Behav Res (2018) 50:826–833
DOI 10.3758/s13428-017-0908-4

The Provo Corpus: A large eye-tracking corpus with predictability norms

Steven G. Luke[1] · **Kiel Christianson**[2,3]

Published online: 18 May 2017

Abstract This article presents the Provo Corpus, a corpus of eye-tracking data with accompanying predictability norms. The predictability norms for the Provo Corpus differ from those of other corpora. In addition to traditional cloze scores that estimate the predictability of the full orthographic form of each word, the Provo Corpus also includes measures of the predictability of the morpho-syntactic and semantic information for each word. This makes the Provo Corpus ideal for studying predictive processes in reading. Some analyses using these data have previously been reported elsewhere (Luke & Christianson, 2016). The Provo Corpus is available for download on the Open Science Framework, at https://osf.io/sjefs.

Keywords Corpus study · Eyetracking · Reading · Predictability

In the present article, we introduce the Provo Corpus, a large corpus of eye-tracking data with accompanying predictability norms. The primary purpose of the Provo Corpus is to facilitate the investigation of predictability effects in reading. Some analyses of the data available in the Provo Corpus are reported in Luke and Christianson (2016). The corpus is publicly available, and can be downloaded from the Open Science Framework at https://osf.io/sjefs.

✉ Steven G. Luke
steven_luke@byu.edu

[1] Department of Psychology and Neuroscience Center, Brigham Young University, 1062 SWKT, Provo, UT 84602-5543, USA

[2] University of Illinois at Urbana-Champaign, Urbana, IL, USA

[3] The Beckman Institute for Advanced Science and Technology, New Haven, CT, USA

Prediction in language processing is a topic that has received considerable attention in recent years. It has been the subject of a number of reviews (DeLong, Troyer, & Kutas, 2014; Huettig, 2015; Huettig & Mani, 2016; Kuperberg & Jaeger, 2016; Kutas, DeLong, & Smith, 2011; Staub, 2015; Van Petten & Luka, 2012) and is a significant component in many models of language processing (Christiansen & Chater, 2016; Dell & Chang, 2014; Pickering & Garrod, 2007, 2013). Predictability is known to influence how we process language, both spoken (Altmann & Kamide, 1999, 2007; Kamide, Altmann, & Haywood, 2003; Staub, Abbott, & Bogartz, 2012) and written (Ashby, Rayner, & Clifton, 2005; Balota, Pollatsek, & Rayner, 1985; Ehrlich & Rayner, 1981; Kennedy, Pynte, Murray, & Paul, 2013; Kliegl, Grabner, Rolfs, & Engbert, 2004; Rayner, Slattery, Drieghe, & Liversedge, 2011; Rayner & Well, 1996).

The most common way to establish the predictability of a given word is through the cloze procedure (Taylor, 1953). In this procedure, participants are presented with a portion of a sentence or passage up to the word of interest and then asked to produce the word that would be most likely to follow. Traditionally, this method has been used to assess the predictability of a single word, usually a noun, in either a highly constraining or a nonconstraining sentence context. Many sets of predictability norms have been made publicly available (e.g., Bloom & Fischler, 1980; Schwanenflugel, 1986). The cloze procedure, along with the predictability norms that have been made publicly available using that procedure, has greatly facilitated research into predictive processes.

A useful method for studying prediction in reading is the collection of eye-tracking data. Participants in these studies read sentences or passages in which the predictability of one or more words is already known (Kennedy et al., 2013; Kliegl et al., 2004; Rayner & Well, 1996) while their eye movements are monitored. These types of studies have revealed much

about how prediction affects reading (Staub, 2015). A few corpora of eye movement data exist (Cop, Dirix, Drieghe, & Duyck, 2017; Kennedy, Hill, & Pynte, 2003; Kennedy et al., 2013; Kliegl et al., 2004; Kliegl, Nuthmann, & Engbert, 2006), with varying degrees of availability.

The Provo Corpus consists of two parts, predictability norms and eye-tracking data. The predictability norms consist of completion norms for every word in 55 paragraphs. The eye-tracking corpus consists of eye movement data from 84 native English-speaking participants, who read all 55 paragraphs for comprehension. Below, we compare and contrast both the predictability norms and the eye-tracking corpus with existing norms and corpora, pointing out similarities and differences. Then we discuss potential uses of the Provo Corpus. Next, we describe in detail the contents of the corpus—first the predictability ratings, and then the eye-tracking data. Finally, we describe how interested parties can get access to the corpus.

