	Std. Dev			
(Intercept) | Recording chain	0.35			
Generation | Recording chain	0.03			

Note: Significance of fixed-effects is computed following Satterthwaite (1946).

Table 4

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

surprisal estimates suggest that the changes made by people are better reflected by the language models that use abstract representations of preceding context. As with the previous analysis, we find that models that are trained on large datasets are more representative of people's linguistic expectations than models trained on the Penn Treebank ($\beta = 0.0042$, *t* value = 6.03, $p < 0.0001$).

3.2.1. Predicting word-level errors

A second way to compare the fit of probabilistic generative language models to human behavior is to evaluate how well they predict which words are misheard: altered or deleted in transmission.[3] We specifically test whether models that represent abstract hierarchical structure provide predictive power, in the form of their conditional probability estimates, beyond lexical models regarding which words are altered or deleted. Instances of deletions and substitutions are identified using dynamic programming, using the edit operations corresponding to the Levenshtein distance, or minimum edit distance between utterances. As is commonly done to estimate Word Error Rate (WER) in natural language processing (Popovic & Ney, 2007), this computation is applied to word sequences rather than character strings (Table 4). We present a separate analysis of participants' edit rates in relation to their flagging behavior in *Technical Appendix 4: Analysis of Flagged Recordings*.

We then construct a mixed-effects logistic regression model including model-based estimates of surprisal, as well as the control variables of position in sentence, age of acquisition ratings ([Dataset] Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), concreteness ([Dataset] Brysbaert, Warriner, & Kuperman, 2014), number of phonemes, number of syllables ([Dataset] Brysbaert & New, 2009), and phonological neighborhood density ([Dataset] Yarkoni, Balota, & Yap, 2008) as fixed effects to predict whether the word changes (1 = change, 0 = no change). We restrict model based estimates to those from larger datasets (i.e., omitting models trained on the Penn Treebank if another model with the same architecture but trained on a larger dataset is available). To deal with the strong correlations between language model predictions, we use a residualization scheme to isolate their respective contributions. Unigram surprisal (negative log probability) is used directly as a predictor. We then predict trigram probability from unigram probability, and use the residuals as a predictor representative of the contribution of a language model which considers the two words preceding the words in question as the preceding context. For each of the remaining models with word-level surprisal estimates (5-gram on the DeepSpeech dataset, BigLM on the One Billion Word Benchmark, the Roark Parser trained on the Penn Treebank) we take the residuals after predicting its surprisal estimates from both unigram and trigram models.

The identities of the listener and the identity of the speaker are treated as random intercepts to account for variability in comprehension performance and speaker intelligibility. The model is fit with 27,290 instances where a participant heard a word and either 1) reproduced it faithfully in their own recording (22,482 cases) 2) produced an identifiable substitute (substitution) or 3) did not produce an identifiable substitution (deletion). Cases 2) and 3) were collapsed into a single category of transmission failure, coded as a 0. We fit the full model and conduct no pruning of predictors.

We report the contributions of lexical properties and the word's position of the sentence before reporting on the contribution of model-based estimates of predictability (Table 5). Words are more likely to change if they are appear later in the utterance ($\beta = 0.0989$, *z* value = 15.57, $p < 0.0001$), or if that word is acquired later in development ($\beta = 0.0544$, *z* value = 3.96, $p < 0.0001$). By contrast, words with more syllables or more phonemes are *less* likely to change ($\beta = -0.249$, *z* value = -4.91, $p < 0.0001$), as are words rated as highly concrete ($\beta = -0.0856$, *z* value = -4.46, $p < 0.0001$). In line with recent regression models investigating recognition accuracy in isolated word (Gahl & Strand, 2016), this analysis suggests that words in sparse phonological neighborhoods — words with high average phonological Levenshtein distance to the 20 most similar wordforms (or PLD-20, per ([Dataset] Yarkoni et al., 2008)) are less likely to change ($\beta = -0.0948$, *z* value = -2.22, $p < 0.05$).

The mixed-effects logistic regression reveals that words with higher unigram surprisal are more likely to change ($\beta = 0.3827$, *z* value = 17.25, $p < 0.0001$), as are words with higher trigram surprisal with unigram surprisal partialed out ($\beta = 0.3729$, *z* value = 19.37, $p < 0.0001$). For all models with abstract structure (with trigram surprisal partialed out), higher residualized surprisal estimates are predictive of a transmission failure. This corroborates the effect of word frequency on recognition seen in Gahl and Strand (2016), and is consistent with the independent contributions of word frequency and word predictability seen in automatic speech recognition systems (Goldwater et al., 2010). An example transmission chain with each word colorized by the probability of successful recovery by the listener under the full model are shown in Fig. 10. This figure shows higher rates of transmission failures for low-frequency "content" words vs. generally high-frequency "function" words, though function words in certain constructions have higher probabilities of transmission failure (e.g. "*in to* the door").

Given that the fit model can be used to estimate the probability of change for each word in the dataset, it can then be evaluated against the

Table 5

Mixed-effects logistic regression predicting whether a word will be changed in transmission on the basis of its surprisal under various language models as well as other word properties.

Fixed effects	Coef β	SE(β)	*z* value	$Pr(>|z|)$
(Intercept)	−2.6554	0.0933	−28.45	**<0.0001**
BNC unigram surprisal	0.3827	0.0222	17.25	**<0.0001**
Residualized BNC trigram surprisal	0.3729	0.0193	19.37	**<0.0001**
Residualized Roark PCFG syntactic surprisal	0.1159	0.0266	4.36	**<0.0001**
Residualized Big LM surprisal	0.1959	0.0181	10.82	**<0.0001**
Residualized DS 5-gram surprisal	0.1593	0.0262	6.07	**<0.0001**
Position in sentence	0.0989	0.0064	15.57	**<0.0001**
Age of acquisition	0.0544	0.0137	3.96	**<0.0001**
Number of phonemes	−0.0609	0.0231	−2.64	**<0.01**
Number of syllables	−0.249	0.0507	−4.91	**<0.0001**
Concreteness	−0.0856	0.0192	−4.46	**<0.0001**
Phonological Neighborhood Density (PLD20)	−0.0948	0.0427	−2.22	**<0.05**
Random effects	Std. dev.			
Listener ID	0.62			
Speaker ID	0.3			

Note: Significance of fixed-effects is computed following Satterthwaite, 1946.

---

[3] While we also collect data regarding insertions in the course of serial reproduction, it is harder to identify the relevant properties of the preceding recording that prompt the insertion of material.

Table 6
Comparison of individual language models for predicting transmission failures. Each model contains the same set of word properties, plus surprisal estimates from one language model.

Model	AUC	AIC
OBWB/Big LM	0.639	23,677
DS/5-gram	0.641	23,687
BNC/Trigram	0.638	23,732
BNC/Unigram	0.604	24,095
PTB/Roark Parser	0.592	24,370

ground truth of whether the word changed or not. The area under the ROC curve corresponds to the probability that in a large random sample of (transmission failure, transmission success) pairs the model assigned a higher probability of change to the word that was actually changed vs. the word that was not changed. The full ensemble model (AUC = 0.650, AIC = 23,504) outperforms simpler ones, for example one based only on word predictability (surprisal) under language models (AUC = 0.632; AIC = 23,900) and one based on invariant word properties (AUC = 0.583, AIC = 24,445). This difference between areas under the respective AUC curves and the difference in AICs suggests that while both sets of predictors contribute to predictions, predictability under language models is a better predictor of transmission failure than invariant properties of words.

The previous ensemble model tests whether each model architecture makes an independent contribution in predicting transmission failures. Another question is how the models directly compare to one another in their ability to predict transmission failures. To this end, we fit a series of models with the same set of non-model-based predictors (e.g., position in sentence, age of acquisition, concreteness, and phonological neighborhood density) and use the per-word surprisal estimates from a single model. We then compare the logistic models in terms of AIC and area under the ROC curve (AUC). This analysis (Table 6) reveals that large LSTMs and smoothed 5-gram trained on a large collection of conversational text best predict transmission failures. The Roark Parser has the least predictive power, though this may reflect that its training corpus (the Penn Treebank) was at least a hundred times smaller compared to those used to fit the other language models.

