

# Measuring local context as context–word probabilities

Lance W. Hahn

Published online: 30 September 2011  
© Psychonomic Society, Inc. 2011

**Abstract** Context enables readers to quickly recognize a related word but disturbs recognition of unrelated words. The relatedness of a final word to a sentence context has been estimated as the probability (cloze probability) that a participant will complete a sentence with a word. In four studies, I show that it is possible to estimate local context–word relatedness based on common language usage. Conditional probabilities were calculated for sentences with published cloze probabilities. Four-word contexts produced conditional probabilities significantly correlated with cloze probabilities, but usage statistics were unavailable for some sentence contexts. The present studies demonstrate that a composite context measure based on conditional probabilities for one- to four-word contexts and the presence of a final period represents all of the sentences and maintains significant correlations (.25, .52, .53) with cloze probabilities. Finally, the article provides evidence for the effectiveness of this measure by showing that local context varies in ways that are similar to the N400 effect and that are consistent with a role for local context in reading. The [Supplemental materials](#) include local context measures for three cloze probability data sets.

**Keywords** Cloze probability · Expectancy · Text corpus · Word prediction · N400 · Distributional cues · Conditional probability

---

**Electronic supplementary material** The online version of this article (doi:10.3758/s13428-011-0148-y) contains supplementary material, which is available to authorized users.

---

L. W. Hahn (✉)  
Department of Psychology, Western Kentucky University,  
1906 College Heights Blvd. #21030,  
Bowling Green, KY 42101, USA  
e-mail: Lance.Hahn@WKU.edu

Our environment often contains statistical regularities that, whether we are aware of them or not, can aid our perception (Geisler, 2008) and can foster learning (Saffran, 2003). Although these regularities have influenced our understanding of human linguistic behavior (Misyak, Chistiansen, & Tomblin, 2010; Redington, Chater, & Finch, 1998) in important ways, statistical regularities have been fundamentally important to computational approaches to language. For example, some of the founding work of computational linguistics includes Shannon's (1948) assessment of theoretical communication based on letter and word usage statistics and Harris's (1954) hypothesis that words that appear in similar or identical contexts may be semantically linked.

Since statistical regularities can provide powerful cues for the recognition and the anticipation of text, an efficient human communicator is likely to use these cues in general communication and reading. Saffran and colleagues (Saffran, 2002, 2003; Saffran, Aslin, & Newport, 1996; Saffran & Wilson, 2003) have demonstrated that the statistical properties of spoken language are not only available to, but can be acquired by, infants. Misyak et al. (2010) provided a link between the ability to learn sequence regularities and reading ability. They combined an artificial-grammar learning task with a serial reaction time task into a single task (the AGL-SRT) to evaluate a participant's ability to utilize nonadjacent conditional probabilities with non-word stimuli. In an elegant example of convergent validity, Misyak et al. found that (1) participants were able to learn sequence regularities, (2) strong performance in the AGL-SRT task was associated with quick processing of difficult subject–object relative clauses, and (3) the AGL-SRT behavior was described well by a simple recurrent network computational model. Reading skill appears to be intimately related to sequence learning. Therefore, to fully understand

the processes underlying reading, we need to understand the constraints provided by word sequences.

### Sentence-level expectancy measured by a cloze procedure

An important topic in discourse processing research has been the estimated constraints imposed by a sentence on the final word of the sentence. These constraints are commonly represented as cloze probabilities. A cloze probability is the probability that a participant will choose to end a sentence stem (e.g., “Dillinger once robbed that \_\_\_\_.”) with a specific word (e.g., “bank”). Bloom and Fischler (1980) produced a set of cloze norms that have been used extensively in discourse processing research.

In one of the first event-related potential (ERP) studies examining the influence of expectancy on N400 magnitude, Kutas and Hillyard (1984) used the Bloom and Fischler (1980) cloze norms to operationally define word expectancy. By using the cloze norms, Kutas and Hillyard established a precedent for using the cloze procedure and the Bloom and Fischler norms in ERP research. Kutas and Hillyard concluded that the N400, a negative potential that occurs roughly 400 ms after a stimulus, declined as cloze probability increased, and that the N400 was sensitive to semantic information. More recently, Federmeier, Wlotko, De Ochoa-Dewald, and Kutas (2007) found that the N400 response to unexpected words was not influenced by the level of sentence constraint, but that the N400 did vary with the expectancy of the word, which was operationally defined as the cloze probability of the target word. The link between changes in N400 and cloze probability is quite robust, and it appears reasonable to expect that any sentence attribute that correlates well with the cloze probability is likely to correlate well with N400 changes.

The expectancy generated by a sentence influences the size of the N400 response generated by the final word, but smaller contexts also influence the N400 response. Coulson, Federmeier, Van Petten, and Kutas (2005) found that target words that are preceded by a semantically associated prime produce smaller N400s than do targets that are preceded by an unrelated prime. The magnitude of the N400 can be influenced by the context generated by a single word. Coulson et al. also found that, when reading a complete sentence, sentence context reduces the influence of word–word associativity within the sentence. Although this additional finding suggests that word–word associations within the sentence may play a secondary role during sentence comprehension, word–word associativity and single-word contexts within a sentence may reflect different underlying processes. The contribution of the few words preceding a final word may simply contribute strength to

the global context (Hess, Foss, & Carroll, 1995), or it may include a subject and verb that nonlinearly influence processing of the final word (Duffy, Henderson, & Morris, 1989).

### Measuring local expectancy with usage statistics

The cloze procedure provides a good measure of sentence context, but a measure of local expectancy is needed in order to examine the possible reading processes that utilize the statistical regularities of word sequences. Measuring the context present in a few words raises some difficult challenges for a behavioral approach like the cloze procedure. The brief or highly ambiguous context provided by a few words (e.g., “that \_\_\_\_”) is likely to produce a very diverse collection of endings. As a result, a very large number of participants would be required to properly describe the distribution of word sequence frequencies. Repeating this procedure several times to capture the local context at varying points within a single sentence would be useful for testing hypotheses about the integration of words into a context, but it would be impractical for a collection of many sentences.

An alternative approach to the behavioral estimate of expectancy is found in computational linguistics (see Jurafsky & Martin, 2009, and Manning & Schütze, 1999, for excellent reviews). The intersection of computational linguistics and semantics has become a fairly active research area (see, e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997). One way to represent a language with the intent of predicting the last word in a sequence is with an  $n$ -gram model. An  $n$ -gram model describes conditional probabilities  $p(w_n | w_{n-k}^{n-1})$ , where the last word in the sequence,  $w_n$ , is preceded by a sequence of  $k$  words  $w_{n-k}, w_{n-k+1}, \dots, w_{n-2}, w_{n-1}$  represented as  $w_{n-k}^{n-1}$ . The probabilities are typically estimated from frequencies describing  $n$ -grams (word sequences) of a given length in a text corpus. For instance, the probability that “bank” will follow the context “robbed that” can be expressed as the conditional probability  $p(w_n = \text{“bank”} | w_{n-k}^{n-1} = \text{“robbed that”})$ . This conditional probability is calculated by dividing the number of times the trigram “robbed that bank” occurs by the number of times “robbed that” begins a trigram. Bigrams (two-word sequences) and trigrams (three-word sequences) are typically used as the basis for these calculations, because long  $n$ -gram sequences produce sparse frequency distributions. However, some proposed methods (e.g., Jelinek & Mercer, 1980 as cited in Chen & Goodman, 1998) combine probabilities based on a range of  $n$ -gram sequence lengths in order to form a language model that uses long  $n$ -gram information when it is available and short  $n$ -gram information when long

$n$ -gram information is unavailable. A fundamental hypothesis of the present work is that expectancy generated by local context can be operationally defined using such conditional probabilities. Since the focus of the present work is to capture the local context provided by several words, combining conditional probabilities for  $n$ -grams of different lengths will be a factor in developing the measure of local context.

This computational approach has several benefits and limitations when compared to the cloze procedure. The computational approach overcomes some of the limitations of the cloze procedure by allowing a more exhaustive and objective examination of the possible endings of a sentence. Creating a list of the words that typically follow “that” and their associated frequencies is unrealistic using a behavioral approach but becomes a feasible task with a computational approach. Typical language usage represented in a large corpus will include many different texts and many different authors. The corpus is likely to represent diverse uses of language but, due to its large size, still produces non-idiosyncratic endings.

Probabilities conditioned on long word sequences, such as a complete sentence, are not likely to be available or useful, because individual sentences are unlikely to be frequently repeated. Only the behavioral cloze procedure is capable of assessing context at the level of a complete sentence. The conditional probability approach could, however, be quite useful for describing local context. Phrases or short word sequences have the potential to provide important local context information and the sequences may be repeated frequently enough in a large text corpus to constitute a reasonable sample size. The local-context measure will not replace the cloze procedure, but it would allow hypotheses to be tested about the general role of local context in reading and how local context contributes to the sentence context.

