N-gram probability effects in a cloze task.

Cyrus Shaoul

Department of Linguistics, University of Tübingen

R. Harald Baayen

Department of Linguistics, University of Tübingen

Department of Linguistics, University of Alberta

Chris F. Westbury

Department of Psychology, University of Alberta

This version prepared on September 11, 2014

Author Note

Please send correspondence to Cyrus Shaoul: cyrus.shaoul@uni-tuebingen.de

Abstract

What knowledge influences our choice of words when we write or speak? Predicting which word a person will produce next is not easy, even when the linguistic context is known. One task that has been used to assess context dependent word choice is the fill-in-the-blank task, also called the cloze task. The cloze probability of specific context is an empirical measure found by asking many people to fill in the blank. In this paper we harness the power of large corpora to look at the influence of corpus-derived probabilistic information from a word's micro-context on word choice. We asked young adults to complete short phrases called *n*-grams with up to 20 responses per phrase. The probability of the responded word and the conditional probability of the response given the context were predictive of the frequency with which each response was produced. Furthermore the order in which the participants generated multiple completions of the same context was predicted by the conditional probability as well. These results suggest that word choice in cloze tasks taps into implicit knowledge of a person's past experience with that word in various contexts. Furthermore, the importance of *n*-gram conditional probabilities in our analysis is further evidence of implicit knowledge about multi-word sequences and support theories of language processing that involve anticipating or predicting based on context.

Keywords: cloze probability, production, formulaic language, multi-word expressions, n-grams

N-gram probability effects in a cloze task.

When we read or hear language, we are concurrently processing what is perceived and predicting what will happen next. Landmark studies of anticipation in language perception have allowed us to understand how we react to expected versus unexpected words (DeLong, Urbach, & Kutas, 2005; Kutas & Hillyard, 1984). Bar (2007) goes as far as to say that "memory-based predictions/association-based predictions" are one of the unifying principles of the brain (p. 280). In this paper we will look beyond the influence of concurrent prediction in linguistic comprehension and investigate the use of anticipation in linguistic production. In particular, we will expose the participants to incomplete fragments of language and invite them to complete them, a type of cloze task (Taylor, 1953). We hypothesize that contextual, probabilistic prediction will play a part in completing these phrases because there is strong evidence that we anticipate upcoming words during sentence processing (Kamide, 2008). A similar process may be taking place during the completion of a cloze task. We propose that the word chosen in a cloze task is the first one provided by the language prediction system, constrained by linguistic context.

This raises larger theoretical questions: what is a language prediction system and on what does it base its predictions? The fundamental theoretical shift that is happening in psycholinguistics is a shift towards theories of language that posit the existence of very simple statistical learning processes at the heart of language acquisition and processing. The theories proposed by Frank and Bod (2011), Elman (2011) and Baayen, Milin, Djurdjevic, Hendrix, and Marelli (2011) rest on the ideas that complex language behavior emerges from a very simple process, and that the statistical structure of the input is sufficient to explain our linguistic abilities. Looking at our language prediction system from this vantage point, all that is needed to predict words in a language stream is exposure to the statistical structure of a language. Using that probabilistic information, a person can begin to predict and anticipate words from context. However, there is an alternative possibility, one that cannot be dismissed outright. That position would be one in which prediction from context is not the dominant influence — sources of information from the micro-context might be negligible and sources of

information other than the micro-context might be the dominant predictor of performance in an *n*-gram cloze task. This is the theoretical question we hope to address in this paper: how does contextual probabilistic information take part in the process of choosing a word in a cloze task?

A first step in getting closer to answering this question is to define the term *cloze task*, first proposed by Taylor (1953). The word *cloze* was chosen because it hearkened back to the gestalt principle of *closure* in the visual sense — if people are exposed to a partial drawing or photograph they perceive a whole by filling in the missing information (Ellis, 1999). The linguistic cloze task was originally devised as a method to measure the readability of texts. A certain number of words in a text were replaced with blanks, and naive readers were asked to fill in the blanks with the words that made sense in the contexts given. If the accuracy on the cloze task was high, the text was considered very readable, implying that texts that were more predictable were easier to read.

In 1951, long before the coining of the term *cloze task*, Claude Shannon calculated the entropy of English using a letter cloze task and a small corpus (Shannon, 1951). He was not addressing psycholinguistic questions, but his influential paper spurred psychologists to look at information, entropy and redundancy in language. One of the first studies to use the cloze task as a window into the psychology of language was carried out by Fillenbaum, Jones, and Rapoport (1963). They deleted words from transcripts of spoken English at 5 different rates (every 2^{nd}, 3^{rd}, 4^{th}, 5^{th} or 6^{th} word) and measured subjects' accuracy for all of these deletion rates. They also counted how often the word was replaced with a word different from the original word but of the same lexical class. They found that performance was better than chance at all the different deletion rates, showing the powerful impact of context on performance. Despite this, there was a great deal of item variability in the accuracy data, with some contexts enabling greater item accuracy than others. The key discovery in this study was that context strongly constrains what subjects will produce when asked to fill in the blanks, paving the way for computational studies of cloze task completion.

Once the computational resources became available to build digital corpora, to find the word frequencies in these corpora (Francis & Kucera, 1982), and to calculate probabilistic

measures from these corpora, researchers began to look at the cloze task as an experimental behavior that could be predicted. Up to this point, human rating norms were the only way to calculate the cloze probability. The second wave of studies attempted to understand the results of cloze experiments based on the distributional properties of language as observed in corpora.

Finn (1977) was the first to use the ideas of entropy and information from Shannon (1948, 1951) to analyze a cloze task. He used data from a cloze experiment done by Bormuth (1966) to calculate the amount of information in the words that were replaced by blanks, taking into account both the orthographic frequency of the words and the number of completions given by subjects. The term *information* in this context refers to the Shannon information, which is proportional to the predictability of the word. Less predictable words have more information. He found that if the words replaced by blanks had high amounts of information (and were low frequency) they were less likely to be correctly chosen in the cloze task. Low information/high frequency words were more likely to have responses that matched the original word. The reason for this was that the constraints on the high frequency words (which were mostly closed class words) by the context of content words (open class words) was strong. In contrast, the constraints placed by open class words (such as *cat*) on the preceding closed class word (*a cat?*, *the cat?*, *some cat?*) were found to be weak. This attempt to understand the relevant sources of variation in the cloze task was limited by the relatively low quality of the lexical probability data that was available at the time, but the theoretical implications of this work are still relevant today.

Beattie and Butterworth (1979) did ground-breaking work on the interactions between lexical frequency and cloze probability. They looked at pauses of more than 200 ms in spontaneous speech and noted the frequency of the words after the pauses. Judges were then asked to fill in a transcription of the speech data where the words after the pauses had been replaced by blanks. Beattie and Butterworth found that the corpus frequency of the missing word was correlated with the cloze probability given by the number of judges choosing that word. As in all the other research on the cloze task, the cloze probability of a word in a certain context was defined as the probability that a panel of judges will choose that word when asked to fill in the blanks. If 41 out of 100 judges pick a certain word in that context, the cloze

probability of that specific response is said to be 0.41. This was the only way to analyze data from experiments that used the cloze task, by comparing the new results to cloze norms. Two problems exist with this definition of cloze probability: 1) The inconsistency of human judges can render cloze norms too noisy to be useful, and 2) When there are a large number of meaningful completions possible, a panel of judges, each providing its solution to the cloze riddle, cannot provide enough completions to give probabilities for all possible meaningful completions. The theoretical significance of this correlation between orthographic frequency and cloze probability was not appreciated at the time, but this result was evidence for frequency effects in the process of generating a response in a cloze task.

McKenna (1986) proposed that the responses in a cloze task could be predicted based on the associative strength of the words in their semantic categories. He built a computational simulation that searched through lists of word categories from the norms given by Battig and Montague (1969). His explanation of the process of completing a cloze was that "At the semantic level, a single schema associated with a key constraint is used as the basis of a memory search. Sub-schemata in the form of individual words are examined in the order of associative strength until a word that meets additional constraints (if there are any) imposed by context is encountered."(McKenna, 1986, p.493) This schema model is far removed from our theoretical stance, but the core concepts of memory retrieval within constraints are relevant to our approach.

There has also been research on a verbal production variant of the cloze task that directly addresses the interplay between frequency and contextual constraint. Griffin and Bock (1998) asked participants to complete a sentence with a word that was cued with a drawing. They manipulated the spoken frequency of the word cued by the picture and the amount of constraint created by the context. They found that when the context was highly constraining the effect of frequency on the word chosen was diminished. When the context was less constraining or even incongruous, high frequency words were chosen over low frequency words. This was evidence that both verbal and written cloze tasks are influence by orthographic frequency of the response word.

A recent study by Smith and Levy (2011) asked subjects to complete sentence-initial

4-grams and compared their responses with the most frequent continuations from the Google Web1T corpus. Encouragingly, they found that subjects' responses were sensitive to corpus probabilities, and the responses from the subjects were more variable than the corpus. They then asked a different group of subjects to read some of these 5-grams that they found in the first experiment, and calculated reading times for these critical 5th words. They found that a model without control covariates showed an effect for cloze probability, but once the covariates (lexical frequency, concreteness, contextual diversity and length) were added, these effects were no longer significant. Smith (2011) approached the question again using ideas from statistical smoothing. He found that probabilities from the Google Web1T and Google Books datasets were predictive of cloze probabilities, but that there was unexplained variability not accounted for my corpus measures. The question of what drives word choice in a cloze task is as of yet unresolved, but there was a tantalizing possibility proffered: that real-word experience and prediction from experience is behind it.