4. Discussion

In this work, we investigate several probabilistic generative language models (PGLMs) in their ability to capture people's linguistic expectations. A serial reproduction task — the game of Telephone, where participants reproduce recordings made by other participants in sequence—reveals participants' prior expectations in a naturalistic spoken word recognition task. We find evidence that people use abstract representations of that context to inform their expectations, in line with the results of Fossum and Levy (2012) for reading.

4.1. The role of abstract structure

Before further interpretation of the results regarding the utility of abstract structure, we clarify the exact nature of our claims and note several methodological challenges that temper the strength and generality of these results.

First a top-level theoretical clarification is in order: Higher-level abstractions are thoroughly implicated in the processes of sentence comprehension and production (e.g., verb argument structure). Our objective in this study, following Frank and Bod (2011), is to investigate whether such abstractions may inform the lower-level process of word recognition.

Second, we highlight an inherent brittleness in using a small cohort of probabilistic generative models to characterize human knowledge

more generally: the encoded expectations reflect a complex interaction of architecture, training data, and fitting procedure. Limited explanatory power for human performance may reflect any of one of the above aspects, or a complex interaction in between them. Further, each language model is drawn from a larger space of possible language models, such that the generalizations we make about model architectures on the basis of a few examples may not be robust.

On the matter of fitting, the encoded expectations may reflect local minima or maxima for some models. While this is not a concern with n-gram models fit with count-based methods, this problem increases in severity for language models which have large numbers of randomly-initialized parameters, especially neural network language models. For these models, continued research focuses on how to improve the speed and robustness of the training procedure. All models may suffer from the problem of overfitting (Geman, Bienenstock, & Doursat, 1992), in which learned distributions reflect properties of the training data at the expense of their generality in extending to new data. The relationship between corpus size and performance is modulated by architecture: Large models trained on small datasets may be more prone to overfitting than small models trained on those same datasets. More abstract structure may allow models to perform better on smaller datasets because they can fall back on expectations for higher-level abstractions in the absence of experience with specific words or sequences. But smaller training corpora may also make it harder for models to arrive at useful abstractions in the course of fitting.

We took several steps to address these pitfalls regarding model performance. First, we confirmed that all models exhibited characteristic levels of performance on standard test datasets, suggesting that the fitting procedures used here are consistent with previous benchmarks. Though we note above the broader shortcomings of evaluating language models in terms of the perplexity on a held-out dataset, perplexity is nonetheless useful insofar as it allows us to check whether the fitting procedure yields models comparable or equivalent to those used in other studies. Second, we treat the dataset on which each model is trained as a separate predictor in our analyses (binarized into those trained on the Penn Treebank and those trained on large datasets known to yield highly performant models). This helps isolate the contribution of the model architecture from the dataset on which the model was trained. Third, we note that deficiencies in model fitting would most likely affect the more sophisticated neural network models or PCFG parsers, which would yield a null result rather than the positive one in favor of structure obtained here.

Another caveat is that the set of models investigated here represents only a small sample of possible models. Sampling a larger set of models may reveal that the differences in fit to human behavior that we find for model architectures are not robust, or that other distinctions in model architecture are predictive of larger differences in performance. We acknowledge this limitation of the current study, and note the possibility of evaluating a larger set of models—reflecting a larger space of architectures—in future work.

The regression analysis presented above yielded a statistically negative coefficient for `generation × abstract structure: used`, suggesting that models using abstract hierarchical structure better predict the changes made by people in the course of the transmission experiment. However, we note that this is a relatively minor quantitative effect. Several of the n-gram models, such as the DeepSpeech Kneser-Ney 5-g and the BNC trigram model, exhibit a pattern of decrease in surprisal estimates that is very similar to models with considerably more sophisticated representations of abstract structure (BLLIP and Big LM). This pattern of near-parity between certain large, higher-order n-gram models and more sophisticated generative models may be explained by a recent theoretical proposal regarding processing difficulty that posits graded use of detailed structural representations in people's linguistic expectations. *Noisy-context surprisal* suggests that while people may use structured, abstract representations of the preceding linguistic context, that such representations are imperfect and in

particular tend to degrade the longer they are kept active in memory (Futrell & Levy, 2017).

Futrell and Levy (2017) highlight the case of *structural forgetting effects*, where people do not effectively use preceding grammatical structures to predict the remainder of the sentence. For example, among the two utterances,

1. *The apartment[1] that the maid[2] who the cleaning service[3] had[3] sent over was[1] well-decorated.

2. The apartment[1] that the maid[2] who the cleaning service[3] had[3] sent over was[2] cleaning every week was[1] well-decorated.

the first is grammatically ill-formed in that no verb corresponds with noun phrase with the head "maid[2]", yet people give it consistently higher grammaticality ratings than the second sentence (originally presented in Gibson & Thomas, 1999). This effect, they argue, arises from an "information locality" effect, whereby structural expectations—which should give infinitely more probability mass to the grammatically correct sentence than the incorrect one—are attenuated for material with longer intervening intervals (in terms of words or time). Under this account, people have fleeting access to the representations (approximated with parse trees) that they form in sentence processing. *n*-gram models can be thought of as an approximation of an abstract, structured generative model, but one with a particularly sharp memory decay function such that an extremely limited sample of the preceding context is used to predict the identity of the next word.

The memory-based effect identified by Futrell and Levy (2017) may be further exacerbated in the current task by high levels of background noise, such that we observe relatively small advantages of abstract structural representations. While imperfect memory imposes noise on the representation of context even under optimal acoustic conditions, this decay may be yet stronger if participants lack peaked estimates regarding what actually constitutes the preceding context.

4.2. An error model for in-context spoken word recognition

Data regarding which words are successfully transmitted from speaker to listener provide for an alternative method for evaluating how well these language models capture people's expectations. Transmission errors are more likely when a word is unpredictable given the preceding context. These models assign divergent probability distributions over continuations in many cases. For example, models with an abstract representation of syntax would assign different probabilities to the continuations "is" and "are" given the preceding context "the quick brown fox and the lazy dog". A hierarchical model generally picks out a subject consisting of the coordinated NPs with heads "fox" and "dogs," such that the plural copular "are" is a more likely continuation (less surprising) compared to "is". An *n*-gram model tracking a limited preceding lexical context, e.g., a trigram model that uses the only the two preceding words, would suggest that "is" is the more probable continuation given the two preceding words "lazy dog"; under this model "are" is more surprising (*"the lazy dog are").

The regression analysis above shows a predictive utility — in the form of statistically significant beta coefficients—for trigram, 5-gram, PCFG (Roark), and LSTM (Big LM) estimates of in-context predictability—suggesting that even the most abstract representations of context are predictive of the collected transmission errors. Nonetheless, we find a continued role of word frequency, consistent with the predictions of the modeling work by Goldwater et al. (2010). A comparison of the individual models reveals that the LSTM (Big LM) and the DeepSpeech 5-gram model best predict the edits; we leave it to future work in the same format as Goodkind and Bicknell (2018) to test whether word predictability estimates from these models also excel at predicting reading times and other behavioral measures of processing difficulty.

The finding that the age of acquisition of a word is predictive of successful transmission—earlier-acquired words are more likely to be recognized than later-acquired ones—extends previous results from isolated visual word recognition (Morrison & Ellis, 2000) into the auditory realm. This is consistent with the hypothesis that such words enjoy a privileged status above and beyond their frequency, perhaps relating the problem of retrieving semantic representations (Austerweil, Abbott, & Griffiths, 2012) given the compositional structure of the lexicon (Vincent-Lamarre et al., 2016) in which lower frequency words are defined in reference to higher frequency ones.

A second notable finding is that once the relationship between word length and unigram surprisal is accounted for, longer words are *more* likely to be recognized. This can be interpreted as evidence that people are better able to recognize words that are more perceptually distinctive, in that words with more syllables have have *fewer* perceptually similar competitors, above and beyond edit-distance-based measures of neighborhood density. These words may also be easier to recover because there is more redundant information about their identity, such that listeners are more likely to be able to recover the intended word if some portion is corrupted in transmission.