Comparison of the Provo predictability norms with other extant norms

A number of predictability norming studies have been published over the years. Notable among these are Bloom and Fischler (1980) and Schwanenflugel (1986). These studies are sentence completion norms: A sentence was presented, minus the final word, and participants were asked to produce the final word. More recent predictability norms have followed a similar procedure (see, e.g., Hamberger, Friedman, & Rosen, 1996; McDonald & Tamariz, 2002).

The predictability norms in the Provo Corpus differ from these other published norms in several significant ways. As we mentioned, the existing norms are all sentence completion norms, meaning that they involve single sentences, in which only the last word in the sentence is normed. The Provo predictability norms are paragraphs, rather than sentences, and norms are provided for each word in the paragraph, rather than just the final word. Although traditional sentence completion norms are well-suited to event-related potential (ERP) and eye-tracking experiments that manipulate the predictability of a single target word in a sentence, the Provo norms are ideal for studies in which responses (such as reading times or ERPs) are examined for every word (see, e.g., Luke & Christianson, 2016; Payne, Lee, & Federmeier, 2015; Smith & Levy, 2013). Furthermore, traditional predictability norms focus heavily on highly constraining sentences (cloze scores > .67), which turn out to be relatively rare in connected texts (Luke & Christianson, 2016); the Provo corpus provides a more naturalistic distribution of predictability. Additionally, where existing predictability norms focus exclusively on content words, especially nouns, the Provo predictability norms include norms for function words as well as for a wider variety of content words (adverbs, adjectives, and verbs are more well-represented).

Comparison of the Provo Corpus with other eye-tracking corpora

Several other eye-tracking corpora exist. Among these, the Ghent Eye-Tracking Corpus is notable, because it is large (participants read an entire novel) and publicly available (Cop et al., 2017). However, two other well-known corpora deserve special mention, because predictability ratings are available for these corpora: the Dundee Corpus and the Potsdam Sentence Corpus.

The Dundee Corpus (Kennedy et al., 2003; Kennedy et al., 2013) is a large corpus of eye movements from ten native English speakers (and ten native French speakers) reading texts from newspaper editorials (56,212 tokens). Texts were presented on-screen in a multiline format. For a subset of the texts (16 four-line paragraphs), predictability data were obtained for each word (272 participants total, making approximately 25 responses per word). The Provo Corpus is similar to the Dundee Corpus in that it is a corpus of texts, but the Provo Corpus drew on both more participants and more texts for its predictability norms.

The Potsdam Sentence Corpus (Kliegl et al., 2004; Kliegl et al., 2006) is a collection of 144 German sentences, with predictability estimates (cloze scores) available for all but the first word in each sentence. These predictability norms were obtained using a cloze procedure, in which 272 native German speakers provided responses, producing a total of 83 complete predictability protocols. The eye-tracking corpus consists of data from 222 participants reading these sentences. Like the Potsdam Corpus, the Provo Corpus contains predictability norms for all words. The Provo Corpus has 134 sentences total, but it differs from the Potsdam Sentence Corpus in that these sentences were presented as part of connected multiline texts, rather than in isolation.

There is an additional, significant difference between the Provo Corpus and these other corpora with predictability ratings. In all three corpora, cloze scores are included for all normed words, and these cloze scores represent the proportions of responses provided by participants in the cloze procedure that matched the target word orthographically (e.g., if the target word was "apple" and the response was "apple," that is a match; "turtle," "fruit," and "red delicious" are not matches). However, some theorists argue that prediction is a graded process (for a review, see Kuperberg & Jaeger, 2016), so even if the context is not sufficiently constraining to permit the prediction of orthography, it may still permit the prediction of morpho-syntactic or semantic information. For example, for the paragraph that begins "With schools still closed, cars still buried and streets still," it is unlikely that most readers

will form a strong prediction that the next word will be "blocked" (the cloze score for this word was only .07 in our predictability norming study). However, readers should be able to predict with some accuracy that the next word will be a verb (it follows a noun and an adverb, after all), that it will be in the past tense (the other verbs in the sentence were), and maybe even that the verb will mean something similar to "blocked," like closed or inaccessible. Indeed, the participants in our predictability norming study produced a verb 79% of the time when given the sentence fragment above. That verb was in the past tense most of the time (72% of all responses) and was semantically related to the target word "blocked" (the two most frequent responses were "closed" and "covered"). With this in mind, the Provo Corpus contains predictability ratings for word class and (where appropriate) inflection, as well as mean semantic relatedness scores (latent semantic analysis; see Landauer & Dumais, 1997, and below for more information) that represent the semantic similarity between the target word and cloze task responses. These additional ratings quantify the predictability of the morpho-syntactic (word class, inflection) and semantic information, permitting a deeper investigation into the graded nature of prediction. See Luke and Christianson (2016) for some examples of analyses using these variables.