Although the cloze procedure and the local-context probabilities differ in scope and method, they should share some common variance because some of the context measured by the cloze procedure comes from the local context that immediately precedes the final word. Thus, if the conditional probabilities are a valid description of local context, then measures of the local context at the ends of sentences should correlate well with cloze probabilities for the complete sentences.

A prerequisite for calculating conditional probabilities for word sequences is a comprehensive set of word sequence frequencies. Brants and Franz (2006) have created an extensive database (the Google Dataset) of  $n$ -gram frequencies based on an extremely large text corpus. The database describes a very large sample (over 95 billion sentences) of common text usage on all English-language public Web pages that were indexed by Google in January of 2006. The database has been used to evaluate the

relationships between common language usage and the embodied representations of word meaning (Connell & Lynott, 2010; Louwerse, 2008; Louwerse & Connell, 2011; Louwerse & Jeuniaux, 2010), antonymy (Mohammad, Dorr, & Hirst, 2008), and semantic associativity (Hahn & Sivley, 2011). One of the most influential word usage databases related to word recognition in psychology is Kučera and Francis’s (1967) set of word frequencies. As a preliminary test of the psychological relevance of the Google Dataset, a correlation was calculated for the 41,664 word frequencies common to both the Kučera and Francis frequencies and the single-word Google Dataset. The word frequencies were significantly correlated ( $r = .966$ ,  $n = 41,662$ ,  $p < .001$ ), with the words being, on average, more than 340,000 times more frequent in the Google Dataset. Thus, the Google Dataset describes a much larger sample than the Kučera and Francis data set, but it nonetheless maintains an extremely high correlation with the traditional database.

There are, of course, some limitations and challenges associated with the Google Dataset. Although we may suspect that the authors of the original Web text share certain demographic attributes, the demographics of the authors are unknown. Also, the text of a Web page may not necessarily be grammatically correct. Text may or may not be in the form of complete, well-formed sentences. The text corpus may also have different topics than those covered by more traditional text corpuses, such as the Wall Street Journal or the Brown text corpus. The high frequency of the word “sex” (Google Dataset, 719 per million; Kučera & Francis, 86 per million) certainly supports this idea. In some cases, this difference might be viewed as a more diverse, uncensored view of written text, which might tap into semantic relationships with less influence from social expectations. However, it is also reasonable to expect the corpus to have certain linguistic idiosyncrasies, such as an overabundance of “links” and “home pages.” Finally, in order to create the Google Dataset, certain shortcuts or limits were necessary in order to make creation of the data set feasible. The data set only includes  $n$ -grams up to five words in length, and each  $n$ -gram must occur at least 40 times in order to be included in the data set. While these limits are unlikely to influence the probabilities generated from short  $n$ -gram frequencies, they do have the potential to influence statistics generated from long  $n$ -gram frequencies (Chen & Goodman, 1998; Hahn & Sivley, 2011; Zhai & Lafferty, 2004).

### The present study

The general hypothesis considered in this article is that the local context of four words or fewer can be captured by a

combination of conditional probabilities, similar to Eq. 1 below, that describe common language usage. Equation 1 gives the conditional probability that is the analogue for a cloze probability limited to a local context of four words. It is based on a five-word  $n$ -gram,  $(w_1 w_2 w_3 w_4 w_5)$  or  $w_1^5$ , with a four-word local context ( $w_1^4 = L_4$ ) provided for the final word ( $w_5 = F$ ). The numerator is the number of times the five-gram with both the four-word context and the final word ( $w_1^4 = L_4 \wedge w_5 = F$ ) occurs, and the denominator is the number of times a five-gram starts with the four-word context ( $w_1^4 = L_4$ ). To test the validity of a local context measure, the measure is compared with the cloze probability that is widely accepted as a measure of sentence context. As a final test of the local context measure's value, it is further tested within the context of N400 results.

$$p(w_5 = F | w_1^4 = L_4) = \frac{C(w_1^4 = L_4 \wedge w_5 = F)}{C(w_1^4 = L_4)} \quad (1)$$

In Study 1, conditional probabilities were calculated for the five-word sequences ending sentences with published cloze probabilities. One problem that arose in Study 1 was that some five-word sentence endings were absent from the Google Dataset, but shorter forms of the same sentence endings were present. This is the problem of sparse frequency distributions, which computational linguists have investigated. For instance, “and made a big wave” was too infrequent to be included in the Google Dataset, but “made a big wave,” “a big wave,” and “big wave” were all included. When only five-word sequences are considered, this sentence ending has a conditional probability of 0, but in Study 2 the shorter word contexts also contributed to the local context measured. This additional information would potentially increase the number of sentence endings described by the local-context measure and could potentially create an overall increase in the correlation between the subjective cloze probabilities and the local measures based on conditional probabilities.

Although the conditional probability for each local context length may individually provide interesting information, it would be useful to have a single composite measure of local sentence context. The composite measure would describe the impact of the local context as a whole. One challenge produced by the inclusion of additional short contexts is the method by which conditional probabilities should be combined into a single composite measure of local context. Although several strategies are possible, two approaches have been used here: (1) an evenly weighted linear combination of the contributing conditional probabilities and (2) the selection of the single maximal conditional probability.

In Study 3, the contextual influence of the ending period was considered. The hypothesis tested was that requiring

the word produced in a cloze procedure to end the sentence poses an additional constraint that may influence word expectancy. For instance, high-frequency function words (e.g., “the”) are unlikely to end a sentence. The conditional probabilities of the second study were recalculated with the requirement that the final target word was immediately followed by a period. Requiring the target word to end the sentence might further align the conditional probabilities with the cloze probabilities.

Finally, in Study 4 the value of a composite measure of local context based on the measures used in the previous studies was evaluated by comparing the relationship between global context and local context with the relationship between global context and the N400. If changes in the N400 are associated with changes in the local context, then the local context might play a role in the N400 effect during sentence comprehension.

## Study 1

The first goal of this work was to compare local context, described by conditional probabilities, with the sentence context, described by cloze probabilities. Sentence stems, stem completions, and cloze probabilities were acquired from three recently published articles that updated or extended Bloom and Fischler's (1980) norms: Arcuri, Rabe-Hesketh, Morris, and McGuire (2001); Lahar, Tun, and Wingfield (2004); and Block and Baldwin (2010). Arcuri et al. found support for a hypothesized regional influence on cloze probabilities by comparing probabilities measured with inner-city British participants and Bloom and Fischler's North American college student participants. In support of research into aging and language comprehension, Lahar et al. used Bloom and Fischler's sentence stems to gather cloze probabilities with participants from four age groups, ranging from “young” (mean age = 22.5) to “old-old” (mean age = 79.2). Finally, Block and Baldwin examined sentence stems from Bloom and Fischler as well as 398 new sentences, in order to create highly constraining sentence stems with high-cloze-probability dominant responses. Block and Baldwin compared ERP responses to these stem completions with responses to incongruent stem completions. Incongruent stem completions created significantly larger N400 responses than did the congruent stem completions. The difference in N400 responses provides convergent evidence that the congruent sentence completions are linked to the sentence stem context in a physiologically important way.

As described in Eq. 1, a conditional probability can be estimated based on five-word  $n$ -gram statistics with the probability of the last word conditioned on the occurrence of the first four words in the  $n$ -gram sequence. This

conditional probability describes the probability of the stem completion word occurring given the local context of the four words that end the sentence stem. The conditional probabilities were calculated for each stem completion for all sentence stems reported in the three recent articles with cloze probabilities. If these conditional probabilities describe the local context present at the end of the sentence, the probabilities should be significantly correlated with the cloze probabilities, because the sentence context described by the cloze probability includes the local context described by the conditional probabilities. To the extent that the sentence completions rely on context prior to the four-word local context, the correlations will be reduced.

## Method

*Materials* Brants and Franz's (2006) Web 1T 5-gram Version 1 was acquired from the Linguistic Data Consortium and is the basis for all of the frequency and probability calculations. For complete details about the data set, see Brants and Franz (2006), but some relevant details are included here. Brants and Franz (2006) created the Google Dataset frequencies from over 95 billion sentences available to Google through public English-language web pages in January of 2006. Text was parsed into lists of tokens (e.g., words) by separating text at spaces or punctuation marks. This parsing created over one trillion tokens. The Google Dataset includes individual tokens occurring at least 200 times and  $n$ -grams occurring at least 40 times. The frequency data are grouped by sequence length from one token (13.6 million unique entries) to five tokens (1.18 billion unique entries). The conditional probabilities in Study 1 were calculated with the five-gram frequency data.