Cloze norms

Can extant databases of cloze probabilities help us answer our questions? The value of cloze probabilities to experiment designers in many fields is high. There have been several popular lists that have been used to help design psycholinguistic experiments where the predictability of a word in context is critical such as event-related potential (ERP) experiments that measure expectancy violations (Kutas & Hillyard, 1984). Bloom and Fischler (1980) produced one of the first set of norms, a set of 329 sentences. Recently a larger set of 498 norms has been released by Block and Baldwin (2010), and the N400 effect was validated for these contexts. The goal of these norms is to find the most highly semantically constrained contexts possible, such as the sentence *She could tell he was mad by the tone of his* _____, which their subject completed with *voice* 99% of the time. 400 of the 498 sentences have a top completion that is dominant (defined as a cloze probability between .67 to .99). They achieved their goal of finding and norming many highly constrained sentences.

These norming studies did not delve into the sources of constraint, nor did they try to understand the source of the variability in their data. The issue with the analysis of cloze

norms is the generalizability of the data. To know the cloze probability of an arbitrary piece of language it would be necessary to collect more human judgments, impractical for large amounts of text.

Are sentences and paragraphs the only type of stimuli that make sense in a cloze task? In many types of reading activities we are not exposed to the context of a whole sentence. For example, reading a narrow column of text will inevitable cause a group of three or four words to be cut off the end of a sentence, and the reader will have to look across and down to the next line to complete the process of reading the sentence. These short groups of words are what we will call *n*-grams. There has been a recent growth in the number of studies investigating the processing of *n*-grams. Arnon and Snider (2010) showed that more frequent *n*-grams were read faster than less frequent *n*-grams (but see Baayen, Hendrix, and Ramscar, 2013 for a different interpretation of their results). Tremblay and Tucker (2011) measured how long it took subjects to read an *n*-gram and also how long it took them to produce the *n*-gram. Probabilistic predictors were able to explain much of the variability in reading the *n*-grams aloud. Sprenger and van Rijn (2013) replicated this effect in the production of Dutch expressions for time. Arnon and Cohen Priva (2013) found reduced phonetic durations for higher frequency *n*-grams and Shaoul, Westbury, and Baayen (2013) found evidence for *n*-gram probability effects in subjective frequency tasks [1].

The Cloze task in relation to other psycholinguistic tasks

In this study we will look at how *n*-gram statistics from a corpus can predict word choice when completing a linguistic fragment, an *n*-gram. The *n*-grams that we will use are 3 or 4 words long, much shorter than sentences used in most previous research on the cloze task. The shorter length reduces the amount of context and increases the number of possible completions of an *n*-gram cloze task. *N*-gram stimuli and sentence stimuli are not comparable, forcing us to look at other types of psycholinguistic paradigms for theoretically relevant work.

One type of research that may be germane is work on free association (Nelson, McEvoy, & Dennis, 2000). The classic free association task is to provide a cue (such as *bread*) to many

[1] See Shaoul and Westbury (2011) for a review of recent *n*-gram processing studies.

subjects and count the frequency of the various responses (such as *butter*). Nelson, McEvoy, and Dennis (2000) characterize free association as a memory task, and point to evidence from cued recall experiments (Nelson, McKinney, Gee, & Janczura, 1998) and false memory experiments (McEvoy, Nelson, & Komatsu, 1999) that support the predictive power of associative strength in memory tasks. They conclude that the probability of a response being given is a manifestation of its associative strength in memory, noting that the "strength of a response reflects the number of its instances in memory, with stronger associations reflecting larger numbers of instances" (Nelson, McEvoy, & Dennis, 2000, p.896).

How similar is the free word association task to the *n*-gram cloze task? The context of a single word is less than that of a 3-gram, but the process of producing "the first word that comes to mind" may be similar for both free word-word association and free *n*-gram-word association. When presented with a short *n*-gram, such as *third most popular*, it is conceivable that a pool of associates emerges, and that the strongest associate is chosen first. One question we hope to address in this paper is what information is producing this emergence of candidates and what factors determine the order in which they are selected from this pool.

Memory researchers that use words as stimuli sometimes ignore the psycholinguistic nature of their research, but it is undeniable that lexical factors influence memory experiments. If we look at the cloze task is as a strategic test of memory retrieval we can gain another theoretical perspective.

In many studies the statistical properties of language have been presented as confounds or nuisance variables because they have been found to influence memory tasks that use words as stimuli. Instead of trying to eliminate probability effects by matching and counter-balancing average probability, Criss, Aue, and Smith (2010) systematically manipulated the contextual variability and orthographic frequency of the cues and targets in a paired associated cued recall task. The contextual variability of the cue words in their experiment was defined as the number of different documents in which words appeared in a corpus. They found that high probability targets were recalled better, independent of the context variability and probability of the cue. The probability of the cue did not influence recall, but cues with low context variability (seen in a smaller number of documents) had the

effect of improving recall performance.

Pickering and Garrod (2007) argue that language comprehension involves making simultaneous predictions at different linguistic levels and that these predictions are generated by the language production system. They report increased muscle activity in the lips and tongue when listening to speech but not when listening to non-speech noise as evidence for an automatic forward-modeling system. In their system comprehension and production are tightly coupled and the motor production system facilitates language comprehension just as the comprehension system facilitates production.

Pickering and Garrod (2007) are persuasive when they argue that language knowledge, stored in memory, exerts its contribution to behavior by way of predictions. The advantages of a constant simulation or *emulation* of the external is becoming a foundational idea in psycholinguistics (Willems & Hagoort, 2007). In looking at how we process and complete *n*-grams, an emulation framework may help us explain how completions are chosen. To understand how written production systems choose a word, we will look at what kind of information could be used by a hypothetical emulator to predict an upcoming word in a stream of words.

The context found in an *n*-gram puts certain constraints on what words can fill an empty slot. Semantic and syntactic constraints are the most studied constraints. The constraint that we hope to add to this list is the constraint of memory: if there is an implicit or explicit memory of seeing or writing an *n*-gram, that *n*-gram should be accessible during the completion of a cloze task. Conversely if there are few or no memory traces for an *n*-gram, the likelihood of predicting a completion using that *n*-gram is much lower. We will assume in these experiments that the corpus-based measures of the *n*-grams are correlated with the participants' language experience, and that the effects of the constraints of experience will be seen in the responses the participants give.

New tools have become available in the current era of psycholinguistics that allow for new approaches to the question of constrained language production. In particular, corpus data-driven research has begun to let us investigate the probabilistic nature of language in new ways. One asset is the enormous corpora of electronic texts, and the computational resources

to calculate lexical statistics across these corpora. A case in point is the Google Web1T corpus (Brants & Franz, 2006) that we will be using exclusively as a source of probabilistic information about language in this paper. It is a corpus made up of one trillion words of English web page text. Using the immense computing infrastructure at Google, the authors were able to count all the occurrences of all the word groups or *n*-grams from two to five words long (but only those that occurred more than 40 times per trillion). These frequencies are a rich source of information about the word grouping patterns of English and give us an unprecedented ability to estimate the probability of word co-occurrence in language. There are, undoubtedly, differences in the language experience of individuals that are not captured by the broad coverage of the web corpus, but the immense number of *n*-grams included in the Web1T corpus make it invaluable to those seeking to understand the influence of *n*-gram probability on lexical processing. By using the information in this large corpus we hope to better understand performance on cloze tasks.

A Cloze Experiment using *n*-grams

The first reported cloze task was created by taking a passage of 200 to 300 words and replacing a certain percentage of the words with blanks (Taylor, 1953). In later psycholinguistic studies the cloze task was often shorter, often single sentences, optionally with the blank locked in the same position (sentence-final for example, as in Schwanenflugel and LaCount, 1988). The amount of context was recognized to be critical to the choices participants made, and it follows that as the amount of context shrinks, the processing required changes. One question we will address in this study is: What happens when the context is reduced almost to the minimum, to three words? The poverty of context should reduce the constraints on the number of meaningful ways to complete the fragment and allow a greater variety of completions than for longer sentences or passages. Another difference between 3-gram contexts and sentence contexts is the level of meaningfulness: most of the 3-grams we will use in our studies are not syntactic constituents and some will remain fragmentary even after filling in the blank (e.g.*nothing to do with the*).

Most cloze experiments have only permitted participants to submit one response per

context, forcing the participant to choose the best response or sometimes the first response they can think of. To address the limitations of a standard, single response cloze task, we allowed each participant to list multiple cloze completions. The participants were given the opportunity to generate as many completions as they could (up to 20) with no time limit to allow us to probe differences in the productivity of *n*-grams. This will enable us to analyze the order in which the completions were generated, giving us a window into the sources of relative accessibility of the responses.

The position of the blank in the stimuli was limited to two locations: to the beginning of the stimulus (prepended responses, or **PRs**) or the end of the stimulus (appended responses or **ARs**). By limiting the location of the responses to either the beginning or the end of a trigram context we are able to calculate the conditional probabilities directly from the *n*-gram probability data in the Web1T dataset, allowing us to investigate how the predictability of a word in context can influence behavior in a cloze task.

Is there prior work that can inform our thinking about this task? A relevant body of work exists in studies of verbal fluency. In these studies the participant is presented with a categorical cue and asked to generate as many members of that category as possible in a fixed amount of time, usually one minute (Ruff, Light, Parker, & Levin, 1997). The categories are often letter categories ("Words that begin with the letter F.") or semantic categories ("Animals"). The letter fluency task is in some ways similar to our experiment because we ask our participants to generate up to twenty members of a category without time limits. Our categories, though, are slightly different ("Words that can co-occur with the 3-gram *chocolate chip cookie*") but the response space is essentially the same as that in the letter cloze task: a subset of all words in the language. Owing to these similarities, we will also attempt to understand our results in the context of research done on verbal fluency. We also analyzed the total number of responses produced, which is one of the commonly used dependent measures in verbal fluency studies.