Potential uses of the Provo corpus

The Provo Corpus is primarily intended for the study of prediction in reading; however, its usefulness is not restricted to this purpose. The Provo Corpus is a large data set of the eye movements of skilled readers reading connected text. As such, it should prove useful for studying other aspects of reading behavior and for evaluating models of eye movement control in reading. The Dundee and Potsdam Corpora have already proven invaluable in this regard (see, e.g., Engbert, Nuthmann, Richter, & Kliegl, 2005; Kennedy et al., 2013; Kliegl & Engbert, 2005; Kliegl et al., 2004; Nuthmann, Engbert, & Kliegl, 2007; Pynte, New, & Kennedy, 2009; Smith & Levy, 2013).

Content of the Provo corpus

Data collection for the Provo Corpus proceeded in two stages. In the first stage, the predictability norms were created; cloze scores were collected via a large-scale online survey for each word in 55 paragraphs taken from various sources. In the second stage, each of these 55 paragraphs was presented to a different set of participants to read while their eyes were tracked, creating a large corpus of eye movement data. Both sets of data (predictability norms and eye-tracking data) are available as part of the Provo Corpus. In the section that follows, we describe the predictability norms in more detail. Then, in the next section, we provide details about the eye-tracking corpus.

Predictability norms

Participants Four hundred seventy-eight participants from Brigham Young University completed an online survey for course credit through the Psychology Department subject pool. The responses from eight participants were discarded because they were not native speakers of English or did not complete the survey. In total, data from 470 people (267 females, 203 males) were included. Participants' ages ranged from 18 to 50 years (*M*: 21). All were high school graduates with at least some college experience, and approximately 10% had received some degree beyond a high school diploma.

Materials Fifty-five short passages were taken from a variety of sources, including online news articles, popular science magazines, and public-domain works of fiction. These passages were an average of 50 words long (range: 39–62) and contained 2.5 sentences on average (range: 1–5). The sentences were on average 13.3 words long (range: 3–52). Across all texts, there were 2,689 words total, including 1,197 unique word forms.

The words were tagged for parts of speech using the Constituent Likelihood Automatic Word-Tagging System (CLAWS; Garside & Smith, 1997). Using the tags provided by CLAWS, words were then divided into nine separate classes. In total, the passages contained 227 adjectives, 169 adverbs, 196 conjunctions, 364 determiners, 682 nouns, 287 prepositions, 109 pronouns, 502 verbs, and 153 other words and symbols. In addition, inflectional information was also coded for the words within each class, where appropriate. Nouns were coded for number, and verbs were coded for tense.

Words ranged from 1 to 15 letters long (*M*: 4.76). A measure of the semantic association between the target word and the entire preceding passage context was obtained using latent semantic analysis (LSA; Landauer & Dumais, 1997). This LSA context score was obtained using the General Reading–Up to First Year of College topic space with 300 factors. LSA cosines typically range from 0 to 1, with larger values indicating greater meaning overlap between two terms. LSA context scores ranged from .03 to .97 ($M = .53$). This variable quantifies the semantic fit of the target word with the preceding context. Target word positions within the passage (sentence number) and within the sentence (word-in-sentence number) were also obtained.

Procedure Participants completed an online survey administered through the Qualtrics Research Suite software

(Qualtrics, Provo, UT). Participants first answered a few demographic questions (gender, age, education level, and language history), then proceeded to complete the main body of the survey. For each question, participants were instructed to "Please type the word that you think will come next." Beneath this instruction was a portion of one of the texts, with a response box for the participant to type in a word. For the first question about a text, only the first word in the text was visible, then the first two words for the second question, the first three words for the third, and so on, until for the last question about a text all words but the final word in the text were visible. Thus, participants provided responses for all but the first word in each text. Participants were required to give a response before proceeding to the next question, and within a text, all questions were presented in a fixed order, so that participants were never given a preview of the upcoming words in a text.