Sentence stems, stem completions, and cloze probabilities were acquired from Arcuri et al. (2001), Lahar et al. (2004), and Block and Baldwin (2010). It is important to note that these three sources are not completely independent, because they each sampled, in varying degrees, from Bloom and Fischler's (1980) 398 sentence stems. The Arcuri et al. and Lahar et al. articles had the most overlap, with only 42% and 26% of the sentence stems unique to each study. Block and Baldwin had the most unique sentence stems, with 83%. The overlap in sentence stems used across the three studies is considerably reduced when sentence completions are considered. The percentages of sentence completions unique to each article were 67%, 76%, and 84%, respectively. The common sentence completions provided a natural way to examine the consistency across data sets in terms of cloze probabilities and in terms of a measure of local context. This comparison is made later in the present article.

Stem completions containing more than one word (e.g., "tea towel") and sentence stems ending with a proper name or with a word containing an apostrophe were excluded from the analysis.  $N$ -grams containing proper names are unlikely to represent general text usage frequencies because of the idiosyncratic use of the proper name. Words containing apostrophes and multiple-word responses produce usage statistics that differ from the other text because the Google Dataset uses apostrophes to separate words, and a multiple-word response requires two positions in an  $n$ -gram rather than a single word position.

Sentence stems were only excluded when the offending text was the last word in the sentence stem. When the last word in a sentence stem is a proper name or contains an apostrophe, all local contextual information will be affected by it, because all contexts under consideration will include the offending text. For example, "he hurt Jake's \_\_\_\_" only has local context that includes "Jake's," but "Jake took out the \_\_\_\_" has some local context ("took out the") that precedes the completion word that might influence the expected completion. Any sequence that includes a proper name is unlikely to represent general text usage frequencies, because the idiosyncratic use of the proper name typically results in the sequence frequency being reduced to 0. Four-word contexts that included the offending text in the first through third positions were included in Study 1 in order to ease the comparison between Study 1 and Study 2. One of the purposes of Study 2 was to overcome missing sequence frequencies that might result from proper names near the beginning of the sequence. The total number of excluded stem completions was less than 1.5% for each of the three sets of cloze norms.

Arcuri et al. (2001) used 150 sentence stems from Bloom and Fischler (1980) with 2 additional new stems, to which 599 reported stem completions were produced. Exclusions for this data set included 1 sentence stem (with 5 stem completions) that ended with a proper name and 3 multiple-word stem completions from 3 different sentence stems. Lahar et al. (2004) used 119 stems that resulted in 812 stem completions by the "young" participants. Two stem completions were excluded due to multiple-word responses. Block and Baldwin (2010) used 498 sentence stems and reported only the single dominant completion for each stem. Seven of the sentences were excluded because the stem ended with a word containing an apostrophe. One sentence was duplicated in the original norms but was not repeated in our calculations. Across the three different source articles, a total of 1,891 cloze probabilities could be compared with the local context measure after using these exclusion criteria.

*Procedure* For each sentence stem, a five-gram file (or files) in the Google Dataset was selected using the index

files provided with the Google Dataset. The computer programs described in this article were written in Perl and run with Strawberry Perl on Dell computers running Windows XP. The computer program scanned the identified file(s) for lowercase word sequences that began with the sequence stem, immediately followed by the target word. Sequences beginning with uppercase words were excluded for two reasons: A capitalized word might indicate the beginning of a sentence that was not part of the sentence under consideration, and an all-uppercase word might be a label rather than a word embedded in a sentence. The total frequency of the four-word local context ending the stem [ $C(w_1^4 = L_4)$ ] was acquired, as well as the frequency of the four-word local context with each sentence completion word associated with a published cloze probability [ $C(w_1^4 = L_4 \wedge w_5 = F)$ ]. The conditional probability defined in Eq. 1 was calculated with these frequencies.

In the cloze norms, some responses (e.g., “boot” and “boots”) were collapsed into a single cloze probability. Conditional probabilities for these collapsed cloze probabilities were calculated in a fashion similar to the combined  $n$ -gram frequency. For instance,  $p(\text{“boot(s)”} \mid \text{“the dirt from her”})$  was calculated as the sum of  $p(\text{“boot”} \mid \text{“the dirt from her”})$  and  $p(\text{“boots”} \mid \text{“the dirt from her”})$ .

## Results and discussion

The conditional probability for each sentence stem completion is available in the [Supplemental materials](#). Because the three cloze probability data sets were acquired with different participants, used some different sentence stems, and frequently produced different sets of completions for the same sentence stems, the data sets were individually compared with the local-context measure. The correlations between the conditional probabilities produced with Eq. 1 and the published cloze probabilities were significant for all three data sets (see Table 1). The correlations between cloze probabilities and conditional probabilities for the Arcuri et al. (2001) and Lahar et al. (2004) were comparable to the correlations between cloze probabilities in two different samples. For instance, Lahar et al. found that “young” and “old–old” participants had cloze probabilities for 25 weakly constrained sentence stems that were correlated with a Pearson correlation of .47.

The lowest cloze probability in a set of norms was constrained by the number of participants to be  $1/n$  for  $n$  participants. Block and Baldwin (2010) had the largest number of participants, with 377, but their sentences were highly constrained, and their lowest reported cloze probability was .17. Of the three data sets, Arcuri et al. (2001) had the sentence ending with the lowest reported cloze probability of .01. The conditional probabilities for the local context tend to be much lower than the cloze probabilities, because more diversity is available in the words that can complete the context. Arcuri et al. (2001), Lahar et al. (2004), and Block and Baldwin (2010) had average cloze probabilities of .22, .13, and .78, with average conditional probabilities of .09, .05, and .25. There are two restricted-range issues evident in these averages and in Fig. 1: (1) The Block and Baldwin sentences have high cloze probabilities, and (2) the conditional probabilities tend to be relatively low in comparison with the cloze probabilities. The restriction in the range of the conditional probabilities is likely to reduce the ability to detect a linear relationship. Furthermore, the reduced range for the Block and Baldwin sentences might reduce the correlations between cloze probabilities and conditional probabilities in comparison with the other two sets of sentences. Adding quadratic and cubic components in a linear regression did not significantly improve the link between conditional probabilities and cloze probabilities, but the magnitude of the simple linear relationship was sufficient to indicate that an important link between conditional probabilities and cloze probabilities exists.

Although the local context and cloze probabilities share some common variance, within each of these data sets there appear to be both sentences with a measurable local context and sentences without a measureable local context for a range of cloze probabilities. This nonlinear aspect, which should arise when local context and sentence context disagree, suggests that the influences of the two sources of context should be empirically separable. By measuring the local context, sentences can be categorized as having low or high local expectancy for the final word. Having a measure of local context will allow ERP researchers to differentiate between short-range and long-range contextual processes.

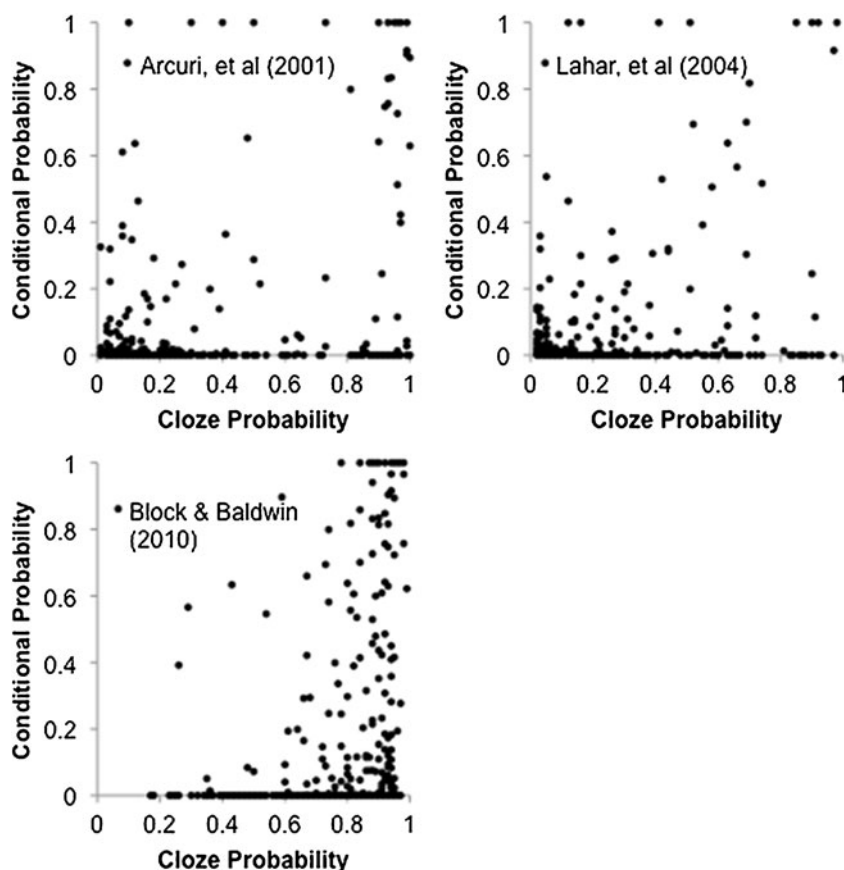
The correlations and sentence totals reported do not include all of the sentences because the relevant local