One of the few studies that used a task similar to ours was conducted by Owens, O'Boyle, McMahon, Ming, and Smith (1997). They were interested in speech recognition systems and had built a weighted average *n*-gram language model that predicted words based

on their probability given the previous context. They used the one million word Brown corpus (Kučera & Francis, 1967) to train their model, then used the model to calculate the top 30 completions for 768 fragments. The same fragments were given to eight human participants asking them to provide a ranked list of completions for fragments of text such as *the republicans must hold a _____under the county*. The results of the statistical model and the humans were compared with the intention of measuring the quality of the model. They found that their model was almost as good as the humans in picking the word that had been deleted from the passage (21% versus 26% correct) and was almost as good for getting the correct answer within the first three tries (72% versus 80% correct). From a psychological point of view it is fascinating that humans performing this task produce such similar responses to a statistical model of memory built on *n*-grams. The corpus used in our study is larger than the one used by Owens et al. by a factor of 10^6 and our methodology is different but our hypothesis is that *n*-gram probabilities will have a similar predictive power for the participants' performance in a multiple-response cloze task.

Another study of interest by Crowe (1998) looked at the change in the responses to a verbal fluency task over time. Subjects were given one minute to complete letter and semantic fluency tasks, and the number of responses was counted over each 15 second interval. Crowe noted that the largest number were produced in the first 15 seconds and progressively smaller numbers of responses were produced for each of the following three 15-second periods. Also, the orthographic frequency of the words was higher for the first words produced, and lower for the later words. This task is similar to our task, and we will be able to analyze the order of production to see if the same pattern appears in our results.

If the memory systems of the participants are capturing the statistical properties of the English language, in particular the probability of a word occurring in a specific micro-context, then the *n*-gram statistics found in the Web1T corpus should help predict the likelihood that the participants will choose a certain completion in this cloze task.

Participants

Using a custom web experiment management system, we recruited 864 undergraduate students from the University of Alberta. All were self-described native speakers of English. All received course credit for their participation.

Materials

We randomly sampled 240 trigrams from the Web1T corpus to cover a broad range of *n*-gram frequencies. We sampled these trigrams without regard for their status as constituents. Most of the trigrams in the corpus are very low frequency, and so some of the stimuli we sampled could appear to be malformed (e.g. *in the to*). Rather than try to filter out these items using arbitrary criteria we left them in the stimulus set and gave them no special treatment. As we will see, the inclusion of these items likely contributed to the production of a large number of idiosyncratic responses.

Procedure

The survey was administered using custom web-based software. All participants completed the web survey at a time and location of their choosing. In the pre-survey instructions they were requested to find a quiet location where they would not be disturbed before starting to complete the survey.

The participants were randomly assigned into one of 30 groups and each group was asked to complete a different survey. Each survey consisted of a set of 8 trigrams for a total of 240 unique trigram contexts. Participants were asked to type in a word either before or after the context, and were given twenty fields to use for each context. If all 864 participants had provided 20 completions for each of the 8 contexts they saw, the maximum total number of observations would have come to 138,240. We did not receive this many responses, perhaps due to our decision not to set any restrictions on the minimum number of responses in the experiment. Participants were allowed to submit surveys with between 0 and 20 responses per item and still receive credit for their participation. We obtained 77,621 data points, 56% of the total possible.

This survey was part of a larger package of surveys that took approximately 50 minutes to complete in total. The other surveys in the package did not contain any tasks that were similar to this task. The *n*-gram completion task was always the last of the web surveys to be administered in the package.

The instructions at the beginning of the survey asked participants to fill in a blank with the first word that came to mind, and to avoid changing the order of their responses once they had typed in a word. They were instructed to only type in one response per line, either before or after the *n*-gram. It was explicitly noted that all completed phrases should make sense. They were also asked not to consult books, web pages or other resources when thinking about how to complete these *n*-grams. The text input field did not allow more than 12 characters to be entered forcing the maximum word length for all responses to be 12 letters.

There were some entries that were excluded because participants typed a word both before and after the *n*-gram. After eliminating 7,925 (10%) of the responses because of this type of double entry error, we were left with 69,696 observations.

Statistical Considerations

Our analyses will focus on cross-subject patterns, and so the pattern of individual, random subject differences in these cloze tasks and verbal fluency tasks needs to be accounted for. We handle random participant and item variation in our data using mixed effects models. In particular we will follow the recommendations of Baayen (2008) to use mixed effects models with crossed random subject and item effects. We will use mixed models whenever there is a possibility of random within-subject or within-item variability. During model selection we tested models with random slopes for our predictors of interest and only retained these random slopes when there was an improvement found.

All analyses were complete with R version 3.0.0 (R Development Core Team, 2009), **lme4** version 1.0-6 for linear mixed models (Bates, Mächler, & Bolker, 2011) and **mgcv** version 1-7.28 for generalized additive mixed models, or GAMMs (Wood, 2006). Whenever we analyze a dependent measure that is a count, such as a response count, we assume that it is Poisson distributed, and we use a logarithmic link function.

In all of our analyses we perform statistical inference using model selection. We used a forward stepwise model building process, building increasingly more complex nested models, and stopping when we found models that had the best balance between fit and complexity. All models were compared using a likelihood ratio test and testing χ^2. The change in the Akaike Information Criterion (AIC, Akaike, 1974) was also calculated as a secondary measure of model quality. We did not report the initial comparison of the null model without random intercepts and slopes to the model that included them because the models with random intercepts and slopes were always superior when tested with the likelihood ratio test.

One issue with these co-predictors is that they are all highly correlated. In an analysis of the colinearity of these predictors we found that the level of multi-colinearity as defined by the condition number[2], κ, was unacceptably high ($\kappa > 50$), likely rendering any regression models unstable. The only method available to us to retain these covariates was to create orthogonal principal components through principle component analysis (PCA). In all of the following analyses, a PCA was carried out and a statistical model that included the most explanatory principal components (henceforth PCs) as predictors was fitted to the data. In all cases the size and direction of the effects that we found in models that did not contain the PCs was the same as when we included them. For this reason, and because effects of PCs are difficult to interpret, the analyses that included the PCs will not be reported in this paper, but will be made available with the raw data as supplementary material. The multi-collinearity of our predictors, κ, was below 30 in all of our final models as suggested by Belsley, Kuh, and Welsch (2004).

Finally, we performed model criticism on all of the models presented in the paper: we located those observations that led to residuals greater than 2.5 standard deviations away from the mean of the residuals and temporarily dropped them from the data. We then reanalyzed the data using this subset and checked to see if there were any large fluctuations in the size or direction of the effects. All of our models did not differ in their interpretation before and after model criticism.

We will analyze the data in the following ways. First, we will consider the popularity of

[2]For an explanation of how the condition number of a matrix of predictors is calculated see Chambers (1992).

each response, as in the standard cloze analysis. Next we will analyze the tendency for participants to prefer to generate PRs or ARs for each stimulus. We will then analyze the size of the response set for each stimulus and the amount of entropy in each response set. Finally we will analyze the order in which the responses were generated by our participants.

Results: Response Frequency

In the first analysis we tried to predict the frequency of responding, or cloze probability, of each produced word in each context. To filter out nonsense responses we removed 4-grams (response plus stimulus) that did not have a corresponding entry in the Web1T corpus. As has been noted by Hahn and Sivley (2011), there is an issue with the Google Web1T data from Brants and Franz (2006): due to technological constraints, it is very computationally expensive to collect frequencies for very rare *n*-grams. For this reason the corpus only contains data for *n*-grams that occurred 40 occurrences per trillion or greater. Since our trigram stimuli were drawn at random from the Web1T corpus, and since most trigrams are rare, many low frequency trigrams were drawn (the full list of stimuli are given in Appendix A). Finding a sensible completion to a very low frequency proved to be extremely difficult for the participants. This may explain why most of the participants' 4-gram responses were in the range of 0 to 39 occurrences per trillion, leaving them out of the Web1T 4-gram data. Without a 4-gram frequency measure there is no point including these responses in a statistical model. A casual inspection of these removed responses showed them to be almost all nonsensical[3]. Following Nelson, McEvoy, and Schreiber (1998), who found that it was necessary to remove rare responses to do any meaningful analysis of this type of free-response data, all idiosyncratic responses were dropped from the dataset. After removing all responses without Web1T frequency data, the number of *n*-gram types dropped from 47,438 types to 8,066 types, translating to a drop from 69,696 responses to 18,357 responses. Of the 51,340 observations that were dropped, 33,114 were singletons. The mean number of observations per type was still quite small at 2.2 ($\overline{\sigma} = 2.6$, range = 2 to 29).

During this process of removing idiosyncratic responses the number of contexts with

[3] A typical nonsense response was *treasure trove of dentists* given the context was *treasure trove of* _____.

data dropped from 240 to 201, meaning that 39 contexts produced 4-gram responses that were *all* absent from the Web1T corpus (a total of 937 observations). A statistical summary of the 39 items that were dropped at this point is provided in Appendix B.

The final data set was split into two sub-groups: PRs (7,461 observations of 3,248 types) or ARs (10,895 observations of 4,818 types). These are the data sets that we will analyze.

At this point we also calculated pointwise mutual information values (PMI, Fano and Hawkins, 1961). PMI measures the degree to which words in an *n*-gram occur together more frequently than would be expected by chance. PMI was calculated using the following formula:

$$PMI_{\text{context}} = log\left(\frac{P_{\text{context}}}{P_{w_1} \times P_{w_2} \times P_{w_3}}\right) \quad (1)$$

where P_{w_1} is the probability of the first word occurring alone and P_{context} is the probability of the three words occurring together[4]. We also calculated the PMI of the PRs and ARs by taking the 4-gram's probability as the numerator and expanding the denominator with the fourth word's probability. This covariate was included during model selection.