4.3. Limitations of serial reproduction

A potential caveat to the generality of the results is that the expectations implicit in the utterances obtained from the Telephone game may be task-dependent, and of limited utility for characterizing linguistic expectations more generally. For example, participants could infer that the task they are performing is qualitatively unlike "normal" language use, and make use of a different set of expectations such that the collected data is not representative of the distributions of interest for language processing. It would be particularly concerning, for example, if the collected utterances demonstrated marked decreases in grammatical acceptability or semantic interpretability over the course of serial reproduction. To evaluate whether participants produced grammatical and semantically-interpretable recordings for the duration of the experiment, authors S.M. and S.N. conducted a follow-up analysis in which they independently coded all utterances from generations 23–25 with binary judgments of grammaticality and semantic interpretability. The latter category was used to distinguish sentences that are structurally well-formed but not interpretable, e.g., "the bus and bus driver were opening the door." Utterances in this set were judged as 87% and 88% grammatical (Cohen's $\kappa = 0.94$) and 78% and 79% semantically interpretable (Cohen's $\kappa = 0.89$). The results of this analysis suggest that participants largely maintain the grammaticality and semantic well-formedness of their responses through the course of the serial reproduction experiment, and suggest that participants are tapping into a similar set of expectations as "normal" language use.

Another important possibility is that participants could be modulating how they use preceding context in word recognition based on the level of noise in the experiment. Because they may be unsure of the preceding context for a particular word, they may prefer shorter, less-structured representations of context in the current experiment, whereas they might rely on that context more heavily in a noise-free environment. This basic logic of noise-modulated expectations is substantiated by the finding of Luce and Pisoni (1998) that participants' reliance on word frequency in isolated spoken word recognition increases as a function of the level of background noise. Audio stimuli here are embedded in relatively high levels of noise, qualitatively dissimilar to the reading tasks of Frank and Bod (2011) and Fossum and Levy (2012). We highlight the importance of characterizing variation in the use of linguistic expectations as a function of environmental noise as an important next step to explore with this paradigm.

One possibility is that participants are providing responses according to the *maximum* a posteriori (MAP) estimate, i.e., choosing the response with the highest posterior probability rather than sampling from possible responses according to their posterior probabilities. If participants respond according to their MAP estimate of what was said by the previous speaker, the equivalence between the stationary distribution obtained in the limit and the participants' prior is no longer guaranteed.

Under certain conditions, the implications of participants choosing according to MAP are known, and the resulting stationary distribution is highly informative regarding the characteristics of the prior. Kirby, Dowman, and Griffiths (2007) introduce a general-purpose method for interpolating between MAP and posterior sampling when hypotheses have a constant likelihood, and demonstrate that the ordering of hypotheses under the prior is preserved under MAP. Griffiths and Kalish (2007) demonstrate the equivalence of MAP to expectation-maximization (Dempster, Laird, & Rubin, 1977) in continuous hypothesis spaces, with a guarantee that the stationary distribution will be centered on the maximum of the prior, with variance increasing according to the rate at which hypotheses change over sequential participants. If the asymptotic normality assumption holds, this same equivalence between the stationary distribution and the prior should be expected here.

While the experimental paradigm presented here can potentially reveal participants' linguistic expectations, it is relatively costly from a sampling perspective. Coverage in the dataset shared here is limited by the small number of initial sentences (reflecting a limited number of semantic contexts), relatively short chains, and relatively strong auto-correlation within those sampling chains. While this method could in principle be used to collect a very large collection of utterances after convergence, from which a new probabilistic model of linguistic expectations could be estimated, a more data-efficient method may be to use the word-by-word responses to fine-tune the parameters of an existing model (D. Ziegler et al., 2019). In future work we will investigate whether this fine-tuning approach can effectively combine the broad coverage of language models learned from large text corpora with the specific behaviors of human participants in the task of in-context spoken word recognition.

Finally, we emphasize the limitations in the generality of these results with respect to individual variation among speakers of English, as well as variation across speakers of different languages. This experiment makes the strong simplifying assumption that the process of serial reproduction yields samples from a single, unified set of linguistic expectations that are shared across all English-speaking participants for the purposes of word recognition. Of course, linguistic expectations should vary significantly as a function of linguistic experience, and may vary between speakers for other reasons like population-level variability in working memory. At a higher level, the expectations of English speakers are certainly not representative of the expectations of speakers in other languages. Given the pronounced typological diversity of languages, expectations may take qualitatively different forms. For example, listeners may rely less on sequential word order in languages with more flexible word order. Future work will be needed to characterize the ways and extent to which expectations vary across natural languages.

5. Conclusion

In this study, we collected data on how utterances change in the course of a web-based game of "Telephone." We use this data, which provides an alternative to corpus-based evaluation methods, to evaluate a range of broad-coverage probabilistic generative language models in their ability to characterize people's linguistic expectations for in-context spoken word recognition. Addressing an outstanding empirical question in speech processing, we find that models that use an abstract (i.e. hierarchical) representation of preceding context to inform expectations regarding upcoming words more strongly reflect the changes made by people. This paradigm offers promise in helping to better understand humans' remarkable language processing abilities.

Acknowledgements

Appendix A. Technical appendix 1: Language models and training procedure

A.1. Language models

A.1.1. n-gram models

The simplest PGLMs that we evaluate here are *n*-gram models. First used to model natural language by Shannon (1948), *n*-gram models typically track forward transitional probabilities for words by conditioning on preceding words. These models make the strong simplifying assumption that the sequence of words in an utterance or sentence are generated following a *Markov chain*, in which the probability of each event (i.e., word) depends only on immediately preceding events. These models can condition on a larger or smaller preceding context: an *n*-gram model tracks conditional probabilities given *n*-1 preceding words: bigram models condition on just the preceding word, while trigram models condition on the two previous words. The probability of an utterance is the product of the conditional probabilities of the constituent words,

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P\big(w_i | w_{i-(n-1)}, \ldots, w_{i-1}\big), \tag{1}$$

where $w_1, \ldots, w_m$ is an utterance comprised of *m* words and *n* is the order of the *n*-gram model. By convention, sentences (or utterances, more generally) are augmented with a start symbol (of probability 1) and an end symbol (the probability of which is tracked by the model, the same as any other word type). These models track statistics over sequences of words, and do not include any explicit encoding of statistical expectations for higher-order abstractions like parts of speech (e.g., nouns and verbs), super-ordinate grammatical categories (e.g., NPs or VPs), or semantic roles (e.g., agents or patients). *n*-gram models may appear to encode expectations related to these abstractions because lexical statistics capture these regularities implicitly: a trigram model will assign almost all of the probability mass following "near a" to nouns and adjectives. However, without support for abstract representations, a trigram model cannot assign higher probabilities to nouns as a class (i.e., nouns not observed following "near a") in predicting the next words in that context. As such, comparing the probability of the serial transmission chains under the *n*-gram models against the probability under the other models—which all posit and track regularities at a level higher than words—allows us to evaluate evidence for the degree to which people may use representations of higher levels of linguistic structure to inform their expectations.

A special case of the *n*-gram model of particular theoretical interest is the unigram, or 1-gram model, where the probability of a word is not conditioned on preceding context. Unigram probability estimates thus largely reflect normalized word frequencies, but may assign nonzero

probability mass to out-of-vocabulary words depending on the choice of smoothing technique, described below. Comparing the predictions of the unigram model with those of higher-order n-gram models allows us to evaluate the degree to which participants' linguistic expectations reflect any amount of preceding context, separate from the treatment of abstraction.

Besides the length of the context on which they condition, a second dimension of variation in the architecture of n-gram models is the choice of *smoothing* technique, or how probability mass is re-allocated to unobserved word sequences during model fitting. Using probabilities directly derived from counts, i.e., the maximum likelihood estimate, of an n-gram model is generally ill-advised because of sparsity: many word sequences observed in a new dataset may not have been observed in the dataset used to fit the corpus. Unless unobserved sequences receive nonzero probability, sentences with zero-probability continuations will be evaluated as also having a probability of 0. This reallocation of probability mass among sequences often reflects simple assumptions about the statistical structure of languages. Smoothing over possible sequences complements the practice of assigning some proportion of the unigram probability mass to newly-encountered tokens.