**Table 1** Correlations between conditional probabilities and cloze probabilities

Data Source	Correlation	Analyzed Completions (Sentences Stems)	Source Completions <sup>1</sup> (Sentences Stems)
Arcuri et al. (2001)	.60*	370 (151)	591 (151)
Lahar et al. (2004)	.58*	488 (119)	810 (119)
Block and Baldwin (2010)	.26*	241 (490)	490 (490)

<sup>1</sup> These sentence and sentence completion counts have been reduced to reflect stimulus exclusions that are described in the text. \*  $p < .001$

**Fig. 1** The conditional probabilities using a four-word context for three published sets of cloze probabilities



context was not always available in the Google Dataset. One limitation of using five-word  $n$ -grams is that, in some cases, the  $n$ -gram frequencies necessary for the calculation were absent from the Google Dataset because the five-word  $n$ -grams did not occur at least 40 times in the text corpus used by Brants and Franz (2006). A conditional probability of 0 was assigned to instances in which the four-word context was present, but the five-word sequence including the context and the sentence completion was absent. However, sentences were excluded from the analysis when both the four-word context and the complete five-word sequence were absent. The magnitude of the missing data can be seen in Table 1, which provides the count of sentence stems and sentence completions analyzed (analyzed completions) and the count of the sentence stems and sentence completions in the original article (source completions). Many sentence stem completions are missing from the analysis.

The loss of sentence stems for long  $n$ -gram sequences is an example of the conflict between focusing on specific details of language usage (using long  $n$ -grams) and focusing on less specific, but more frequent, details of usage (using short  $n$ -grams) recently described by St. Clair, Monaghan, and Christiansen (2010). Their solution to this conflict, which was applied to grammatical categorization, was to have a flexible trigram representation that utilizes

bigram statistics. A similar approach was used in Study 2, with the conditional probabilities associated with one-word, two-word, and three-word contexts augmenting the four-word context probabilities.

A contributing factor to this problem is the use of proper names in the sentences. Because the presence of a proper name in a word sequence reduces the likelihood of the sequence generally occurring, these sequences are under-represented in the Google Dataset. Across all three data sets, the majority of proper names occurred in the first of the four positions in the four-word context, which suggests that including context measures with fewer than four words should reduce this problem.

## Study 2

For several reasons, it is interesting to consider the context provided by word sequences shorter than the four-word contexts used in Study 1. In some instances, more relevant context may be provided by a few words that are near the end of the sentence than by the full four-word context. For instance, in a sentence that ends with “had to clean his house,” the two-word context “clean his” provides almost as much context for “house” as the longer contexts “had to clean his” and “to clean his.” Since “clean his” has a higher

frequency (26,242) than “had to clean his” (227), the percentage of times “house” follows “clean his” (4.9%) may be more representative of the constraints imposed than the percentage based on a three-word (4.5%) or a four-word (0%) context.

The influence of context as a function of the size and location of the context provided has been both an empirical and computational consideration. Tulving and Gold (1963) found that the time required to recognize the final word in a sentence decreases steadily across a range of context lengths from zero to eight words. On the other hand, contextual effects have been proposed that do not diminish across the span of a sentence (Foss, 1982) or that are primarily attributed to the context words of the sentence (Duffy et al., 1989).

Hahn and Sivley (2011) have found that the distance between a word and a contextually relevant word is not a simple inversely proportional relationship, with adjacent words having the largest contextual influence. Hahn and Sivley looked for statistical regularities in the usage statistics of over 62,000 free association word pairs. The conditional probability that a semantic associate would occur within a four-word sequence, given that the sequence began with the target word, was modestly, but significantly, correlated with the association strength between the words based on free association. However, the weakest correlation occurred when the two words were adjacent in the sequence. This result suggests that context might be weakest if it only includes a word that immediately precedes the target word. The influence of context is most likely due to several mechanisms that may be sensitive to proximity, word class, syntactic structure, and context size.

In a computational model of semantic memory (HAL), Lund and Burgess (1996) inferred word–word associations from a text corpus using co-occurrence statistics within a moving ten-word window. The co-occurrence statistics were weighted inversely proportional to the distance between the co-occurring words. In other words, near neighbors were more heavily weighted than distant neighbors. Although the success of HAL suggests that this weighting is generally a good approach, Hahn and Sivley’s (2011) results suggest that a heavy weight for the adjacent word might be counterproductive when considering semantic relationships.

An important variable in this literature is the length of the contextual sequence within a local context. In this study, the effect of context is measured for context lengths from one to three words, with the goal of improving the measure of local context in Study 1 by including these short contexts. Although the value of context is hypothesized to generally increase with increases in context length, by considering a range of contexts it should be possible to

include sentence stems that were excluded in Study 1 due to missing frequency information.

One challenge that arises from considering a range of context lengths is the manner in which the different conditional probabilities for a single stem completion should be chosen or combined into a single composite measure of local context. The three new conditional probabilities and the conditional probability from Study 1 could be combined in a variety of ways. One possibility is that the strongest contextual element will provide the largest influence. This strategy is represented by taking the maximum conditional probability as the sole contextual influence. Another possibility is that the contextual influences combine in an additive fashion. This possibility is modeled as the sum of the four probabilities. Note that this sum is not a probability and has a range from 0 to 4.

## Method

*Materials* As in Study 1, the conditional probabilities were calculated from the Google Dataset frequency data. The frequency data were matched to the length of the context being described. For example, the four-word  $n$ -gram frequencies were the basis for the conditional probabilities evaluating three-word contexts preceding the final word of the sentence, and the three-word  $n$ -gram frequencies were the basis for the conditional probabilities evaluating two-word contexts preceding the final word of the sentence.

*Procedure* The general procedure used in Study 1 was also used in Study 2. Equations 2, 3, and 4 provide the conditional probabilities that a final word (F) will occur conditioned on a one-word context ( $L_1$ ), two-word context ( $L_2$ ), and three-word context ( $L_3$ ). The conditional probabilities for all word sequences ending the target sentences were calculated by dividing the frequency of the stem sequence with the ending word by the frequency of the stem sequence with any additional word. The size of the  $n$ -gram used to estimate the probabilities varied from bigrams to four-grams in Eqs. 2–4, in order to accommodate the size of the context.

$$p(w_2 = F | w_1 = L_1) = \frac{C(w_1 = L_1 \wedge w_2 = F)}{C(w_1 = L_1)} \quad (2)$$

$$p(w_3 = F | w_1 = L_2) = \frac{C(w_1 = L_2 \wedge w_3 = F)}{C(w_1 = L_2)} \quad (3)$$

$$p(w_4 = F | w_1 = L_3) = \frac{C(w_1 = L_3 \wedge w_4 = F)}{C(w_1 = L_3)} \quad (4)$$



The *probability sum* and *probability maximum* composite measures of context defined in Eqs. 5 and 6 were calculated for each sentence stem completion. A Perl computer program calculated the probabilities, and Microsoft Excel was used to organize the probabilities and calculate the composite measures.

$$P_{\text{sum}}(F|L_4) = p(w_2 = F|w_1^1 = L_1) + p(w_3 = F|w_1^2 = L_2) + p(w_4 = F|w_1^3 = L_3) + p(w_5 = F|w_1^4 = L_4), \tag{5}$$

$$P_{\text{max}}(F|L_4) = \max[p(w_2 = F|w_1^1 = L_1), p(w_3 = F|w_1^2 = L_2), p(w_4 = F|w_1^3 = L_3), p(w_5 = F|w_1^4 = L_4)] \tag{6}$$

Results and discussion

The conditional probabilities for each sentence stem completion are available in the [Supplemental materials](#). As hypothesized, the correlation between conditional probability and cloze probability decreased as the length of the context sequence decreased (see Table 2). Conversely, as the sequence length decreased, the number of sentence stem sequences with frequency information in the Google Dataset increased. This is not surprising, because a long word sequence will be less common than a short word sequence that is more likely to occur in a variety of contexts. Typically, long sequences provide a stronger constraint but are less likely to occur than short sequences.

The bigram frequencies required to calculate conditional probabilities based on one-word contexts were available for all sentence completions. As a result, the composite measures provided values for all of the normed stimuli. The composite measures succeeded in increasing the number of sentences with a context measure, but the correlations were smaller than the correlations for conditional probabilities based on the four-word context. As can be seen in Table 3, the two composites were significant for all three data sets, and the composite measures were fairly indistinguishable from one another.