We hypothesized that if many participants produced the same response for an item, that response had been seen in that context by that participant at some time in the past, and that this contextual cue aided in retrieving a memory involving that context. The *n*-gram probabilities in the Web1T corpus are a rough approximation of the probability of experiencing an *n*-gram in each participant's personal experience. Consequently we expected that the conditional probability of that response given each context in the Web1T corpus should be predictive of the response. The conditional probabilities for the responses were calculated in the following manner:

$$P(\text{Response Word}|\text{Context}) = \frac{P(\text{Response Word} \bigcap \text{Context})}{P(\text{Context})} = \frac{P(\text{4-gram})}{P(\text{Context})} \quad (2)$$

We entered this variable along with our other predictors and began fitting models to the data.

Each response might be influenced by various Web1T probabilities (word, bigram and trigram) and since these probabilities can be highly inter-correlated, we tested for excessive

[4]The PMI calculation assumes that the unigram probabilities P_{w_1}, P_{w_2}, and P_{w_3} are independent. In the case of word probabilities in a corpus, this assumption is very likely incorrect, but for our purposes it should allow us to estimate how unlikely it is to find a combination of certain words in an *n*-gram.

multi-collinearity. We found that all of our predictors of interest had an acceptable level of collinearity ($\kappa < 30$). The other frequency predictors were entered into a PCA, and the five most predictive principle components were also entered into our models, after transforming them using a Johnson transformation (Chou, Polansky, & Mason, 1998) so that they would be more symmetrically distributed. In all analyses these PCs were not found to be predictive and were left out of all models.

We detected non-linear relationships in our models, and so we chose to use GAMMs to better understand the data. We included random intercepts for each context (as a penalized smooth) in all of our analyses to take into account the shared variation due to each context. We did not include random intercepts for each participant due to the nature of our experimental design, where each subject only responded to 8 items out of the full set of 240 items. There was a striking similarity between the variables retained in our models for the PR and AR datasets: we found reliable interactions between the effect of the conditional probability of each response word given the context and the probability of the word that the participant supplied as well as the probability of the *n*-grams created by affixing the word to the context.

The other predictors that were temporarily entered into models but did not improve the models were: the PC's for the remaining frequency variables, the context PMI and the response 4-gram PMI. The whole *n*-gram frequency was not included in any of the models because it was already included in the calculation of the conditional probability.

We built models that included all the variables of interest and only retained models that were both less complex and better fitting. For the PRs, the best model contained interactions between the conditional probability of the word given the context interacting with two probability measures: the word probability and the probability of the bigram that was formed by combining the supplied word with the first word of the context. We compared two GAMMs, one with linear interactions of these variables, and another with the non-linear tensor smooths. The χ^2 test showed that the more complex model was a better model : ($\chi^2(3) = 34.6$, $p = 6.5 \times 10^{-15}$). The model comparison is shown in Table 1. The only linear term in the model was word length. Word length had a negative relationship with response counts: shorter words were easier to generate than long words. To understand smooth

components and the non-linear interactions we plotted countour plots of the tensor smooths in Figure 1. Contour plots contain contour lines that define a path of constant altitude. Lines that bunched close together imply a steep gradient, whereas lines that farther apart show a shallower gradient. In these plots, the contour lines represent a certain number of responses produced. In the left-hand plot, we see that for higher probability words, the gradient of the conditional probability effect is steeper than for that of the lower probability words (seen in the curvature of the countour lines). For the bigram probability (in the right-hand plot), the peak production zone was in the upper right corned of the plot (highly probably bigram and highly probable word in context). For the low probability bigrams, the conditional probability of the word in the context did not modulate the number of responses.

We applied the same analysis to the ARs. During model selection, a third measure entered the best model: The probability of the final 3-gram in the completed 4-gram. This probability also interacted with the conditional probability measure, and it so it was included in the model. Once again we compared the model with linear interactions of these three variables with a model containing three tensor smooths. The χ^2 test showed that the more complex model was a better model : ($\chi^2(3) = 61.5$, $p = 1.7 \times 10^{-26}$). The effect of word length was in the same direction as for the PRs, with longer words being produced less often.

The smooth effects are reported in Table 2 and visualized in Figure 2. This first plot shows a facilitation in responding medium probability words: high and low probability words had less responses across the range of conditional probabilities than medium probability words. For the final trigram probabilities, the pattern was slightly different. For this non-linear interaction, the peak production was for high probability *n*-grams combined with high conditional probability contexts. For the trigrams, an example would be the high probability trigram *on behalf of* created by added *of* to the end of *and on behalf*.

Discussion: Response Frequency

Our models took into account the length of the words produced by our subjects, and we found that subjects preferred shorter words. More importantly, we found reliable interactions between the component *n*-gram probabilities and the conditional probability of the word given

	Model 1	Model 2	Model 3	Model 4
Linear Terms, $\hat{\beta}$ ($SE_{\hat{\beta}}$)				
Intercept	$0.70\ (0.04)^{***}$	$1.96\ (0.12)^{***}$	$2.83\ (0.18)^{***}$	$0.60\ (0.04)^{***}$
Response Length (Letters)	$-0.32\ (0.02)^{***}$	$-0.24\ (0.02)^{***}$	$-0.24\ (0.02)^{***}$	$-0.23\ (0.02)^{***}$
$\log P$(First word|Context)		$0.26\ (0.02)^{***}$	$0.34\ (0.03)^{***}$	
$\log P$(First word)		$0.06\ (0.01)^{***}$	$0.02\ (0.02)$	
$\log P$(First bigram)		$0.01\ (0.00)^{***}$	$0.01\ (0.00)^{***}$	
$\log P$(First word|Context) $\times$ $\log P$(First Word)			$0.09\ (0.01)^{***}$	
$\log P$(First word|Context) $\times$ $\log P$(First Bigram)			$0.01\ (0.00)^{***}$	
Smooth Terms, *edf* (*ref edf*)				
Smooth for per-context random intercepts	$171.13\ (219.00)^{***}$	$166.80\ (219.00)^{***}$	$171.17\ (219.00)^{***}$	$171.00\ (219.00)^{***}$
$\log P$(First word|Context) $\otimes$ $\log P$(First Word)				$6.83\ (7.43)^{***}$
$\log P$(First word|Context) $\otimes$ $\log P$(First Bigram)				$2.66\ (2.88)^{***}$
AIC	12559.83	11942.19	11882.28	11828.05
Log Likelihood	-6106.78	-5799.30	-5762.97	-5731.53

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1

Model coefficient estimates in the GAMM for response count for PRs. Standard errors are shown in parentheses next to the estimate for each linear coefficient. For smooths, the effective degrees of freedom are shown, with the maximum degrees of freedom possible shown in parentheses. The symbol $\times$ denotes a linear interaction while the symbol $\otimes$ denotes a non-linear interaction modeled with a tensor product smooth.

the context, as measured in the Web1T corpus. Despite the differences between these two experiments, the same forces were at play. This evidence supports the hypothesis that implicit contextual memory drives word choice.

Results: Response Position

Our participants always had a choice to make in this experiment: to place their responses before or after the stimulus *n*-gram. Did any of our probabilistic measures influence their choice to produce a PR or an AR? To measure this we counted the total number of responses for each stimulus in each position and then determined if the item had a majority of PRs or ARs. As an illustration of three contexts with majority PRs were *gallons per day*, *in the moment*, and *on this page*, while three with majority ARs were *to that contained* and

	Model 1	Model 2	Model 3	Model 4
Linear Terms, $\hat{\beta}$ ($SE_{\hat{\beta}}$)				
Intercept	$1.82\ (0.04)^{***}$	$2.25\ (0.11)^{***}$	$2.81\ (0.20)^{***}$	$0.70\ (0.03)^{***}$
Response Length (Letters)	$-0.07\ (0.01)^{***}$	$-0.13\ (0.01)^{***}$	$-0.13\ (0.01)^{***}$	$-0.13\ (0.01)^{***}$
$\log P$(Last Word|Context)	$0.19\ (0.01)^{***}$	$0.16\ (0.01)^{***}$	$0.26\ (0.03)^{***}$	
$\log P$(Last Word)		$-0.07\ (0.01)^{***}$	$-0.11\ (0.01)^{***}$	
$\log P$(Last trigram)		$0.06\ (0.01)^{***}$	$0.12\ (0.01)^{***}$	
$\log P$(Last Word|Context) $\times$ $\log P$(Last Word)			$-0.01\ (0.00)^{***}$	
$\log P$(Last word|Context) $\times$ $\log P$(Last Trigram)			$0.01\ (0.00)^{***}$	
Smooth Terms, *edf* (*ref edf*)				
Smooth for per-context random intercepts	$160.99\ (213.00)^{***}$	$155.56\ (213.00)^{***}$	$155.33\ (213.00)^{***}$	$155.35\ (213.00)^{***}$
$\log P$(Last Word|Context) $\otimes$ $\log P$(Last Word)				$5.91\ (6.49)^{***}$
$\log P$(Last Word|Context) $\otimes$ $\log P$(Last Trigram)				$5.33\ (5.72)^{***}$
AIC	17743.03	17641.86	17619.99	17498.50
Log Likelihood	-8707.52	-8660.37	-8647.67	-8580.66

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2

Model coefficients for GAMM predicting the response frequencies of ARs. Standard errors are shown in parentheses next to the estimate for each linear coefficient. For smooths, the effective degrees of freedom are shown, with the maximum degrees of freedom possible shown in parentheses. The symbol $\times$ denotes a linear interaction while the symbol $\otimes$ denotes a non-linear interaction modeled with a tensor product smooth.

which may be and *with obtaining the*.