Here we use two smoothing schemes when higher-order n-gram models. For larger datasets, we use modified Kneser-Ney smoothing (Chen & Goodman, 1998). This smoothing scheme shifts probability mass to unobserved n-grams that are expected on the basis of the prevalence of constituent lower-order n-grams, e.g., assigning "near San Antonio" a small non-zero probability because "near San" (as in "near San Francisco") and "San Antonio" are both relatively probable bigrams. For smaller datasets, we use Good-Turing smoothing (Gale & Sampson, 1995), which shifts probability mass from n-grams seen once to ones that are not encountered in fitting. We build several new n-gram models using the SRI Language Modeling Toolkit (SRILM, Stolcke, 2002), and use a proprietary 5-gram model from the DeepSpeech project (Hannun et al., 2014) which was estimated using KenLM (Heafield, 2011). We continue here with the enumeration of language models; the linguistic datasets used to fit the models are treated in greater detail below.

A.1.2. Probabilistic context-free grammars (PCFGs)

The remainder of the probabilistic generative language models under consideration here use abstract representations above the level of the word to derive word-level expectations. The first class of model, probabilistic context-free grammars (PCFGs), posit that utterances reflect a latent structure composed of abstract, hierarchical categories like nouns, verbs, noun phrases, and verb phrases. The task of predicting the next word is thus not just dependent on the previously-observed words as in the n-gram models, but also depends on a listener's beliefs about the grammatical structure consistent with the previously-encountered words, and how that grammatical structure narrows the set of possible continuations. For example, having heard words that likely form a verb phrase for a transitive verb (e.g., "bought a"), a listener might expect a noun phrase containing the object of that verb.

Under a PCFG, each hypothesis about the hierarchical structure of an utterance is captured in terms of a parse tree, which describes how an utterance could be generated from a context-free grammar. A context-free grammar is a linguistic formalism that describes sentences as a hierarchy of constituents, starting with a root "sentence" node. This sentence node is composed of a tree of nonterminals: grammatical categories such as nouns, verbs, noun phrases, verb phrases, each of which in turn consists of other grammatical categories or words. All terminals, or leaf nodes in the hierarchy, are (observable) words. A CFG captures the intuition that the same (observable) linear sequence of words may reflect several possible (unobservable) interpretations of the relationship between constituents. For example, "the girl saw the boy with the telescope" has two high probability interpretations depending on the attachment point of the prepositional phrase — whether "with the telescope" modifies *boy* or how the girl *saw*. While CFGs cannot capture certain human linguistic phenomena (Stabler, 2004), probabilistic CFGs, or PCFGs, are commonly used as generative probabilistic models of language given their principled account of higher-order structure. Unlike typed dependency parsers (e.g. De Marneffe, MacCartney, & Manning, 2006), which track typed pairwise relationships between words, PCFGs provide a probability distribution over hierarchical parses for an utterance.

The parameters of PCFGs can be learned in an unsupervised fashion from linear word representations alone, or fit using corpora that have been annotated by linguists with gold-standard most-probable parses. Because of their stronger performance, we focus here on the predictions of PCFGs derived from supervised training. Given the size of the hypothesis space of possible grammars for a language, as well as parses for a sentence, PCFGs vary in the techniques that they use to find the highest probability parses. Here we use two PCFGs, the Roark (2001) parser and the BLLIP parser (Charniak & Johnson, 2005), which approach the problem of inference in this large hypothesis space in very different ways.

A.1.3. Roark parser

The Roark parser (Roark, 2001) is a widely used *top-down* parser, meaning that it maintains and updates a set of candidate parse trees as it moves from left to right in an utterance. In combination with other architectural decisions, this allows the Roark parser to calculate probabilities for each sequential word conditioned on the preceding words. This incremental property makes it especially well-suited for modeling online sentence processing, where people update their interpretation of the utterance as they receive new acoustic data (Roark et al., 2009).

A.1.4. BLLIP parser

Unlike the top-down approach of the Roark Parser, the BLLIP (or Johnson-Charniak) parser uses bottom-up inference in combination with a secondary scoring function on the yielded highest probability trees (Charniak & Johnson, 2005). The bottom-up approach means that the parser starts by positing grammatical categories directly above the level of words for the entire sentence, then iteratively identifies possible higher-level grammatical categories up to the sentence root. As such, it cannot be used to produce probability estimates incrementally after each word. The BLLIP parser uses bottom-up parsing to generate a set of high probability parses which are then re-ranked using a separate discriminative model with a researcher-specified set of ad-hoc features. Though bottom-up parsing violates the constraints of online auditory language processing in that it requires that the complete signal be received previous to parsing, we include the parser in analyses tracking utterance-level changes.

For both PCFG language models, the probability of an utterance reflects a marginalization over the possible parse trees that generate that utterance. Because of computational constraints, these parsers typically track a relatively small number of the highest probability parse trees (in this case, 50 for each model). While this represents a truncation of the true set of possible parse trees, the first few (i.e., the first one to five) are overwhelmingly more probable than the remainder, and thus provide a reliable estimate of utterance probability. In the case of the Roark parser, surprisal estimates for individual words can be derived by querying the probability of the next word given the restricted set of highest probability provisional parse trees (Roark et al., 2009).

A.1.5. Recurrent neural network language model (RNN LM)

The expectations derived from PCFGs reflect hierarchical parses of sentences, but listeners could also develop expectations by noting abstract commonalities in the usage of words, for example that the words "five" and "six" tend to appear in very similar lexical contexts. We include here two recurrent neural network models (RNNs) that can infer partially syntactic, partially semantic higher-level regularities in word usage, and can use these regularities in the service of prediction. As with PCFGs, at least some RNN architectures can track long distance dependencies (Linzen et al., 2016). While models lacking abstract representations of context could in principle capture long-distance dependencies (e.g., a 9-gram model), the combinatorial richness of language means evidence is far too sparse to infer these regularities when tracking sequences of words when treated as discrete symbols.

Recurrent neural networks are able to capture and use these higher-level regularities because they use the state of the hidden layer of the network at the previous timestep, often referred to as a "memory," in addition to newly-received data to make predictions. This hidden layer from the previous timestep is a lossy—and thus generalizable—representation of the preceding context. By projecting words in the preceding context into a lower-dimensional space, the observation of a specific word sponsors the activation of words with similar lexical distributions in the context, imbuing the model with robustness in prediction even if a particular sequence of words has not been seen. Since early demonstrations of their utility for language prediction with small-scale, synthesized linguistic corpora (Elman, 1990), an extensive literature has developed architectures to deal with web-scale natural language prediction tasks, particularly developing techniques to prevent overfitting given the massive number of parameters, and to manage their computational complexity (Mikolov, Kombrink, Burget, Černocký, & Khudanpur, 2011).

The first RNN we use here, "RNN LM," was first described in Mikolov et al. (2010), while additional performance optimizations for training can be found in Mikolov et al. (2011). We train a network with 40 hidden nodes that uses backpropagation through time (BPTT, Werbos,1990) for the two preceding timesteps. The vocabulary is limited to the 9999 highest-frequency word types, with the remainder of types assigned to an $_<$unk$_>$ type.

A.1.6. Big language model (Big LM)

One of the major challenges with recurrent neural networks is the problem of *vanishing gradients* where the error signal necessary to update connection weights in the network becomes too small to use standard gradient descent techniques. This problem becomes especially prevalent when recurrent neural networks track longer histories of preceding events. The RNN language model described above, for example, can only update weights using the model state at two preceding timesteps. One solution that has received significant attention is to use a special kind of node in the neural network's hidden layer, or *long short-term memory* (LSTM) cells (Hochreiter & Schmidhuber, 1997). While non-LSTM RNNs directly copy the previous state of the hidden layer at each timestep, the LSTM uses these special cells that regulate how information is propagated from one timestep to the next. These cells have input, output, and forgetting "gates" that regulate how the cell's state changes, such that it can maintain state for longer intervals than a typical RNN. Such networks have been widely useful in sequence prediction tasks (Gers & Schmidhuber, 2001), with particular successes in language modeling (Sundermeyer, Schlüter, & Ney, 2012; Zaremba et al., 2014).