Correlations between the two composite measures were calculated in order to determine whether a further analysis into

**Table 3** Correlations between cloze probabilities and combined conditional probabilities

	Probability Sum	Probability Maximum
Arcuri et al. (2001)	.52*	.52*
Lahar et al. (2004)	.50*	.51*
Block and Baldwin (2010)	.27*	.25*

\*  $p < .001$

the benefits of one measure over the other would be fruitful. The correlations were extremely high, with  $r = .95$  for Arcuri et al. (2001),  $r = .96$  for Lahar et al. (2004), and  $r = .94$  for Block and Baldwin (2010). As a result, it is unlikely that one measure would create a more informative description of the context than the other measure. The sum of probabilities, however, is mathematically awkward because the range of the measure depends on the size of the context, with the maximum value being the length of the largest context considered. The maximum-probability approach produces a well-defined range, and as a result, it will be the composite measure adopted for the rest of this article.

Study 3

Choosing a word to end a sentence in a cloze procedure is constrained by the words preceding the sentence ending, but it may also be constrained by the position of the word at the end of the sentence. For instance, high-frequency function words such as “the” and “an” may appear at almost any position in a sentence, but they are unlikely to end a sentence. This suggests that a conditional probability that requires a sequence to end with a period might reveal an important constraint. The goal of this study was to determine whether producing a final word in a sentence is constrained by the position of the word at the end of the sentence and whether that constraint is a substantial statistical regularity in English text that can contribute to a measure of local context.

Method

*Materials* As in the previous studies, the conditional probabilities were calculated with the Google Dataset frequency

**Table 2** Correlations between cloze probabilities and conditional probabilities for three context lengths and three cloze probability data sets

	One-Word Context	Two-Word Context	Three-Word Context
Arcuri et al. (2001)	.09 (591)	.36* (591)	.44* (546)
Lahar et al. (2004)	.22* (810)	.29* (810)	.42* (736)
Block and Baldwin (2010)	.063 (490)	.18* (488)	.19* (412)

Correlation (N), \*  $p < .001$

data. The frequency data was matched to the length of the context being described. For example, the five-word frequency data were used as the basis for the conditional probability  $p(w_4 = F | w_1^3 = L_3 \wedge w_5 = \text{''.'''})$  evaluating the three-word contexts preceding the final word of the sentence and period.

*Procedure* The procedure used in Study 2 was also used in Study 3, except that a period was included in the final position of each sequence. The conditional probabilities  $p(w_2 = F | w_1^1 = L_1 \wedge w_3 = \text{''.'''})$ ,  $p(w_3 = F | w_1^2 = L_2 \wedge w_4 = \text{''.'''})$ , and  $p(w_4 = F | w_1^3 = L_3 \wedge w_5 = \text{''.'''})$  for all word sequences ending the target sentences were calculated with Eqs. 7, 8, and 9. Equations 7, 8, and 9 restate the conditional probability defined in Eqs. 2, 3, and 4, with a period required in the final position in the  $n$ -gram. Note that because the longest  $n$ -grams included in the Google Dataset have five elements, the inclusion of the period in the final location in the sequence is at the cost of losing one preceding context word.

$$p(w_2 = F | w_1^1 = L_1 \wedge w_3 = \text{''.'''}) = \frac{C(w_1^1 = L_1 \wedge w_2 = F \wedge w_3 = \text{''.'''})}{C(w_1^1 = L_1 \wedge w_3 = \text{''.'''})} \tag{7}$$

$$p(w_3 = F | w_1^2 = L_2 \wedge w_4 = \text{''.'''}) = \frac{C(w_1^2 = L_2 \wedge w_3 = F \wedge w_4 = \text{''.'''})}{C(w_1^2 = L_2 \wedge w_4 = \text{''.'''})} \tag{8}$$

$$p(w_4 = F | w_1^3 = L_3 \wedge w_5 = \text{''.'''}) = \frac{C(w_1^3 = L_3 \wedge w_4 = F \wedge w_5 = \text{''.'''})}{C(w_1^3 = L_3 \wedge w_5 = \text{''.'''})} \tag{9}$$

**Results and discussion**

The conditional probability for each sentence stem completion ending with a final period is available in the [Supplemental materials](#). As can be seen by comparing the correlations in Study 2 with the correlations in Study 3, the pattern of results for word sequences ending with a period (Table 4) was very similar to the pattern for those without

an ending period (Table 2), except that the correlations with cloze probabilities are slightly higher for conditional probabilities that include the period constraint. The increase in correlation only reached significance ( $z = 2.38, p < .01$ ) for the three-word context in the Lahar et al. (2004) sentences, but the increase occurred for all of the data except the two-word context condition with the Block and Baldwin (2010) sentences (see Fig. 2). The correlations between the composite measures and cloze probabilities were slightly lower with the ending period than without the ending period, but this reduction is most likely due to the absence of the four-word context in the composite measures with the period. In summary, the period does provide constraining information, but the loss of the four-word context information resulted in a small overall reduction in the correlation between the composite measures and the cloze probabilities.

**Study 4**

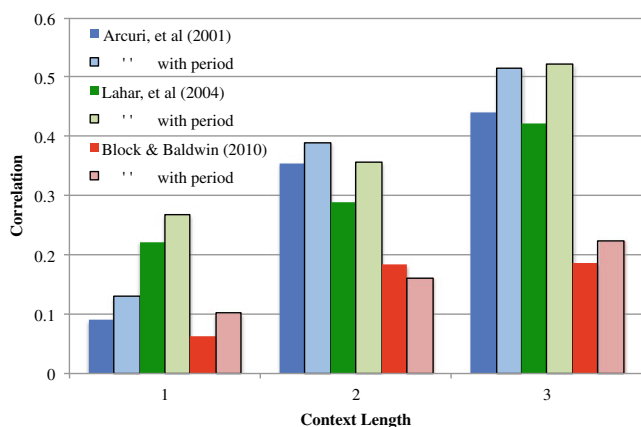
The previous three studies suggested that important local contextual information can be gathered across a range of context lengths and from the constraint that the sequence ends with a period. These different contextual elements can be described in terms of conditional probabilities and combined into a single local-context measure by simply taking the largest of the conditional probabilities. The goal of this study was to describe the local-context measure that is the culmination of the previous three studies and to test the potential importance of local context in reading. The composite measure, local-context probability ( $LCP_{max}$ ), is defined in Eq. 10 for a context, L, preceding the final word of a sentence, F. The equation uses four forms of the context ( $L_1, L_2, L_3,$  and  $L_4$ ), with the subscript number indicating the number of words included in the context.

Sentence endings with different cloze probabilities have a well-documented impact on the N400 produced for the final word in a sentence. If these sentence endings also have different local context strengths as measured by  $LCP_{max}$ , then it is possible that local context played a role in producing the N400 change. Evidence that local context is different in sentences that produced different N400s

**Table 4** Correlations between cloze probabilities and conditional probabilities for context with an ending period

Sentence Source	One-Word Context	Two-Word Context	Three-Word Context	Probability Sum	Probability Maximum
Arcuri et al. (2001)	.13* (591)	.39* (584)	.51* (472)	.49* (591)	.50* (591)
Lahar et al. (2004)	.27* (810)	.36* (810)	.52* (641)	.47* (810)	.47* (810)
Block and Baldwin (2010)	.10 (490)	.16* (477)	.22* (353)	.24* (490)	.23* (490)

N is in parentheses. \*  $p < .001$



**Fig. 2** Correlations between cloze probabilities and conditional probabilities for the three data sets and three context lengths, with and without the additional requirement that the sequence end with a period

would be evidence that local context may influence the N400.

$$LCP_{max}(F|L_4) = \max[p(w_2 = F|w_1^1 = L_1 \wedge w_3 = \text{''.'''}), p(w_3 = F|w_1^2 = L_2 \wedge w_4 = \text{''.'''}), p(w_4 = F|w_1^3 = L_3 \wedge w_5 = \text{''.'''}), p(w_5 = F|w_1^4 = L_4)] \tag{10}$$

**Method**

*Materials* The conditional probabilities of the previous three studies are combined using Eq. 10 to describe the local context for the sentence stems. Since there is some disagreement between the correlations of local context with Block and Baldwin (2010) and the correlations with the Arcuri et al. (2001) and Lahar et al. (2004) sentences that are based on Bloom and Fischler (1980), a new set of 282 sentence stems used in Federmeier et al. (2007) was acquired. The new set of sentences have the additional, and at least equally important, attribute of being categorized in terms of N400 response when the stem was followed by an expected word or followed by a plausible but unexpected word. Four of the Federmeier et al. sentence stems were excluded because they ended with a word that contained an apostrophe.