We fit a generalized linear model (GLM) predicting whether each context with have majority PRs or not (the binomial outcome). The best model for retained two of our predictors, the unconditional probability of the final word in the stimulus and the PMI of the context trigram. In Table 3, we see the following: contexts with higher probability final words were more likely to have a majority of **ARs** and contexts with higher PMI were more likely to have a majority of **PRs**.

Discussion: Response Position

To understand these effects, we looked into the stimulus properties. The highest frequency words were often closed class words, implying that this effect may be connected to

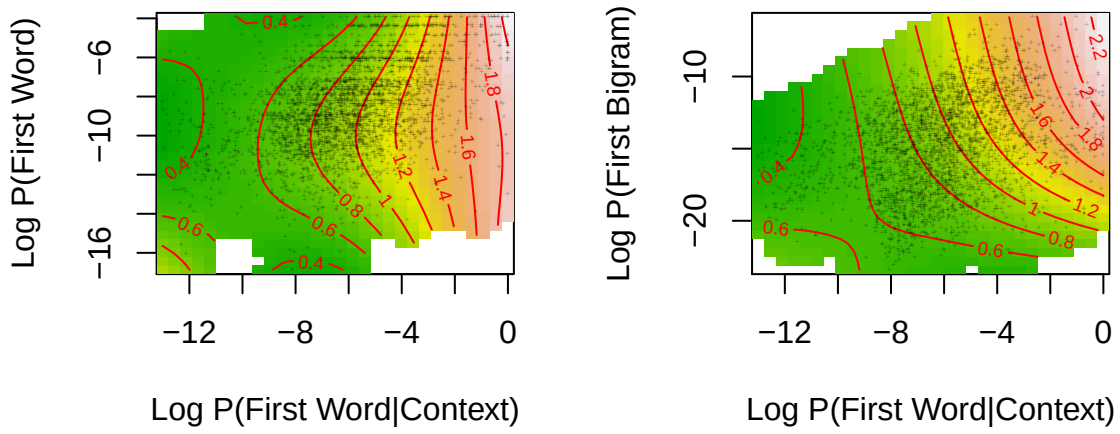

Figure 1. Contour plots of the tensor smooths from the GAMM models for PR frequency. Red contour lines represent the number of responses produced (log-transformed). All data points are plotted with a semi-transparent plus symbol. In this data, the first word the 4-gram is the response word.

	Model 1	Model 2	Model 3
Intercept	$0.32\ (0.03)^{***}$	$-0.03\ (0.06)$	$0.05\ (0.06)$
log probability final word		$-0.06\ (0.01)^{***}$	$-0.05\ (0.01)^{***}$
Context PMI			$0.10\ (0.03)^{***}$
AIC	319.40	283.75	274.52
Log Likelihood	−157.70	−138.87	−133.26

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3

Coefficient estimates for the GLM predicting majority PRs. Standard errors are shown in parentheses next to the estimate for each linear coefficient.

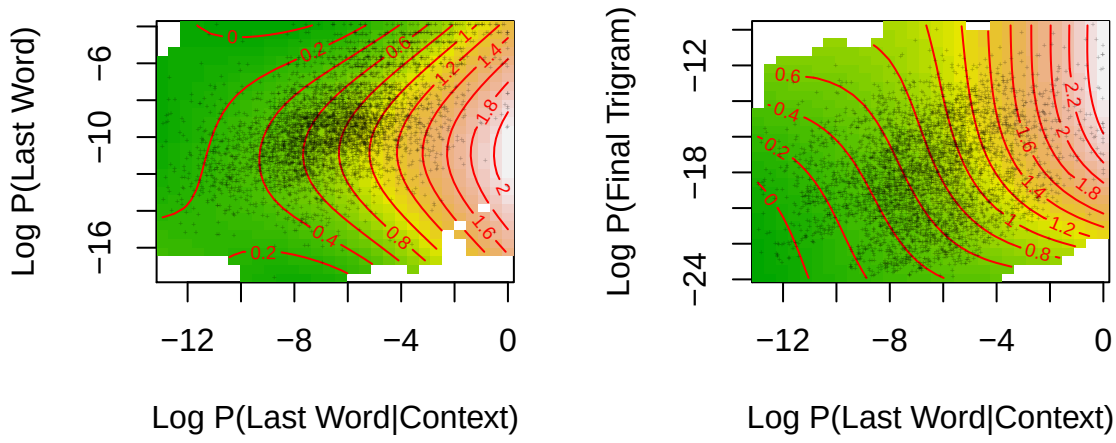

Figure 2. Contour plots of the tensor smooths from the GAMM models for AR frequency. Red contour lines represent the number of responses produced (log-transformed). All data points are plotted with a semi-transparent plus symbol. In this data, the last word the 4-gram is the response word.

the class of the final word. If a closed class, or even high-probability open class word, appeared at the end of one of our contexts it biased our participants to add responses to the end of the *n*-gram. In the same stroke, low-probability open class words at the end of a stimulus inspired fewer ARs and more PRs. As for the PMI effect, looking at high PMI contexts, such as *obstructive pulmonary disease*, it becomes clear that these *n*-grams are more lexicalized and are easier to modify with a PR than an AR.

Results: Response Order

In this final analysis we looked at the order in which the responses were generated by the subjects. Each response has a position in the response list from 1st to 20th, which correspond to the order in which the participant generated the response. What models can predict the position for each response? We used the same set of 18,357 responses that we used in the previous analysis (leaving out the same 39 stimuli listed in Appendix B).

Due to the presence of non-linear effects, we analyzed the data using GAMMs. To

better understand the contrast PRs and ARs, we split our dataset in two: 7,464 PRs and 10,893 ARs. We will report the results for these subset separately but the process that we used to build our models was identical.

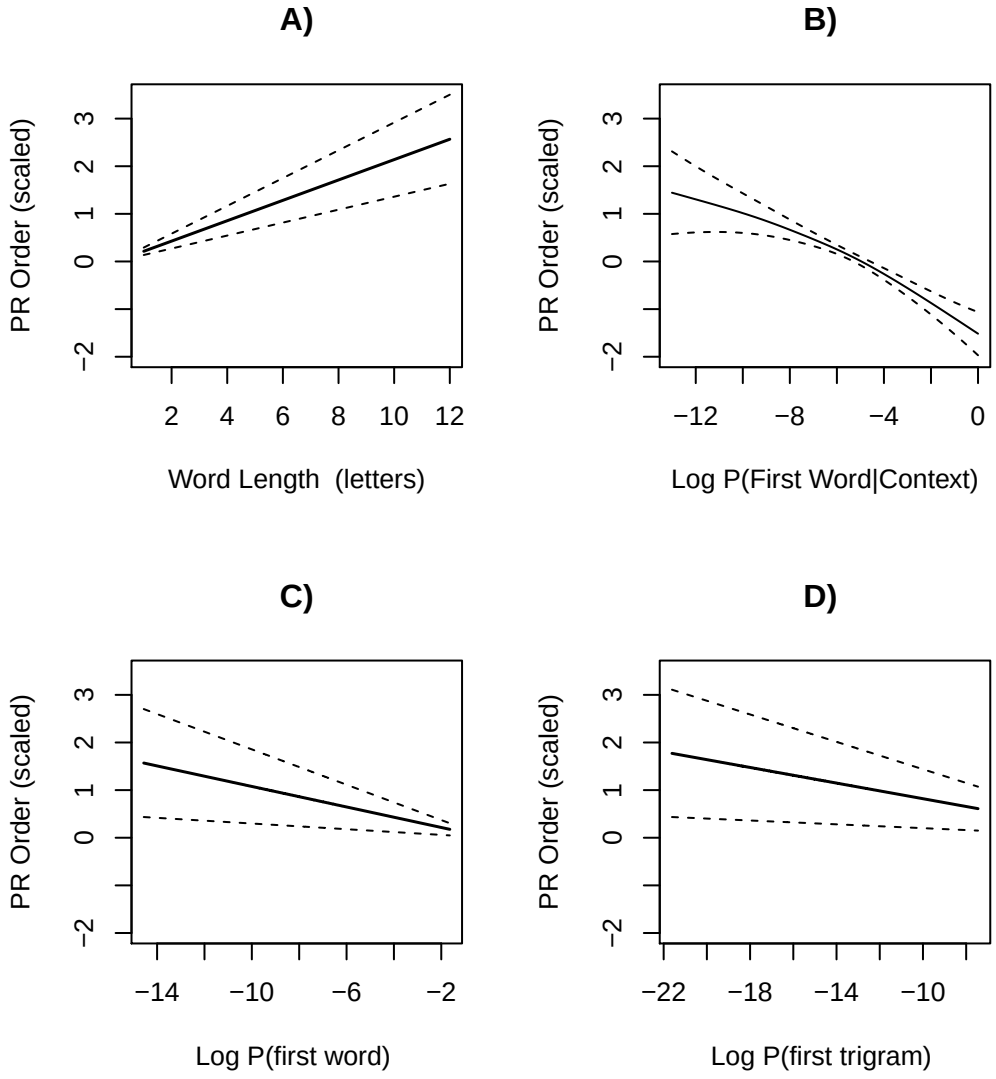

Figure 3. Partial effects of four covariates in the GAMM model predicting response order for PRs. Dotted lines represent 95% confidence intervals. A) Response length, B) log conditional probability of the first word given the subsequent context, C) log first word probability (the responded word) , and D) Log first trigram probability (the response word and the two following words).

First we will present our results for PRs and then for ARs.