One challenge for RNNs, left unaddressed by the use of LSTMs, is the difficulty of scaling to larger vocabularies, a necessity for modeling large naturalistic language corpora. Evaluating the network's loss function, which requires generating a probability distribution over all words, becomes prohibitively computationally expensive owing to the normalizing term in a softmax function. This has lead researchers to limit vocabularies to sizes much smaller than those typical of *n*-gram models or PCFGs, often in the range of 5000 to 30,000 word types.

Jozefowicz et al. (2015) address this limitation in their "Big LM" architecture by representing words as points in a lower-dimensional continuous space using a convolutional neural network (CNN). These models can represent words as embeddings (real-valued vectors) that capture perceptual similarity between words on the basis of shared structure, such as word lemmas or part of speech markings like *-ing* (Ling et al., 2015). The LSTM is then used to make predictions in this lower dimensional space, reducing the number of computations in the softmax function from the size of the vocabulary to the dimensionality of the CNN-derived embeddings.

Given the significant technical resources necessary to train such a model (24–36 graphics processing units) and that model hyperparameters have not been made publicly available, we use the model distributed by Jozefowicz et al. (2015). This model uses 4096-dimensional character embeddings to represent words, makes use of backpropagation through time for the previous 20 timesteps, and has two layers with 8192-dimensional recurrent state in each layer.

A.2. Training data and model fitting

All of the above language models are fit or trained using corpus data represented as written text. Ideally, we would evaluate all combinations of language models and corpora to identify how model performance reflects an interaction of model architecture and training data. However computational limitations, licensing restrictions, and limited public availability of codebases make this goal unfeasible at present. Instead, wherever possible we evaluate each language model trained on two datasets: one on a large corpus for which it is known to produce competitive or state-of-the-art results, and the other on the (publicly-available) contents of the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993).

The Penn Treebank is a standard training, validation, and test dataset in the public domain, consisting of about 930,000 word tokens in the training set. For those models that take advantage of the validation dataset for setting hyperparameters, these models have additional access to 74,000 tokens. All sentences are annotated with gold-standard parse trees, making this dataset appropriate for training the two supervised PCFG models. While the thematic content and register of the Wall Street Journal are not representative of conversational English, all language models should be equally disadvantaged by this shortcoming.

For the instance of each model architecture trained on a larger dataset, we use a variety of datasets with known strong performance in the psycholinguistics or NLP literature. For *n*-gram models we use the British National Corpus (BNC; [Dataset], 2007), an approximately 200 million word dataset commonly used in psycholinguistics. The BNC consists of material from newspapers, academic and popular books, college essays, and transcriptions of informal conversations. The BLLIP parser is trained on the Google Treebank (Bies, Mott, Warner, & Kulick, 2012) in addition to the Penn Treebank. The Google Treebank contains approximately 255,000 word tokens in 16,600 sentences with gold-standard parse trees, taken from blogs, newsgroups, emails, and other internet-based sources. Big LM is trained on the One Billion Word Benchmark of Chelba et al. (2013), a large collection of news articles. The DeepSpeech project (Hannun et al., 2014) provides a smooth 5-g model trained on a proprietary dataset including Librispeech (Panayotov, Chen, Povey, & Khudanpur, 2015) and Switchboard (Godfrey et al., 1992). Only the Penn Treebank was used in the cases of the Roark parser and the RNN LM.

A.3. Computation of lexical surprisal

Estimating surprisal — the unexpectedness of a word, in the form of its negative log probability under a predictive model — for each individual word token requires that a model produce conditional word probabilities. These probabilities are straightforwardly accessible because *n*-gram models encode these continuation probabilities directly.

The Roark parser produces an estimate of lexical continuation probability that marginalizes over the set of highest probability candidate parses given the words seen so far. The bottom-up scoring procedure used by the BLLIP means that it assigns probabilities to parse trees that generate the whole sentence, rather than successive words. For this reason, we analyze changes in sentence probability under the BLLIP, but do not use it as a predictor in analyses requiring word-level surprisal estimates.

The RNN LM yields a vector of activations over the vocabulary which is translated with a softmax function into a probability distribution. Big LM produces a probability distribution over continuations, though the softmax function is computed over the lower-dimensional representations of words. In both cases the prediction by the neural network is conditioned on the state of the model at preceding timesteps. We treat the probabilities obtained from models in the same way as the conditional probabilities from the *n*-gram models.

A.4. Computation of sentence probabilities

For all models we omit the end of sentence marker in the computation of utterance probability. We adopt this strategy because utterance-final punctuation was not collected in the Telephone game, and those models trained on datasets with punctuation assign very high surprisal to end-of-sentence markers without preceding punctuation. In the case of *n*-gram models, sentence probability is simply the product of conditional word probabilities.

Evaluating the probability of a sentence for the two PCFG models, the Roark and BLLIP parsers, requires marginalizing over the set of possible parse trees. For these two models, we sum over the probabilities of the top 50 parse trees yielded by each model. While there may be many more trees that would yield the same string of words, the first few (1–5) highest ranking parses typically account for nearly all of the aggregate probability mass for that utterance.

As first noted by Bartlett (1932), messages decrease in length over the course of serial reproduction. Shorter utterances are necessarily more probable, in that they are comprised of a smaller number of events. To investigate the effects on utterance probability separate from utterance length, we normalize the negative log probability by the number of words in each utterance. This measure, which we call "average per-word surprisal" has an intuitive interpretation in terms of bits of information. However, this measure may disguise an important difference between models in the interpretation of utterance probabilities among the PCFGs. For the non-incremental models (e.g., BLLIP), the probability reflects the set of parse trees consistent with the complete utterance. By contrast, the set of parse trees changes after observing each word under the incremental models. Garden path sentences provide a case where the latter class of models might assign high probability to the initial sequence of words "the horse raced past the barn", then a low probability to the continuation "fell," which is likely to be low probability under the set of parses in the beam. By contrast, a bottom-up, non-incremental model would assign a low probability to all expansions necessary to generate this sentence, in that it only considers the subset of parses compatible withe the full utterance.

Appendix B. Technical appendix 2: Experiment interface

Participants are directed from Mechanical Turk to an externally-hosted version of the Telephone Game. The app guides the participant through adjusting their speakers and microphone, after which participants proceed through a sequence of four practice trials. The first practice trial tests whether the participant's speakers or headphones are working properly by having them transcribe a ten-word sentence. The second practice trial tests whether a participant's microphone is working by having them listen to a recording of a sentence and repeating its content. At the beginning of this trial, the participant grants permission in the web browser for the use of their microphone. Participants are prompted to repeat the second practice trial until their recording is similar to the gold-standard transcription of that sentence, as evaluated by a normalized Levenshtein-Damerau distance of 0.2 between the gold-standard transcription and the output of an automatic speech recognition system run on their recording (described in greater detail below).

After completing the practice trials, the participant begins a series of 47 test trials, including 40 target trials and 7 randomly interspersed fillers. The 40 target trials consist of either sentences from a pool of initial recordings (described below) or the most recently-contributed recording of another participant. The seven filler utterances are the complement of the four filler utterances used for the practice trials, and are drawn from an inventory of nine audio recordings with similar properties to the initial stimuli recordings, plus two standard test sentences from the TIMIT corpus meant to elicit variation in American English dialects (Garofolo et al., 1993). If the participant flags the utterance that they heard as inappropriate in (2) or their own recording as compromised in some way (4), the trial ends early, and they progress to the next trial. Otherwise, if the recording passes the set of automated tests, then the state of the experiment is updated after each target trial to point to that newest recording as the appropriate stimulus for the next participant.

An example screenshot is presented in Fig. 3. We now present in greater detail the five steps an experimental trial and the automated filters outlined above.