As in the previous studies, the new conditional probabilities for the new sentences were calculated with the Google Dataset frequency data. The frequency data were matched to the length of the context being described.

*Procedure* The conditional probabilities that comprise the LCP<sub>max</sub> measure defined in Eq. 10 were calculated using the procedure described in Study 3. The four probabilities  $p(w_2 = F|w_1^1 = L_1 \wedge w_3 = \text{''.'''})$ ,  $p(w_3 = F|w_1^2 = L_2 \wedge w_4 = \text{''.'''})$ ,

and  $p(w_4 = F|w_1^3 = L_3 \wedge w_5 = \text{''.'''})$ ,  $p(w_5 = F|w_1^4 = L_4)$  were entered into a Microsoft Excel worksheet for each sentence completion, and LCP<sub>max</sub> was assigned the maximum of the four probabilities.

**Results and discussion**

As with the previous measures, LCP<sub>max</sub> describes local context in a way that is significantly correlated with the broader sentence context described by cloze probability. LCP<sub>max</sub> did produce a modest but consistent improvement over the previous composite measures of local context (see Table 5). LCP<sub>max</sub> also produced a significant correlation with the cloze probabilities for the Federmeier et al. (2007) sentences that was only slightly smaller than those for the Arcuri et al. (2001) and Lahar et al. (2004) sentences. The addition of quadratic and cubic factors in a linear regression did not significantly improve the link between LCP<sub>max</sub> and cloze probabilities. Each of the data sets plotted in Fig. 3 include local contexts at or near the ceiling of 1. This highlights the fact that the LCP<sub>max</sub> measure, by definition, assumes the value of the largest probability included within the scope of the local context. It also indicates that some of the sentences end with clichéd phrases (e.g., “pigs wallowed in the mud”) and highly constraining local context (e.g., “the engine failed to start”).

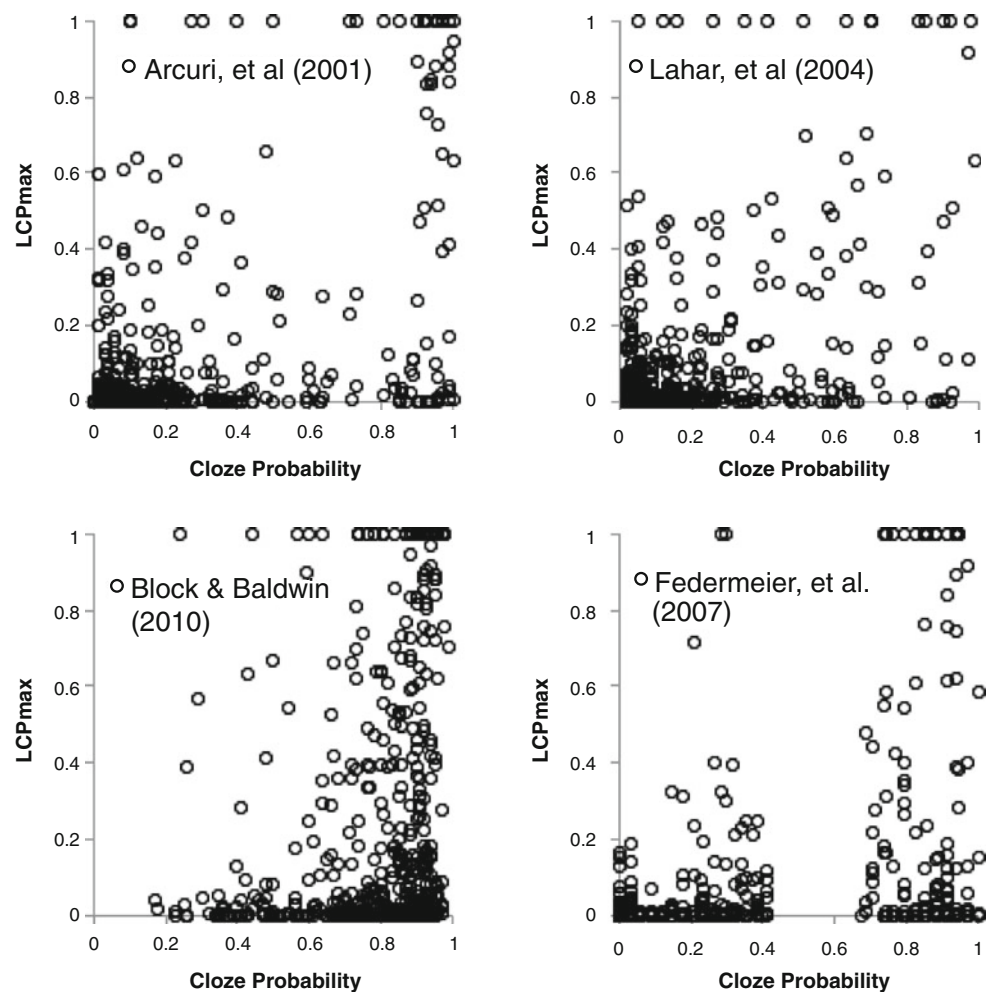
Federmeier et al. (2007) categorized their sentence stems into weakly constrained and strongly constrained stems, with both an unexpected ending and an expected ending for each stem. The stratification of sentences into weakly constrained and strongly constrained sentences is apparent from the gap in cloze probabilities from .42 to .67 in Fig. 3. Federmeier et al. found that (1) expected and unexpected endings produced significantly different N400 effects, (2) expected endings for strongly and weakly constrained sentences produced significantly different N400 effects, and (3) unexpected endings were not significantly different when preceded by the two levels of sentence constraint. Since N400 values are not currently available for individual sentences, we cannot reanalyze the N400s as a function of the local context for individual sentences. However, it is possible to determine whether or not local context varies

**Table 5** Correlations between cloze probabilities and the LCP<sub>max</sub> measure of local context

Sentence Source	LCP <sub>max</sub>
Arcuri et al. (2001)	.53* (591)
Lahar et al. (2004)	.52* (810)
Block and Baldwin (2010)	.25* (490)
Federmeier et al. (2007)	.47* (556)

N is in parentheses. \* p < .001

**Fig. 3** The  $LCP_{max}$  values for four published sets of cloze probabilities



with the categorization of each sentence stem as weakly constraining or highly constraining, and whether or not local context varies with the categorization of each sentence ending as unexpected or expected. Federmeier et al. summarized the observed relationship between N400 amplitudes and cloze probabilities as a stratification, with strongly constrained expected endings producing smaller N400s than weakly constrained expected endings, and unexpected endings producing the largest N400s. If the pattern of results linking these levels of constraint and expectancy to the N400 effect was similar to the pattern of results between these attributes and local context, this would be consistent with the idea that local context in the sentences contributes to the N400 effect. However, even with very similar results, the correspondence between the  $LCP_{max}$  and cloze probabilities makes it impossible to attribute N400 changes to either local or sentence contextual influences without additional experiments that separate the two factors.

A repeated measures ANOVA was conducted on the  $LCP_{max}$  for each sentence completion. As in the original Federmeier et al. (2007) article, sentences were grouped

into strongly expected, strongly unexpected, weakly expected, and weakly unexpected conditions for analysis. The constraint main effect [ $F(1, 136) = 32.5, p < .0001$ ], the expectancy main effect [ $F(1, 136) = 81, p < .0001$ ], and the interaction between constraint and expectancy [ $F(1, 136) = 31, p < .0001$ ] were all significant. As in the original analysis of N400 amplitudes, a comparison of strongly unexpected and weakly unexpected sentences did not produce a significant difference [ $F(1, 136) = 0.63, p = .429$ ]. This pattern of significant main effects and an interaction for constraint and expectancy, with a nonsignificant difference between unexpected endings that differed in sentence constraint, is the same pattern of results Federmeier et al. reported with the N400 effect.

## General discussion

The broad focus of this article is an examination of local context for statistical regularities that may aid and constrain normal reading. The two main goals were (1) to develop an informative measure of local context and (2) to determine

the potential importance of the local context measure for understanding sentence comprehension. Since cloze probabilities measure contextual influence at the sentence level, and since they have been intimately linked to discourse processing, cloze probabilities have been used as a validity test for the development of the local-context measure. A valid measure of local context should positively correlate with cloze probabilities.