	Model 1	Model 2	Model 3	Model 4	Model 5
Linear Terms, $\hat{\beta}$ ($SE_{\hat{\beta}}$)					
Intercept	$5.41\ (0.13)^{***}$	$3.88\ (0.18)^{***}$	$0.20\ (0.42)$	$0.38\ (0.41)$	$2.83\ (0.52)^{***}$
Response Length (Letters)		$0.39\ (0.03)^{***}$	$0.31\ (0.03)^{***}$	$0.21\ (0.04)^{***}$	$0.21\ (0.04)^{***}$
log P(First Trigram)			$-0.25\ (0.03)^{***}$	$-0.21\ (0.03)^{***}$	$-0.08\ (0.03)^{**}$
log P(First Word)				$-0.17\ (0.04)^{***}$	$-0.11\ (0.04)^{**}$
Smooth Terms, *edf* (*ref edf*)					
Smooth for per-participant random intercepts	$460.25\ (821.00)^{***}$	$460.57\ (821.00)^{***}$	$454.72\ (821.00)^{***}$	$455.62\ (821.00)^{***}$	$452.99\ (821.00)^{***}$
Smooth for per-context random intercepts	$116.80\ (219.00)^{***}$	$109.94\ (219.00)^{***}$	$118.59\ (219.00)^{***}$	$110.99\ (219.00)^{***}$	$99.65\ (219.00)^{***}$
Smooth for log P(First Word|Context)					$1.90\ (2.37)^{***}$
AIC	44230.89	44095.10	43998.18	43993.60	43967.59
Log Likelihood	-21536.40	-21474.04	-21421.78	-21425.19	-21424.26

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4

Coefficient estimates for the GAMM for response order of PRs. Standard errors are shown in parentheses next to the estimate for each linear coefficient. For smooths, the effective degrees of freedom are shown, with the maximum degrees of freedom possible shown in parentheses.

	Model 1	Model 2	Model 3	Model 4
Linear Terms, $\hat{\beta}$ $(SE_{\hat{\beta}})$				
Intercept	$5.50\ (0.13)^{***}$	$4.80\ (0.17)^{***}$	$0.65\ (0.39)$	$3.87\ (0.48)^{***}$
Response Length (Letters)		$0.16\ (0.03)^{***}$	$0.10\ (0.03)^{***}$	$0.07\ (0.03)^{**}$
$\log P$(Last trigram)			$-0.31\ (0.03)^{***}$	$-0.13\ (0.03)^{***}$
Smooth Terms, *edf* (*ref edf*)				
Smooth for per-participant random intercepts	$128.29\ (837.00)^{***}$	$148.96\ (837.00)^{***}$	$221.56\ (837.00)^{***}$	$304.06\ (837.00)^{***}$
Smooth for per-participant random slopes for $\log P$(Last word|Context)	$492.18\ (838.00)^{***}$	$474.09\ (838.00)^{***}$	$393.60\ (838.00)^{***}$	$308.14\ (837.00)^{***}$
Smooth for per-context random intercepts	$123.91\ (213.00)^{***}$	$121.70\ (213.00)^{***}$	$125.23\ (213.00)^{***}$	$106.10\ (213.00)^{***}$
Smooth for $\log P$(Last Word|Context)				$4.53\ (5.54)^{***}$
AIC	64944.40	64918.62	64833.98	64808.77
Log Likelihood	-31725.82	-31711.57	-31672.60	-31677.56

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5

Coefficients in the GAMM for response order for ARs. Standard errors are shown in parentheses under each linear coefficient (for smooths, the effective degrees of freedom are shown).

PRs. The estimated regression coefficients for both parametric and smooth terms are shown in Table 4. The first two predictors did not have any detectable non-linearity when fitting with the GAM. The unconditional probability of the chosen word had a linear, negative effect on the order it was produced, indicating that responses with a high probability word in the first position were produced before responses with a low probability word. There was a similar relationship between the probability of the first trigram (the word produced plus the two following words) and the response order: if the trigram created was higher in probability, it was produced earlier than if it was lower in probability. The relationships between predictors and response order are plotted in Figure 3. The number of letters in the word that the participant typed in was predictive: shorter words were produced earlier. The final predictor was the conditional probability of the response word given the context. This non-linear relationship was also in the negative direction. The greater the conditional probability, the earlier the response of was generated. Neither bigram, trigram, quadragram frequencies nor PMI were retained during this model selection process as they did not

contribute any explanatory power to the model.

ARs. For the ARs, three predictors were retained during model selection, listed in Table 5. In the best model we found a strong by-subject effect of the conditional probability, and so this random slope smooth was included in the final model. These non-linear relationships are depicted in Figure 4. The first predictor was length of the typed responses, and in line with the PRs, the longer responses were produced later that the shorter responses. The next predictor was the probability of the final trigram, the last two words of the context plus the responded word. It had a negative relationship with the order in which the responses were produced. Responses with a higher probability final trigram (*bottom* ***part of the***, where the participant responded with *the*) were produced earlier than responses with a lower probability final trigram (*chocolate* ***chip cookie lover***, where the participant responded with *lover*). The final predictor in this model was the conditional probability of the final, responded word given the preceding trigram. This effect was strong and negative, showing that words that were more likely to occur after the trigram in the Web1T corpus given the trigram context were generated earlier. As with the PRs, the other *n*-gram frequencies and PMI did not enter the model. It is particularly interesting to compare these results with those from the PRs, where the word's unconditional probability was a predictor of order of production, which was not the case for the ARs.

Discussion: Response Order

The results from the analysis of response order in the PRs and ARs provide evidence that the search process the participants used to generate responses in this cloze task was sensitive to the conditional probabilities of the *n*-grams that were created. The predictors that remained in the models can help us understand this search process. For the PRs, high probability words were the first to be generated, in particular high probability words that were part of a high-probability *n*-gram. For the ARs, words that had a high conditional probability given the preceding context were generated first, in particular those word that created a high probability trigram.

In this analysis, an *n*-gram probability influenced the order of responses produced, but

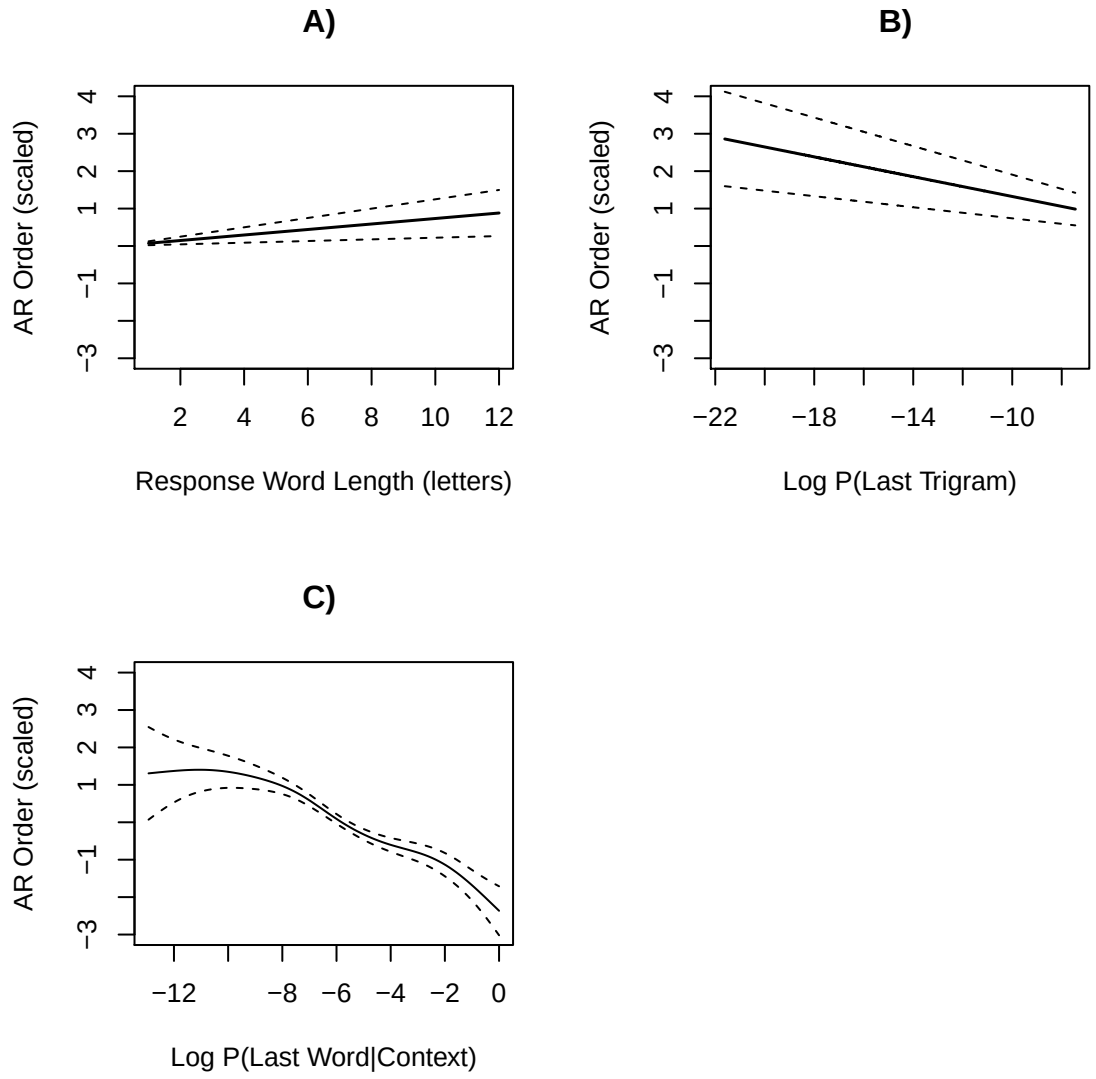

Figure 4. Partial effects from the GAMM predicting the order of generation of ARs. Dotted lines represent 95% confidence intervals. A) Response length in letters, B) corpus probability of the final trigram (two words from the context concatenated with the response word) and C) log conditional probability of the final word given the preceeding three word context.

did not interact with other measures (including unconditional word probability). Unlike the process of choosing a response from the set of possible words, the order in which that word was generated was purely an additive effect of probabilities. The probability of the trigram containing the response word influenced the choice process in both PRs and ARs: lower probability trigrams were generated later. Similarly, lower conditional probability responses were generated later.