1. Listening to a recording. The participant clicks a button to start the trial, which triggers a three-minute audio timer that remains visible to the user through Step 5, below. This three minute timer ensures that the web experiment (given the dependencies between participants) is not blocked by an inactive participant; if time expires, another participant receives this stimulus. Participants receive friendly feedback to encourage them to complete trials more quickly if this timer expires. Besides the button to start playing the audio recording, there are no controls to pause, repeat, or move location in the recording. The audio is presented embedded in noise recorded from a coffee shop (average SNR dB across recordings = −6.8). Pilot experiments established that this naturalistic source of noise introduced a high rate of edits without introducing significant participant attrition, as was observed with similar amplitude white noise. Each utterance has between 500 and 1000 ms of noise-only padding preceding and following the speech content.

2. Flagging the upstream recording. Though the participant cannot pause or move within the recording that they hear, they may flag the recording at any time. If the participant chooses to flag a recording, they are then asked to choose one of a set of provided reasons: Contains speech

errors, Speech starts or stops abruptly, Contains obscenities, and Other. Upon choosing Other, the participant could provide a free-form text response. The trial ends early if any of these options are chosen.

3. Recording best guess of what was heard. If the participant does not flag the audio recording, they are immediately prompted to record their best estimate of what was said, with the specific prompt of "Repeat the sentence you just heard as best you can." After clicking the Record button, the waveform for the recording is drawn in real time. The participant may listen to and re-record their response as many times as desired. If a recording is more than eight seconds long, the participant is prompted to record again. When finished, the participant clicks "Submit." Because the HTML5 audio recording specification leaves the choice of sampling rate and bit depth to the client, all recordings are normalized upon receipt by the server to 16 kHz, 16 bit PCM WAV files. The acoustic properties of the recording environments vary between participants in that each participant records their responses in an uncontrolled environment.

4. Self-flagging the new recording. After submitting a recording, a participant is given the opportunity to "self-flag" it in case of a speech error or other problems such as unexpected background noise. If the participant chooses to self-flag the recording, they are prompted to provide a reason, with the same set of candidate reasons as the upstream flagging procedure in Step 2, above. If a participant self-flags, the trial ends early.

5. Transcribing the new recording. Finally, if the participant has submitted a recording and has attested that it is of good quality, they are then prompted to provide a written transcription thereof. The inclusion of punctuation or of a misspelled word (as determined with the Linux utility Aspell) returns an error message to the participant, who may then edit the transcription and resubmit. After submitting, the trial ends. The participant is then provided with a button entitled "Click to play next audio recording"; because there is no timer on this screen, they may pause between each trial.

B.1. Automated filters

After the participant submits the written transcription, they advance to the start button for the next trial. The web application asynchronously applies a set of automated filters to the audio file and the transcription on the server. We employ these filters to determine if a recording is of sufficient quality to be used as input to other downstream participants. In principle, these filters could be implemented with human participants, but using automated tools allows us to direct participants' effort (and commensurate compensation) towards data collection. Of note from a design and data analysis perspective, these filters *must* be applied at the time of collection: because future participants hear responses from earlier participants, responses must be filtered in real time to maintain the continuity and integrity of the recording chains. These filters include:

- Is the file silent?
- Is the transcription provided by the new participant between 20% longer and 20% shorter (in terms of the number of non-space characters) than the transcription of the input sentence they received? If utterances become too short, they become difficult to characterize with language models.
- Is the transcription provided by the new user more than 2 words longer or 2 words shorter than the transcription of the input sentence? This follows the same logic as above.

A second set of tests pertains to the audio quality and the intelligibility of the recording. For this we use an automatic speech recognition system to generate a transcription of the newly-recorded audio file. Specifically, we use one of the most advanced publicly-available automatic speech recognition systems, DeepSpeech (Hannun et al., 2014), to check:

- Is the DeepSpeech-generated transcription of the participant's audio file similar to the transcription they provided? This guards against the possibility that a user would provide an acceptable transcription, but an unrelated audio file (e.g., one filled with obscenities).
- Is the DeepSpeech-generated transcription of the participant's audio file similar to the transcription provided by the upstream participant? This prevents the introduction of material like "I didn't hear the last sentence."

The above checks are operationalized by testing whether the normalized Levenshtein-Damerau distance (Navarro, 2001) between the strings in question is less than a precomputed threshold. We use the threshold of 0.58, arrived at by computing the normalized Levenshtein distance between each sentence in a large corpus and a number of candidate transcriptions, including the correct one and several unrelated foils. On this test corpus with highly dissimilar sentences, this threshold yields a negligible false positive rate; in practice, this threshold is extremely permissive (allows changes) and flags only highly deviant recordings. The length filter was motivated by the finding in pilot work that utterances decreased in length very rapidly without such constraints, yielding many 1–3 word utterances of questionable linguistic status.

If a recording fails any of the above tests, the participant receives feedback at the end of the following trial. This asynchronous evaluation allows for efficient speech recognition (which is computationally costly and requires the use of a graphics processing unit on the server) and prevents participants from having to wait for the web application to recognize and validate their responses. Utterances that are rejected by any of these filters are retained in that they are potentially relevant to a number of research questions outside of the scope of the current study. If a recording passes the above tests, the recording is combined with a randomly selected interval of cafe background noise (with the same acoustic properties as above) so that it may be used as a stimulus for later participants. The implications of flagging the upstream stimulus, self-flagging, and automated filtering for the stimuli heard by later participants are described below in relation to the process of serial reproduction.

After submitting a response for the final stimulus, participants are prompted to take an optional demographic survey detailing age, gender, level of education, current geographical location, proficiency with English, and information regarding previous residence for the purpose of dialectal analyses (outside of the scope of this work).

Appendix C. Technical appendix 3: Experiment transmission structure

The succession of stimuli and responses (the latter constituting stimuli for later participants) can be conceived of in terms of a directed acyclic graph, or DAG. Considering the succession of recordings for each sentence as a graph—a collection of nodes representing recordings and edges representing the history among participants—allows us to concisely represent the data collected in the course of the experiment and to operationalize the logic for both automated and participant-based flagging of recordings.

Per the specification of the interface above, a recording takes one of five possible states: accepted by fiat because it is an initial stimulus ("protected"), provisionally accepted ("accepted"), flagged by a downstream participant ("downstream-flagged"), flagged by the participant who recorded the sentence ("self-flagged"), or flagged by one of several automated methods ("auto-flagged"). When a participant starts a test trial indexed by a particular initial sentence, they are provided with either the most recent accepted recording or the initial recording itself (if no previous recordings have been accepted). In other words, if a participant flags the input recording they heard, the following participant will then hear the previously

accepted recording in the graph for that sentence. The sequence of recordings appropriate for analysis, or *recording chain*, is then the initial sentence recording and the subsequent succession of accepted recordings. Flagged and self-flagged recordings comprise the complement of the nodes in the graph. We follow the convention established in the iterated learning literature of referring to the sequential position of a recording within a chain as its *generation*, though note that a participant in this setup may contribute recordings for different stimuli at different generations, unlike most iterated learning experiments.

The process by which participants are assigned to stimuli can then be considered in terms of *threading*. Each participant must provide recordings for each of the 40 test sentences. Because of the need for strictly successive recordings, only one participant at a time may listen to and record a response to the same recording. We implement these constraints by implementing a *mutex*, a data structure that limits concurrent access to a resource, in this case the recording chains. In combination with multiple independent chains for each stimulus, this setup allows participants to listen and record sentences continuously while maintaining a strictly sequential relationship between the contributed recordings. The three-minute timer on each trial means that the controller (the web app) will re-assign a stimulus to another participant if a response is not recorded within three minutes. This setup also means that the order in which a participant contributes to each of the recording chains is randomized across participants.

These dynamics mean that any recording may be flagged and removed for the duration of the experiment, even if another participant records a downstream utterance. For example, it is possible that sentence s_1 recorded by participant p_1 is provisionally accepted, and that a succeeding participant p_2 records a downstream repetition, s_2. If, however, p_3 flags s_2, then p_4 will hear s_1, and may in principle flag that recording. We find that such cases of retroactive flagging (and loss of analyzable data) are relatively rare, but that this mechanism provides an automated and scalable method to produce chains of interpretable utterances appropriate for analysis. This architecture also means that if two participants p_1 and p_2 are progressing through the experiment at the same time, then p_1 may provide the stimulus recording heard by p_2 for some sentences and p_2 may provide the stimulus recording heard by p_1 in others.