#### Measuring local context

Context plays an important role in sentence and discourse processing. Although recent research has primarily focused on sentence-level context, there is support for the idea that contextual influences come from portions of a sentence (Duffy et al., 1989; Tulving & Gold, 1963).  $LCP_{max}$  is a measure of the local context provided at the end of a sentence. It is strongly correlated with a sentence's cloze probability, but has a smaller contextual focus. The  $LCP_{max}$  measure and the local-context measures included in the [Supplemental materials](#) will enable researchers to separate sentence context from local context at the end of a sentence and will enable the influence of local context to be directly examined. Furthermore, with the proper computational resources, conditional probabilities describing local context can be calculated fairly easily for any arbitrary sentence fragment. Behavioral approaches such as the cloze procedure, on the other hand, require a researcher to gather responses for each context in hopes that the targeted stem completion will be included in the response distribution. For example, the local expectancies created by “a big” for “splash” and “wave” can be quickly estimated based on conditional probabilities, but analogous behavioral estimates would require many participants in order to be confident that “splash” and “wave” would be produced in the set of responses. One important benefit of the computational approach is that evaluating local context is feasible, whereas it would not be feasible with a behavioral approach.

A measure of local context should capture the relevant elements of context, as well as combine the elements in a form that maximizes both the information available and the usefulness of the measure. In Study 1, local context was described as the conditional probability that a word completing a sentence would occur given the local context of the four words ending the sentence stem. Although this measure correlated well with the cloze probabilities associated with the sentence stem completions, the text corpus frequencies did not include all of the sentence stem endings, and as a result, many sentence stems could not be described by this measure (see Table 1). One theoretical solution to this limitation would be to acquire a more complete sample of English text—a larger text corpus

database. However, given that the Google Dataset includes over 95 billion sentences and that the generative nature of language is unlikely to ever be fully captured, a more reasonable approach is to further consider the local context provided by word sequences that are shorter and more frequent than the four-word sequences.

Study 2 demonstrated that measures of local context for a complete list of sentences can be acquired by considering contexts ranging from one to four words. Two composite measures were considered: (1) summing the conditional probabilities for all four context lengths and (2) using the maximum conditional probability as the representative of local context. Both composite measures of local context correlated well with cloze probability. The maximum-local-probability approach was adopted because it captures important contextual variation, has a defined range from 0 to 1, and maintains a direct link to the original  $n$ -gram frequency data.

The period ending a sentence provides an additional element of context that modestly, but consistently, improves the correlations between the local context measure and cloze probabilities. Because the longest  $n$ -grams available in the Google Dataset contain five words or punctuation marks, including a period after the final word in a sentence is only possible with at most three preceding words of context.  $LCP_{max}$ , defined by Eq. 10, is a composite measure of local context that describes local context for up to three words preceding a target word followed by an ending period, but also includes the full context provided by the four words preceding the ending word in the absence of a period. This measure of local context represents the culmination of the first three studies. With the exception of sentence stems that end with proper names or a word that contains an apostrophe, it describes the local context for all of the sentences from the four cloze data sets and is significantly correlated with the cloze probabilities ( $r = .25-.53$ ).

Despite the uniformly significant correlations between cloze probabilities and  $LCP_{max}$ , there was some variability across the sets of sentences. In particular, sentences from the Block and Baldwin (2010) study produced consistently lower correlations between  $LCP_{max}$  and their cloze probabilities than did the sentences from the other studies. Because understanding this difference may illuminate some of the strengths and shortcomings of a corpus-based approach, several possible explanations of the difference are considered.

One possible explanation for this difference might be that Block and Baldwin (2010) constrained the sentence completions primarily with sentence-level constraints that included only limited local context influence. Since a long sentence introduces more distant context than does a short sentence, correlations between  $LCP_{max}$  and cloze probabil-

ities were calculated for short sentences (5–9 words) and for long sentences (10–14 words). The short sentences did have a higher correlation ( $r = .28$ ,  $n = 303$ ) than the long sentences ( $r = .22$ ,  $n = 187$ ), but the difference was not significant ( $z = 0.68$ , n.s.). Block and Baldwin introduced nearly 400 new sentences in addition to the nearly 100 sentences used to replicate Bloom and Fischler (1980). To determine whether the reduced correlations might be limited to only the new sentences, correlations were calculated between  $LCP_{max}$  and cloze probabilities for the new sentences and for the sentences Block and Baldwin (2010) used to replicate the Bloom and Fischler norms. The Bloom and Fischler replication sentences had a lower correlation ( $r = .22$ ,  $n = 99$ ) than did the new sentences ( $r = .28$ ,  $n = 391$ ), but it was not significant ( $z = .56$ , n.s.). This result is not consistent with the new sentences being less reliant on local context than the replicating sentences.

The Block and Baldwin (2010) norms are also different from the Arcuri et al. (2001) and Lahar et al. (2004) norms in that they only include a single dominant response for each sentence stem. As a result, all of the variance in their cloze probabilities is between different sentence contexts. The other two sets of norms provide multiple completions for each sentence stem. Comparing cloze probabilities or  $LCP_{max}$  across sentences for only the dominant responses should be done with caution, because the probability of the completion is somewhat dependent on the number of alternative completions available—the sentence stem constraint. A stem completion that occurs 5% of the time is a dominant response if the alternative completions are numerous and have smaller frequencies, but it may be a weak response if the alternative completions are few and have larger frequencies. Comparing only dominant responses focuses on the between-sentence variability and ignores the within-sentence variability that influences final-word selection. To determine whether the low correlations for the Block and Baldwin sentences was due to only dominant stem completions being included in the analysis, correlations were calculated for only the dominant stem completions in the Arcuri et al. and Lahar et al. norms. The correlations between cloze probabilities and  $LCP_{max}$  decreased from .53 and .52 to .45 and .41, respectively. Some of this reduction may be due to the dramatic reduction in

stem completions (from 591 to 152 and 810 to 119, respectively).

Finally, as is apparent in Fig. 3, Block and Baldwin (2010) differs from the other three studies in the range of cloze probabilities. Block and Baldwin intentionally focused on high-constraint sentences, and in doing so, they reduced the range of cloze probabilities. This reduction in range could also be the reason that the correlation for this set of sentences is lower than for the other data sets.

There does not appear to be a single explanation for the low Block and Baldwin (2010) correlations. Sentence length, the introduction of new sentences, the use of only dominant stem completions, and the focus on highly constraining sentences are all likely to contribute to the differences between the Block and Baldwin correlations and the correlations with the other cloze norms. To provide some sense of how correlations between  $LCP_{max}$  and cloze probability compare with cloze probability measures from different norming experiments, correlations were calculated for the Bloom and Fischler (1980) sentence completions that were common to at least two of the three published cloze probability norms (see Table 6). The correlations are significant across all of the data sets. Although the  $LCP_{max}$  and cloze correlation is lower for the Block and Baldwin sentences than for the other sentences, the correlation is quite similar to the correlation between the Arcuri et al. and Block and Baldwin studies when both measure cloze probabilities.

#### The impact of corpus-based estimates of local context

Cloze probabilities have been an important measure of sentence context for investigations of discourse processing with ERPs. The N400 response changes distinctively as a function of the expectancy at the end of a sentence, and both expectancy and contextual constraint are commonly defined in terms of cloze probabilities. The traditional cloze probability, however, describes the context provided by the whole sentence stem and provides little insight into whether the important contextual elements occur early or late in the sentence. Measuring the local context present at the end of a sentence using a behavioral approach would be difficult, but, as has been demonstrated in this article, a computational approach can work quite well. Although this article

**Table 6** Correlations between context measures for sentence completions common to at least two of the published cloze norms

	$LCP_{max}$	Block and Baldwin (2010) Cloze	Lahar et al. (2004) Cloze
Arcuri et al. (2001) cloze	.57** (196)	.31* (44)	.87** (161)
Lahar et al. (2004) cloze	.52** (197)	.64** (45)	
Block and Baldwin (2010) cloze	.31** (80)		

Correlation ( $n$ ). \*  $p < .05$ , \*\*  $p < .01$

has focused on the end of the sentence, because the cloze probability provides a well-established measure of context with which the new local-context measure can be compared, this computational measure of local context can be easily applied to any local span of a sentence by removing the period constraint from the probabilities.  $LCP_{\max}$  represents an initial use of corpus-based probabilities to create a more acute measure of context than the cloze probability.

In addition to demonstrating that a single-word context could influence the N400 response to a subsequent word, Coulson et al. (2005) pitted this single-word association effect against the whole-sentence context measured by cloze probabilities. Associated and unassociated word pairs were embedded at the ends of sentences that were either congruent or incongruent with the final target word. With these sentence stimuli, both the associativity of the embedded word pair and the sentence congruity could be active, but only the sentence congruity generated a significant change in the N400. As measured by ERPs, word associativity had very little influence on reading the final word of a sentence, even though associativity does impact the processing of a word when the word pair is presented without the sentence context.