The design of this experiment allowed us to calculate the conditional probability in the Web1T corpus of each response provided by the participants. We found support for the effects of conditional probability in the response frequencies and the in the order of generation of the responses.

Our results mirror those of Crowe (1998) for the PRs. As in his letter and semantic verbal fluency tasks, we found evidence that higher probability words were produced earlier. The same was not true for the ARs where the word probability of the responded word did enter into the model because it did not help explain the order data. The amount of constraint created by the letter categories and semantic categories in the verbal fluency tasks is much less than the amount of constraint in the *n*-gram cloze task and this may explain the difference between the ARs and PRs. Crowe (1998) proposed a mental store of high probability responses that eventually get depleted and force participants to use a different, slower search strategy to complete the task. Our data do not support this model since we found a continuous, linear, negative trend for the order of the subjects' responses, with no abrupt change in strategy evident.

In our results, we replicated several results reported by Smith (2011): the importance of the 3-word context, the tendency to produce shorter words over longer words and effect of unigram probability. Smith concluded that he had found evidence for a human analog of statistical smoothing, where the probability of a previously unseen occurrence is estimated from the known probability of similar events that have been seen. Our data do not contradict this theory, but they do not provide any more support than the data already presented by Smith. More research will be required before it becomes clear whether language users are doing anything akin to statistical smoothing when choosing words in a cloze task.

In the final section we discuss how the data our experiment fits into the larger landscape of psycholinguistic theory.

General Discussion

We performed two experiments in which we asked participants to complete *n*-grams that required either a letter or word to be added to them. The only constraints in this task were the other words in the *n*-gram, the linguistic micro-context. We uncovered evidence that probabilistic measures derived from large samples of language predict which words participants choose to complete these cloze *n*-grams and the order in which they generate them. For the first time we applied frequencies, conditional probabilities and PMIs extracted from a one trillion word corpus of English to this type of psycholinguistic task. The relationship between the corpus measures and the observed behavior imply that a similar kind of probabilistic information is available to the language system when choosing a completion in a cloze task. Since the conditional probability of a word in context can only be learned from linguistic experience, and since experiences are stored in memory, we submit that the process of completing the cloze task was a memory process. We will discuss the theoretical considerations from research on linguistic memory and cognition to better understand how this probabilistic information was used in the process of completing fragments.

As we mentioned in the introduction of this paper the cloze task could be thought of as a cued recall task without an explicit study list; participants had implicit learning due to their lifetime exposure to the stimuli. In this view, the context 3-gram is what the participant used when probing their memory. The required processing in a cued recall task and our cloze task appear, in this light, to be convergent. Comparing results, we see that in the cued recall task reported by Criss et al. (2010) high probability targets were recalled more frequently. The unconditional probability of the cue (the whole *n*-gram probability) was not a strong predictor of response order for the ARs, which is also congruent (in their studies, cue word probability did not affect target recall probability). The similarity of the pattern suggests that a similar process is taking place during cued recall and the cloze task.

In contrast we noticed fewer similarities between the *n*-gram cloze task and the lexical

free association task. Nelson, McEvoy, and Dennis (2000) proposed a theory of the free association process that models word choice as a system that samples a word from a distribution of candidate words when participants are given a certain free association cue. However, the size of the context given in our *n*-gram cloze task was greater than that of the single word cue in the free association task. With the increased constraint, it is difficult to see how the sampling process proposed by Nelson, McEvoy, and Dennis (2000) is directly relevant to the *n*-gram completion process. This may also explain the difficulty in aligning our results with the results from verbal fluency research.

How does our research fit in to other research on *n*-grams? One clear difference is that whole *n*-gram probability was not found to be predictive in our models of response frequencies and response order, whereas conditional probability was. These results appear to show a task-specific difference between the *n*-gram cloze task and other *n*-gram tasks. In whole *n*-gram reading the raw probability of *n*-grams indeed predicted changes in behavior (Arnon & Snider, 2010). In our data, the probability of the context (in this case a 4-gram) was used to calculate the cloze conditional probability, leading to a strong correlation between the conditional probability and the whole *n*-gram probability. Indeed, Shaoul, Westbury, and Baayen (2013) found weaker sensitivity to the probability of 4-grams and 5-grams in subjective frequency tasks, pointing to a limit on the size of *n*-gram that can influence processing.

Why did the conditional probability rather than the whole *n*-gram probability prove to be consistently crucial our models? We propose that the *n*-gram probability benefit that has been found in recent psycholinguistic experiments Arnon and Snider (2010), e.g. is not a consequence of the raw count of exposure. Rather, much like the effect of word probability, the effect of *n*-gram probability is epiphenomenal. Baayen (2010) argued persuasively that the information from the linguistic context is what dictates facilitation, not merely exposure. Our results provide further evidence that context, captured in measures of conditional probability, is key for processing *n*-grams.

This research is exploratory in nature and not a definitive adjudication. This is one of the first *n*-gram cloze task experiments reported. We have not as of yet fit any computational

psycholinguistic models to this data but we believe it would be beneficial to attempt to computationally simulate the data we have collected. To spur the development of computational models of word generation, all the raw data collected has been made available on the Mind Research Repository[5].

With minimal modification, some current computational models of multi-word reading, such as the NDR (Baayen, Milin, et al., 2011), the Bayesian Reader (Norris & Kinoshita, 2008), simple recurrent networks (Mirman, Graf Estes, & Magnuson, 2010) and neural networks (Dilkina, McClelland, & Plaut, 2010) are potentially capable of modeling this data. All of these computational models are trained on a corpus of text and build a large network of probabilistic relations between form and meaning. The NDR is particularly promising because it has already been applied to multi-word input (Baayen, Hendrix, & Ramscar, 2013). The NDR can generate predictions about upcoming words in a stream using the conditional probability of the sub-lexical information which is contained in the models association network. We hope to apply the NDR model to this data to see how well it can simulate the participants' cloze task performance, but to do this, the granularity of the model will need to be augmented. From sub-lexemes cues that discriminate lexemes, we will need to move a higher level, where context lexemes help discriminate lexemes. From there, perhaps to more abstract levels, where lexeme groups may eventually aid in the discrimination between events or experiences.

Computational simulations and models will undoubtedly reveal more about the workings of the cloze task. If these prediction-based computational models turn out to be good models of performance, it will validate the concept of the automatic forward-modeling system put forward by Pickering and Garrod (2007).

The existence of micro-context effects has implications for models of word selection in language production. Our results are only directly applicable to isolated short *n*-grams, but it is conceivable that during the production of longer utterances, and the writing of text, the probabilities from the micro-context of the previous few words has an effect on the next word produced. Once that word is chosen, the micro-context moves forward and begins to influence

[5] http://openscience.uni-leipzig.de/

the choice of the next word, and so on. The question that can now be addressed is: how do the micro-context (the local quadragram, for example) and the macro-context (the sentence or paragraph) interact when language is being produced?

Future research with cloze tasks will not only help us understand how we read groups of words and choose words during production, but also how we learn the meanings of *n*-grams. There is plentiful evidence that infants use the statistical patterns of language to learn how to speak and understand speech (Saffran, Aslin, & Newport, 1996). Infants have been shown to use statistical learning when learning both artificial and natural languages (Hay, Pelucchi, Estes, & Saffran, 2011) and recently implicit statistical learning has been seen in adults as well (Conway, Bauernschmidt, Huang, & Pisoni, 2010). The work of Ramscar and Gitcho (2007) on the changes in the architecture of response selection over the lifespan could be another fruitful area for future cloze task research. The *n*-gram cloze task can provide valuable data about how our word selection processes operate, and computational models can help us link language acquisition theories and language processing models.

There are powerful anticipatory processes at play in single word processing, *n*-gram processing, and sentence processing. The impact of context in the *n*-gram cloze experiments was pervasive and suggestive of a link between contextual memory and word selection. Without the computational infrastructure and extensive *n*-gram probability data that is now readily available we would not have been able to attempt to understand the processes underlying word choice in a cloze task. As the large data trend continues to progress other previously intractable problems in psycholinguistics may soon become tractable as well.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Arnon, I. & Cohen Priva, U. (2013). More than words: the effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*(3), 349–371.

Arnon, I. & Snider, N. (2010). More than words: frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H. (2010). Demythologizing the word frequency effect: a discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436–461.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: an explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, *56*(3), 329–347.

Baayen, R. H., Milin, P., Djurdjevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289.

Bates, D., Mächler, M., & Bolker, B. (2011). *lme4: linear mixed-effects models using S4 classes*. Retrieved from http://cran.r-project.org/web/packages/lme4/.

Battig, W. & Montague, W. (1969). Category norms of verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology*, (392).

Beattie, G. & Butterworth, B. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech*, *22*(3), 201.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression diagnostics: Identifying influential data and sources of collinearity*. Hoboken, NJ, USA: Wiley-Interscience.

Block, C. & Baldwin, C. (2010). Cloze probability and completion norms for 498 sentences: behavioral and neural validation using event-related potentials. *Behavior research methods*, *42*(3), 665–670.

Bloom, P. & Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory and Cognition*, *8*(6), 631–642.

Bormuth, J. (1966). Readability: a new approach. *Reading Research Quarterly*, 79–132.

Brants, T. & Franz, A. (2006). *Web 1T 5-gram version 1*. Philadelphia, PA USA: Linguistic Data Consortium.

Chambers, J. M. (1992). Linear models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical Models in S* (Chap. 4). NY, NY, USA: Wadsworth & Brooks.