Appendix D. Technical appendix 4: Analysis of flagged recordings

15.3% of participant-contributed audio recordings were flagged by downstream participants as too noisy, ungrammatical, or otherwise unintelligible. Specifically, 4.5% were classified by participants as "Contains speech errors" (ungrammatical), 3.5% as "Speech stops or starts abruptly / speech sounds cut off," and the remainder were flagged for a variety of user-specified reasons regarding audio quality, e.g., "too hard to hear," "drowned out," "can't hear it."

One possibility that would complicate interpretation of the results presented here would be if participants who flag a high proportion of recordings behave differently (i.e., make fewer or more edits to what they hear) than those who flag a low proportion of recordings. This could arise if some characteristic of the participant affected both the flag rate and the edit rate. In principle, a participant's robustness to noise (e.g., robustness in speech recognition related to age or language experience) could affect both: a participant with better hearing may flag few sentences and produce a high proportion of matches. However, we do not find a statistically significant correlation between the proportion of words changed (deleted or substituted) and the proportion of input sentences flagged by a participant (Fig. 11, $R^2 = 0.006$, $p = 0.157$). This suggests that participants who flag many recordings do not behave differently than those who flag few recordings.
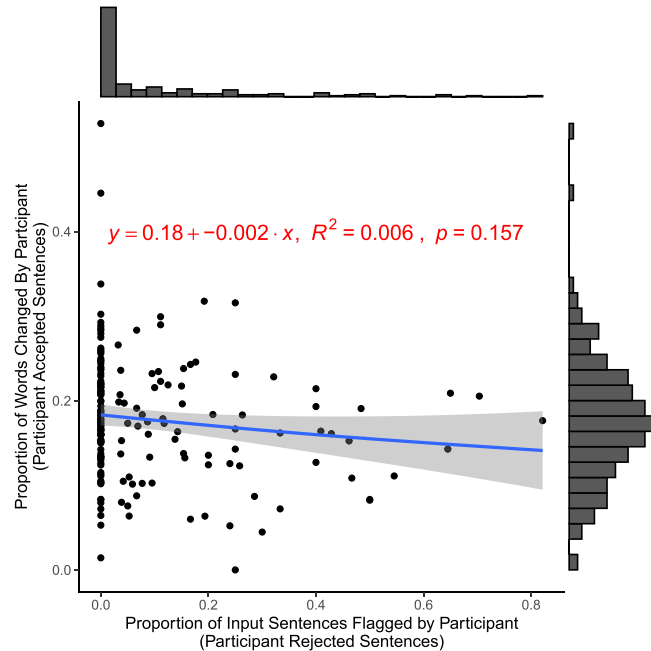

Fig. 11. Relationship between participant flag rate (i.e., number of recordings rejected) and participant edit rate (proportion of words changed among accepted recordings). Most participants change around 1 in 5 words that they hear, and flag none of the recordings that they hear. The non-significant correlation suggests that there is no relationship between flag rate and edit rate by participant.

Appendix E. Technical appendix 5: Similarity of language models

In the main text we present the pairwise similarity of all models trained on the Penn Treebank in order to isolate the contribution of model architecture (vs. training data) to surprisal estimates. Here we present the same measure of pairwise similarity, but for all 11 models (Fig. 9). This reveals

that *n*-gram models trained on the PTB constitute one well-defined cluster separate from all other models. Another cluster emerges among Kneser-Ney trigrams and 5-g trained on large datasets and the Big LSTM trained on the One Billion Word Benchmark. The BLLIP PCFGS constitute a 2-model cluster. The Roark Parser, the RNNLM and unigram probabilities from the BNC constitute a separate set from the BLLIP PCFGs, but pattern more closely with them than the large models or the PTB-trained *n*-gram models.
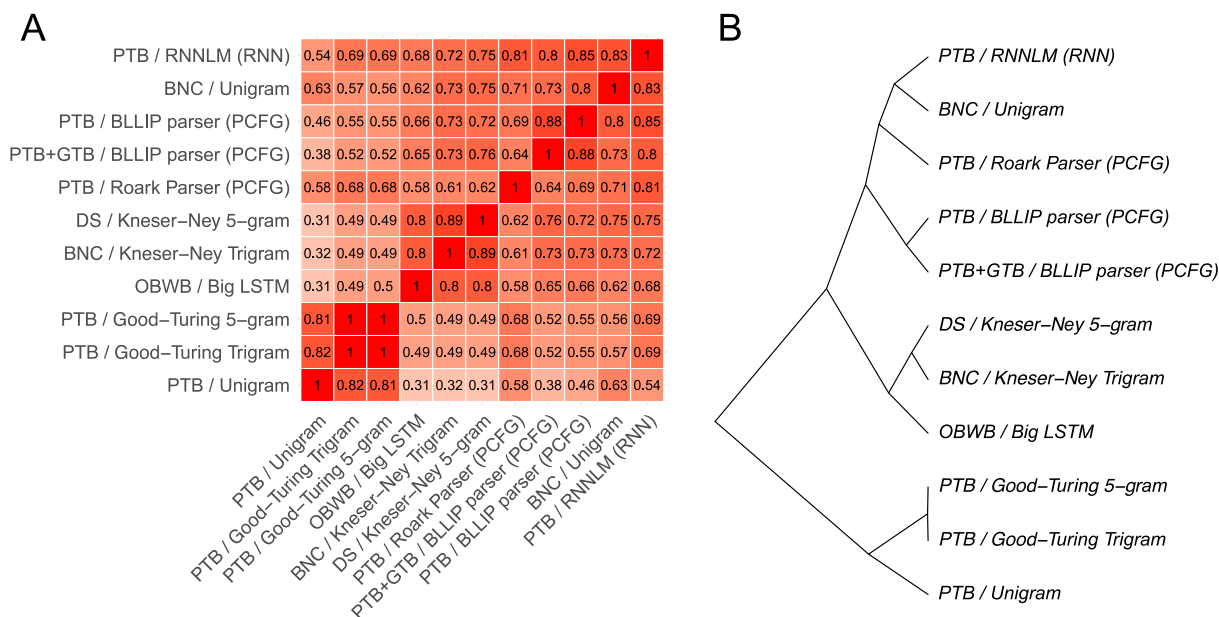

Fig. 12. Model similarity across 11 language models. *A*. Spearman's rank correlation for pairwise combinations of sentence probabilities across the 3193 sentences from the experiment across all models. *B*. The resulting dendrogram, derived from the rank correlations using Ward's method.

References

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

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Austerweil, J. L., Abbott, J. T., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. In *Advances in Neural Information Processing Systems* (pp. 3041–3049).

Bartlett, F. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Benichov, J., Cox, L., Tun, P., & Wingfield, A. (2012). Word recognition within a linguistic context: Effects of age, hearing acuity, verbal ability and cognitive function. *Ear and Hearing, 32*(2), 250.

Bies, A., Mott, J., Warner, C., & Kulick, S. (2012). *English Web Treebank*. Philadelphia, PA: Linguistic Data Consortium.

Broadbent, D. (1967). Word-frequency effect and response bias. *Psychological Review, 74*(1), 1.

[Dataset] Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

[Dataset] Brysbaert, M., Warriner, A., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 173–180).

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring Progress in statistical language modeling. In *arXiv preprint*. arXiv:1312.3005.

Chen, S., & Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language, 13*(4), 359–393.

Connine, C., Blasko, D., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraint. *Journal of Memory and Language, 30*(2), 234–250.

Connine, C. M., & Clifton, C. (1987, May). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology. Human Perception and Performance, 13*(2), 291–299.

De Marneffe, M., MacCartney, B., & Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *, Vol. 6. Proceedings of LREC* (pp. 449–454).

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193–210.

Demberg, V., & Keller, F. (2009). A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *, 31. Proceedings of the Annual Meeting of the Cognitive Science Society*.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–22.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint*. arXiv:1810.04805.

Du Bois, J., Chafe, W., Meyer, C., Thompson, S., & Martey, N. (2000). Santa Barbara corpus of spoken American English. In *CD-ROM*. Philadelphia: Linguistic Data Consortium.

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. (2016). Recurrent neural network grammars. In *Proceedings of NAACL*.