Although the statistical regularities within an  $n$ -gram are modestly related to semantic associativity (Hahn & Sivley, 2011),  $LCP_{\max}$  is not a measure of semantic associativity, but rather an estimate of the likelihood that a word will end a sentence based on the statistical regularities associated with the local context. For example, Coulson et al. (2005) provided “bits” and “chocolate” as an example of an unassociated word pair, but “bits of” does provide some local context for “chocolate” [ $p(w_3 = \text{“chocolate”} \mid w_1^2 = \text{“bits of”} \wedge w_4 = \text{“.”}) = .0016$ ]. Although this remains an empirical question, an extension of Coulson et al.’s findings from sentence context to local context would have local context also overshadowing the influence of word–word associativity. However, the reduction in context from a whole sentence to a few words might allow for a modest word–word associativity effect. With the new measure of local context developed in this article, it is now possible to ask such empirical questions.

The ability to describe local context will also make questions regarding the impact of changing context size and proximity empirically approachable. Electrophysiological studies have often focused exclusively on the final word of a sentence, because context can be operationally defined as a cloze probability. With local context now operationally defined, contrasting the influence of local context with word associativity and with sentence context can be examined with ERP studies like Coulson et al.’s (2005) contrast of word associativity with sentence context. The association between local context and Federmeier et al.’s (2007) N400s found in Study 4 suggests that local context

is important, but a true experiment will be needed to determine whether sentence context or local context is responsible for N400 changes. The [Supplemental materials](#) include sentences both low and high in local context, which would be necessary for such a study.

Finally, it is clear that young children (Saffran et al., 1996) and adults (Remillard, 2010) are sensitive to statistical regularities over a range of items. It also appears that being skilled at learning an artificial grammar in a sequential reaction time task is associated with strong reading skills (Misyak et al., 2010). With this new measure of local context based on inherent statistical regularities, it should now be possible to more fully examine the intersection between statistical regularity learning and how English text is integrated into semantic memory in terms of both typical and atypical reading.

**Author note** I am grateful to Kara Federmeier for sharing sentence stimuli and cloze probabilities. This work was made possible with computational support from the College of Education and Behavioral Sciences, the Educational Technology Center, and the Department of Psychology at Western Kentucky University. [Supplemental materials](#) may be downloaded along with this article from [www.springerlink.com](http://www.springerlink.com).

## Supplemental materials

Supplemental data set files provide the calculated conditional probabilities that are the basis for these studies. All of the calculations were based on the frequencies available in Brants and Franz’s (2006) Google Dataset. For each calculation, the frequencies were acquired from the Google data set that had the same number of words as the word sequence of interest. For example, the conditional probability  $p(\text{“bread”} \mid \text{“loaf of”})$  was derived by dividing the frequency of “loaf of bread” in the three-word  $n$ -gram Google data set by the frequency of “loaf of” beginning the  $n$ -grams in the three-word  $n$ -gram Google data set. In some instances, the sentence stem sequence was absent from the Google Dataset. When this occurred, the reported conditional probability was not available.

The conditional probabilities for Study 1 and Study 2 did not require that the sequence end with a period. For all sentence stems with contexts  $L_1$ ,  $L_2$ ,  $L_3$ , or  $L_4$  preceding the sentence completion (F), the reported probabilities, from the shortest to the longest context, are  $p(w_2 = F \mid w_1^1 = L_1)$ ,  $p(w_3 = F \mid w_1^2 = L_2)$ ,  $p(w_4 = F \mid w_1^3 = L_3)$ , and  $p(w_5 = F \mid w_1^4 = L_4)$ .

The conditional probabilities for Study 3 required that the sequence end with a period. Since the longest  $n$ -grams in the Google Dataset have five elements, and the  $n$ -gram must include the final response word and the period, the largest context available had three words. For all sentence stems with contexts  $L_1$ ,  $L_2$ , and  $L_3$  preceding the sentence

completion (F), the reported probabilities from the shortest to the longest context are  $p(w_2 = F \mid w_1^1 = L_1 \wedge w_3 = \text{“.”})$ ,  $p(w_3 = F \mid w_1^2 = L_2 \wedge w_4 = \text{“.”})$ , and  $p(w_4 = F \mid w_1^3 = L_3 \wedge w_5 = \text{“.”})$ . The composite measures described in the text are also included.

Column Number	Attribute
1	Sentence number in original cloze probability norms
2	Complete sentence
3	Four-word context stem ( $L_4$ )
4	Response word (F)
5	$p(w_2 = F \mid w_1^1 = L_1)$
6	$p(w_3 = F \mid w_1^2 = L_2)$
7	$p(w_4 = F \mid w_1^3 = L_3)$
8	$p(w_5 = F \mid w_1^4 = L_4)$
9	$P_{\text{sum}}(F \mid L_4)$
10	$P_{\text{max}}(F \mid L_4)$
11	$p(w_2 = F \mid w_1^1 = L_1 \wedge w_3 = \text{“.”})$
12	$p(w_3 = F \mid w_1^2 = L_2 \wedge w_4 = \text{“.”})$
13	$p(w_4 = F \mid w_1^3 = L_3 \wedge w_5 = \text{“.”})$
14	$P_{\text{sum}}(F \mid L_4 \wedge \text{“.”})$
15	$P_{\text{max}}(F \mid L_4 \wedge \text{“.”})$
16	LCP <sub>max</sub>

## References

- Arcuri, S. M., Rabe-Hesketh, S., Morris, R. G., & McGuire, P. K. (2001). Regional variation of cloze probabilities for sentence contexts. *Behavior Research Methods, Instruments, & Computers*, 33, 80–90. doi:10.3758/BF03195350
- Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, 42, 665–670. doi:10.3758/BRM.42.3.665
- Bloom, P. A., & Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory & Cognition*, 8, 631–642.
- Brants, T., & Franz, A. (2006). *Web 1T 5-gram (Version 1) [Software]*. Philadelphia: Linguistic Data Consortium.
- Chen, S. F., & Goodman, J. (1998). *An empirical study of smoothing techniques for language modeling* (Tech. Rep. TR-10-98). Harvard University.
- Connell, L., & Lynott, D. (2010). Look but don't touch: Tactile disadvantage in processing modality-specific words. *Cognition*, 115, 1–9. doi:10.1016/j.cognition.2009.10.005
- Coulson, S., Federmeier, K. D., Van Petten, C., & Kutas, M. (2005). Right hemisphere sensitivity to word- and sentence-level context: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 129–147. doi:10.1037/0278-7393.31.1.129
- Duffy, S. A., Henderson, J. M., & Morris, R. K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 791–801. doi:10.1037/0278-7393.15.5.791
- Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, 1146, 75–84.
- Foss, D. J. (1982). A discourse on semantic priming. *Cognitive Psychology*, 14, 590–607. doi:10.1016/0010-0285(82)90020-2
- Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59, 167–192. doi:10.1146/ann.v.psych.58.110405.085632
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114, 211–244.
- Hahn, L. W., & Sivley, R. M. (2011). Entropy, semantic relatedness and proximity. *Behavior Research Methods*, 43. doi:10.3758/s13428-011-0087-7
- Harris, Z. S. (1954). Distributional structure. *Word*, 10, 146–162.
- Hess, D. J., Foss, D. J., & Carroll, P. (1995). Effects of global and local context on lexical processing during language comprehension. *Journal of Experimental Psychology: General*, 124, 62–82. doi:10.1037/0096-3445.124.1.62
- Jelinek, F., & Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.
- Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307, 161–163. doi:10.1038/307161a0
- Lahar, C. J., Tun, P. A., & Wingfield, A. (2004). Sentence-final word completion norms for young, middle-aged, and older adults. *Journals of Gerontology*, 59B, P7–P10.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.
- Louwerse, M. M. (2008). Embodied representations are encoded in language. *Psychonomic Bulletin & Review*, 15, 838–844.
- Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, 35, 381–398. doi:10.1111/j.1551-6709.2010.01157.x
- Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nat. of conceptual processing. *Cognition*, 114, 96–104.
- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28, 203–208. doi:10.3758/BF03204766
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, 2, 138–153. doi:10.1111/j.1756-8765.2009.01072.x
- Mohammad, S., Dorr, B., & Hirst, G. (2008, October). *Computing word-pair antonymy*. Paper presented at the Conference on Empirical Methods in Natural Language Processing, Waikiki, Hawaii.
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425–469. doi:10.1016/S0364-0213(99)80046-9
- Remillard, G. (2010). Implicit learning of fifth- and sixth-order sequential probabilities. *Memory & Cognition*, 38, 905–915. doi:10.3758/MC.38.7.905



- Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*, 172–196. doi:10.1006/jmla.2001.2839
- Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*, 110–114.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. doi:10.1126/science.274.5294.1926
- Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy*, *4*, 273–284.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(379–423), 623–665.
- St. Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*, 341–360. doi:10.1016/j.cognition.2010.05.012
- Tulving, E., & Gold, C. (1963). Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology*, *66*, 319–327.
- Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, *22*, 179–214. doi:10.1145/984321.984322