Chou, Y. M., Polansky, A. M., & Mason, R. L. (1998). Transforming non-normal data to normality in statistical process control. *Journal of Quality Technology*, *30*(2), 133–141.

Conway, C. M., Bauernschmidt, A., Huang, S., & Pisoni, D. (2010). Implicit statistical learning in language processing: word predictability is the key. *Cognition*, *114*(3), 356–371.

Criss, A., Aue, W., & Smith, L. (2010). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language*.

Crowe, S. (1998). Decrease in performance on the verbal fluency test as a function of time: evaluation in a young healthy sample. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 391–401.

DeLong, K., Urbach, T., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2010). Are there mental lexicons? The role of semantics in lexical decision. *Brain Research*, *1365*, 66–81.

Ellis, W. (1999). *A source book of Gestalt psychology*. London, UK: Psychology Press.

Elman, J. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, *6:1*, 1–33.

Fano, R. M. & Hawkins, D. (1961). Transmission of information: A statistical theory of communications. *American Journal of Physics*, *29*, 793.

Fillenbaum, S., Jones, L., & Rapoport, A. (1963). The predictability of words and their grammatical classes as a function of rate of deletion from a speech transcript1. *Journal of Verbal Learning and Verbal Behavior*, *2*(2), 186–194.

Finn, P. (1977). Word frequency, information theory, and cloze performance: a transfer feature theory of processing in reading. *Reading Research Quarterly*, 508–537.

Francis, W. & Kucera, H. (1982). *Frequency analysis of English usage*. Boston, MA, USA: Houghton Mifflin Company.

Frank, S. L. & Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, *22*(6), 829–834.

Griffin, Z. & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, *38*(3), 313–338.

Hahn, L. W. & Sivley, R. M. (2011). Entropy, semantic relatedness and proximity. *Behavior Research Methods*.

Hay, J., Pelucchi, B., Estes, K., & Saffran, J. (2011). Linking sounds to meanings: infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106.

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass*, *2*(4), 647.

Kučera, H. & Francis, W. (1967). *Computational analysis of present-day American English*. Dartmouth, NH, USA: Dartmouth Publishing Group.

Kutas, M. & Hillyard, S. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

McEvoy, C. L., Nelson, D. L., & Komatsu, T. (1999). What is the connection between true and false memories? the differential roles of interitem associations in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1177.

McKenna, M. C. (1986). Cloze procedure as a memory-search process. *Journal of Educational Psychology*, *78*, 433–440.

Mirman, D., Graf Estes, K., & Magnuson, J. (2010). Computational modeling of statistical learning: Effects of transitional probability versus frequency and links to word learning. *Infancy*, *15*(5), 471–486.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28*(6), 887–899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* http://www.usf.edu/FreeAssociation/.

Nelson, D. L., McKinney, V., Gee, N., & Janczura, G. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, *105*(2), 299.

Norris, D. & Kinoshita, S. (2008). Perception as evidence accumulation and bayesian inference: insights from masked priming. *Journal of Experimental Psychology: General*, *137*(3), 434–455.

Owens, M., O'Boyle, P., McMahon, J., Ming, J., & Smith, F. (1997). A comparison of human and statistical language model performance using missing-word tests. *Language and Speech*, *40*(4), 377.

Pickering, M. & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

R Development Core Team. (2009). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramscar, M. & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Science*, *11*(7), 274–279.

Ruff, R., Light, R., Parker, S., & Levin, H. (1997). The psychological construct of word fluency. *Brain and Language*, *57*(3), 394–405.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Schwanenflugel, P. & LaCount, K. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 344.

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

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, *30*(1), 50–64.

Shaoul, C. & Westbury, C. F. (2011). Formulaic sequences: Do they exist and do they matter? *The Mental Lexicon*, *6*(1), 171–196.

Shaoul, C., Westbury, C. F., & Baayen, R. H. (2013). The subjective frequency of word n-grams. *Psihologija*, *46*(4), 497–537.

Smith, N. J. (2011). *Scaling up Psycholinguistics*. (Unpublished Doctoral Dissertation) Downloaded in December, 2013 from http://vorpus.org/. San Diego, CA, USA: University of California, San Diego.

Smith, N. J. & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the 33rd annual meeting of the cognitive science conference* (pp. 1637–1642).

Sprenger, S. & van Rijn, H. (2013). It's time to do the math: computation and retrieval in phrase production. *The Mental Lexicon*, *8*(1), 1–25.

Taylor, W. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433.

Tremblay, A. & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, *6*(2), 302–324.

Willems, R. & Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action: a review. *Brain and Language*, *101*(3), 278–289.

Wood, S. (2006). *Generalized additive models: An introduction with R*. New York, NY, USA: CRC Press.

Appendix A

Experimental Stimuli

Contexts	Contexts (cont)	Contexts (cont)
about what the	hang to dry	potentially toxic compounds
account at official	happily ever after	press release can
a couple weeks	have more to	preyed upon the
act on behalf	have not been	printable on your
affair with a	hereby certifies that	printable telephone numbers
aggregate amount payable	homeland security threats	quality of and
a minimum threshold	identify the potential	quality to your
a more moderate	if you are	quite a variety
an agent who	illustrative purposes only	rarely the case
and all my	image of an	reliable but not
and are of	implication or otherwise	reliable migration from
and in the	including air conditioning	relocating overseas canadians
and into the	including more than	repeats of the
and on behalf	increase in the	residential real estate
and the return	in into the	reverse chronological order
an income stream	instructor led training	revolves around the
an unbelievably low	internally powered by	scripts for the
are health care	in the advance	shall be certified
a report detailing	in the greatest	shines white light
are to facilitate	in the helping	shipping rates for
asbestos attorney lawyer	in the moment	small engine repair
as to claim	in the to	sooner rather than
as to state	into the website	source software community
at game show	in tracking the	speak not of
authorize appropriations for	is against a	specs of the
backed sequin belt	is approaching a	speeds of up
bacterial flagellum is	is designed in	strictly prohibited without
barriers to entry	is to elucidate	submitting to a
because it did	it just comes	subsequent to this
bird flu virus	its jurisdiction the	talk about what
bottom part of	java mortgage calculator	technical to make
brutally murdered his	judicial paperwork that	text have been
business and for	keen interest in	the entire class
butchering technique of	kiln lime production	the film have
but to a	liberates toxic gas	the on position
by telling the	locally advanced or	the on site
by the thread	mad doctoring skills	the response you
calories you burn	make it the	the same the
can be huge	member at the	the special meaning
can be left	molecular mass of	third most popular
can be required	more rather than	those with an
center offers a	more support than	thoughts with the
chocolate chip cookie	more will determine	to and establish
chronic pain condition	musical or comedy	to be impeached
click here if	my bloody valentine	to force the
combined shipping rates	not can not	to into the
comes back to	obligated to pay	to please contact
comparing store ratings	obstructive pulmonary disease	to rental cars
comply with a	of an underlined	to say goodbye
components is not	of cruises aboard	to that contained

compulsive behaviors which	of it comes	to the years
constitute endorsements of	of members for	to those offered
cooling plant setting	of the what	treasure trove of
credited alongside another	of things past	trusted source for
details of and	of ulcers caused	tucked away in
did not like	of your password	two consecutive years
died last week	on by his	under difficult circumstances
dietary supplements have	one able to	unequally yoked with
dietary supplements with	one iota of	urinary tract infection
distinction between public	one of this	used to love
ditch and bank	on store shelves	usher dashboard confessional
doting grandpa of	on this page	virtually all of
ears perked up	organizations all over	was above the
electric mixer until	our cover showed	way it was
electromagnetic waves of	our staff who	website shall be
explores essences of	outdoor activities such	weight loss vitamin
fits nicely into	outline of your	when he fought
flattens out the	particles in the	wherever they are
floodlit tennis court	payment details and	which equals the
for something new	pending renewal or	which may be
for the most	per day or	who carried out
friends and the	performance of three	whoever posted them
from the finest	perked me up	with obtaining the
front porch of	per year of	with too many
fullest extent of	physical high all	with you can
gallons per day	pill weight loss	workshops held in
genetically modified organisms	place to consider	worthy of more
glutton for punishment	pleads guilty to	years of credited
grain leather upper	polyester blend fabric	you realize that
hanging out in	polyphonic ring tones	you would on

Appendix B

Analysis of dropped experimental items

The full list of items that were eliminated from the dataset is provided in Table B1. Descriptive statistics for these two groups are given in Table B2. The effect sizes reported are all greater than 0.8, showing there are clear differences between the two groups of words. The dropped stimuli were on average lower in probability and higher in PMI than the retained stimuli.

Items Dropped from Exp.2	
account at official	explores essences of
asbestos attorney lawyer	to those offered
backed sequin belt	at game show
cooling plant setting	implication or otherwise
ditch and bank	relocating overseas canadians
homeland security threats	unequally yoked with
java mortgage calculator	constitute endorsements of
judicial paperwork that	of an underlined
kiln lime production	of the what
reliable migration from	our cover showed
technical to make	hang to dry
usher dashboard confessional	internally powered by
in the to	shines white light
printable telephone numbers	as to claim
liberates toxic gas	to rental cars
including air conditioning	the film have
mad doctoring skills	not can not
whoever posted them	to into the
the on site	compulsive behaviors which
of cruises aboard	

Table B1

Context Stimuli dropped due to lack of quadragram probability data in the Web1T corpus.

Statistic	Mean for Dropped Stimuli	Mean for Retained Stimuli	Cohen's d', 95% CI
Log Probability	-4.05×10^{-6}	-1.52×10^{-6}	1.39 (1.3 , 1.5)
PMI	11.61	7.12	0.85 (0.7 , 1)

Table B2

Comparison of dropped and retained stimuli. Bootstrapped 95% confidence intervals for Cohen's measure of effect size, d', are included.