Edmiston, P., Perlman, M., & Lupyan, G. (2018). Repeated imitation makes human vocalizations more word-like. *Proceedings of the Royal Society B: Biological Sciences, 285*(1874), 20172709.

Egan, J. (1948). Articulation testing methods. *Laryngoscope., 58*(9), 955–991.

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

Fine, A., Jaeger, T., Farmer, T., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One, 8*(10), Article e77661.

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 61–69).

Frank, S., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science, 22*(6), 829–834.

Frank, S., Bod, R., & Christiansen, M. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences, 279*(1747), 4522–4531.

Frank, S., & Christiansen, M. (2018). Hierarchical and sequential processing of language: A response to: Ding, Melloni, Tian, and Poeppel (2017). Rule-based and word-level statistics-based processing of language: Insights from neuroscience. *Language, Cognition and neuroscience, 33*(9), 1213–1218.

French, N., & Steinberg, J. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America, 19*(1), 90–119.

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *, Vol. 1. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1* (pp. 688–698).

Futrell, R., & Levy, R. (2019). Do RNNs learn human-like abstract word order preferences?. In *, Vol. 2. Proceedings of the Society for Computation in Linguistics (SCiL)* (pp. 50–59).

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gahl, S., & Strand, J. (2016). Many neighborhoods: Phonological and perceptual neighborhood density in lexical production and perception. *Journal of Memory and Language, 89*, 162–178.

Gale, W., & Sampson, G. (1995). Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics, 2*(3), 217–237.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., & Pallett, D. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1–1.1. In *NASA STI/Recon technical report, 93*.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*(1), 1–58.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.

Gers, F., & Schmidhuber, J. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks, 12*(6), 1333–1340.

Gibson, E., Bergen, L., & Piantadosi, S. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of National Academy of Science USA, 110*(20), 8051–8056.

Gibson, E., Desmet, T., Grodner, D., Watson, K., & Ko, D. (2005). Reading relative clauses in english. *Cognitive Linguistics, 16*(2).

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes, 14*(3), 225–248.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *, Vol. 1. IEEE International Conference on Speech, and Signal Processing, ICASSP-92* (pp. 517–520).

Goldwater, S., Jurafsky, D., & Manning, C. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication, 52*(3), 181–200.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on Cognitive Modeling and Computational Linguistics (cmcl 2018)* (pp. 10–18).

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 369–376).

Griffiths, T., Christian, B., & Kalish, M. (2008). Using category structures to test iterated learning as a method for revealing inductive biases. *Cognitive Science, 32*, 68–107.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science, 31*(3), 441–480.

Griffiths, T., Kalish, M., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363*(1509), 3503–3514.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *, Vol. 1. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1195–1205).

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001* (pp. 1–8).

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., & others. (2014). Deep speech: Scaling up end-to-end speech recognition. In *arXiv preprint.* arXiv:1412.5567.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation* (pp. 187–197).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of english words. *The Journal of the Acoustical Society of America, 29*(2), 296–305.

Hudson, P., & Bergman, M. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language, 24*(1), 46–58.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295–1306.

Jaeger, H. (2001). The echo state approach to analysing and training recurrent neural networks-with an erratum note. In *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 148(34)* (p. 13).

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning* (pp. 2342–2350).

Jurafsky, D., & Martin, J. (2009). *Speech & language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Prentice-Hall.

Kalikow, D., Stevens, K., & Elliott, L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61*(5), 1337–1351.

Kamide, Y., Altmann, G., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133–156.

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*(2), 102–110.

Kirby, S., Dowman, M., & Griffiths, T. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences, 104*(12), 5241–5245.

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

[Dataset] Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990.

Lash, A., Rogers, C., Zoller, A., & Wingfield, A. (2013). Expectation and entropy in spoken word recognition: Effects of age and hearing acuity. *Experimental Aging Research, 39*(3), 235–253.

Levesque, H., Davis, E., & Morgenstern, L. (2011). The Winograd schema challenge. In *, Vol. 46. AAAI spring symposium: Logical formalizations of commonsense reasoning* (p. 47).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009, Dec). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings. National Academy of Sciences. United States of America, 106*(50), 21086–21090.

Ling, W., Luís, T., Marujo, L., Astudillo, R., Amir, S., Dyer, C., … Trancoso, I. (2015). Finding function in form: Compositional character models for open vocabulary word representation. In *arXiv preprint.* arXiv:1508.02096.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics, 4*, 521–535.

Linzen, T., & Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 692–697). Austin, TX: Cognitive Science Society.

Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication, 22*(1), 1–15.

Luce, P., & Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1–36.

Luce, P., Pisoni, D., & Goldinger, S. (1990). *Similarity neighborhoods of spoken words*. The MIT Press.

Marcus, M., Marcinkiewicz, M., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY, USA: Henry Holt and Co., Inc.

Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1–2), 71–102.

McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86.

McQueen, J. M., & Huettig, F. (2012, Jan). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America, 131*(1), 509–517.

Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Mikolov, T., Kombrink, S., Burget, L., Černocký, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *2011 IEEE international conference on Acoustics, speech and signal processing (ICASSP)* (pp. 5528–5531).

Miller, G., Heise, G., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41*(5), 329.

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

Morrison, C., & Ellis, A. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology, 91*(2), 167–180.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*(2), 165–178.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR), 33*(1), 31–88.

Norris, D., & McQueen, J. (2008, Apr). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357–395.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience, 31*(1), 4–18.

Owens, E. (1961). Intelligibility of words varying in familiarity. *Journal of Speech and Hearing Research, 4*(2), 113–129.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Petrov, S., & Klein, D. (2007). Learning and inference for hierarchically split PCFGs. In *, Vol. 22. Proceedings of the National Conference on Artificial Intelligence* (p. 1663).

Popovic, M., & Ney, H. (2007). Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 48–55).

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics, 27*(2), 249–276.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 324–333).

Rohde, H., & Ettlinger, M. (2012). Integration of pragmatic and phonetic cues in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(4), 967–983.

Sanborn, A., Griffiths, T., & Shiffrin, R. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology, 60*(2), 63–106.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin, 2*, 110–114.

Savin, H. (1963). Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America, 35*(2), 200–206.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(July 1928), 379–423.

Shannon, C. (1951). Prediction and entropy of printed English. *Bell Systems Technical Journal, 30*, 50–64.

Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In , *33. Proceedings of the Annual Meeting of the Cognitive Science Society*.

Smith, N., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319.

Stabler, E. (2004). Varieties of crossing dependencies: Structure dependence and mild context sensitivity. *Cognitive Science, 28*(5), 699–720.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In , *Vol. 2. Proceedings International Conference on Spoken Language Processing* (pp. 901–904).

Storkel, H., Armbrüster, J., & Hogan, T. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research., 49*(6), 1175–1192.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268* (5217), 1632–1634.

Theunissen, M., Swanepoel, J., & Hanekom, D. W. (2009). Sentence recognition in noise: Variables in compilation and interpretation of tests. *International Journal of Adiology, 48*(11), 743–757.

Treisman, M. (1978). A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review, 85*(6), 525.

Van Schijndel, M., & Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1597–1605).

Vincent-Lamarre, P., Massé, A. B., Lopes, M., Lord, M., Marcotte, O., & Harnad, S. (2016). The latent structure of dictionaries. *Topics in Cognitive Science, 8*(3), 625–659.

Vitevitch, M., & Luce, P. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3), 374–408.

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE, 78*(10), 1550–1560.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The microsoft 2017 conversational speech recognition system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5934–5938).

Xu, J., & Griffiths, T. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology, 60*(2), 107–126.

Yap, M., Balota, D., Sibley, D., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the english lexicon project. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 53–79.

[Dataset] Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review, 15*(5), 971–979.

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. In *arXiv preprint. arXiv:1409.2329*.

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

Ziegler, D., Stiennon, N., Wu, J., Brown, T., Radford, A., Amodei, D., … Irving, G. (2019). Fine-tuning language models from human preferences. In *arXiv preprint.* arXiv:1 909.08593.

Ziegler, J., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language, 48*(4), 779–793.