Each participant was randomly assigned to complete five texts, giving responses for an average of 227 different words. For each word in each text, an average of 40 participants provided a response (range: 19–43).

Content of predictability norms file Responses were edited for spelling. When a response contained contractions or multiple words, the first word was coded. Each survey response was then tagged for its part of speech using CLAWS, and the responses were divided into word classes and coded for inflection, as we described previously for the target words. Responses and targets (the word that actually appeared in that position in the text) were compared to see whether they matched in three different ways: orthographically (cloze score), by word class, and (for nouns and verbs) by inflection. Responses and the target were considered to match orthographically if the two full word forms were orthographically identical. For the purposes of this comparison, all letters were in lowercase. A word class match was coded if the response and target belonged to the same word class, and an inflectional match was coded if the words belonged to the same word class and carried the same inflectional suffix. LSA (Landauer & Dumais, 1997) was also used to provide an estimate of the relatedness of the responses and targets for all content word targets. The LSA cosine between each response and target was obtained using the General Reading topic space via the Web-based LSA interface (http://lsa.colorado.edu). Note that this procedure, which compared the response and target words, is different from the LSA procedure described previously, in which the target words were compared to the entire preceding passage. Comparing two words provides an estimate of the semantic relatedness of these two words, while comparing the target word with its context estimates the contextual fit of the target word. Thus, the corpus provides measures of the contextual fit of the target word and of its semantic predictability. Most of these variables can be found in the eye-tracking corpus file, described in the next section. Table 1 lists and defines the variables in the Provo predictability norms.

Eye-tracking data

Participants Eighty-four participants from Brigham Young University completed the eye-tracking portion of the study. All participants were native speakers of American English with 20/20 corrected or uncorrected vision. They received course credit through the Psychology Department subject pool. None had participated in the predictability norming survey.

Apparatus For the eye-tracking portion of the study, eye movements were recorded via an SR Research EyeLink 1000 Plus eye-tracker (spatial resolution of 0.01°) sampling at 1000 Hz. Participants were seated 60 cm away from a monitor with a display resolution of 1,600 × 900, so that approximately three characters subtended 1° of visual angle (the monitor was 40 × 24 deg of visual angle). Head movements were minimized with a chin and forehead rest. Although viewing was binocular, eye movements were recorded from the right eye. The experiment was controlled with the SR Research Experiment Builder software.

Table 1 Predictability norm variables, with descriptions

Variable	Description
Word_Unique_ID	A unique ID number for each word (each token) in the data set
Text_ID	The text number (paragraph 1–55)
Text	The entire text from which the target word is taken
Word_Number	The ordinal position of the word in the text
Sentence_Number	The ordinal number of the sentence in which the current word is located
Word_In_Sentence_Number	The ordinal position of the current word within the current sentence
Word	The target word, with punctuation, capitalization and contractions removed
Response	The response produced by the participant in the cloze procedure
Response_Count	Number of participants who produced a given response
Total_Response_Count	The total number of responses provided on the cloze task for this word token
Response_Proportion	How often a given response was provided, as a proportion of all responses. Can be used as a measure of predictability. Response_Proportion = Response_Count/Total_Response_Count

Table 2 Eye-tracking corpus variables, with descriptions

Variable	Description
Participant_ID	A unique ID number for each participant
Word_Unique_ID	A unique ID number for each word (each token) in the data set
Text_ID	The text number (paragraph 1–55)
Word_Number	The ordinal position of the word in the text
Sentence_Number	The ordinal number of the sentence in which the current word is located
Word_In_Sentence_Number	The ordinal position of the current word within the current sentence
Word	The word as it appeared on the screen
Word_Cleaned	The word, with punctuation and capitalization removed
Word_Length	The length of the current word, in letters
Total_Response_Count	The total number of responses provided on the cloze task for this word token
Unique_Count	The total number of unique responses provided on the cloze task for this word token
OrthographicMatch	Cloze probability: The proportion of responses that were an orthographic match with the target word (e.g., Target = "apple" and response = "apple")
OrthoMatchModel	The same as OrthographicMatch, but instead of raw proportions this variable is the fitted values generated by a logit mixed-effects model that included only random by-word intercepts. These values correlate with OrthographicMatch ($r = .999$) but never include 0 or 1. This variable is provided for users who desire to log transform the predictability measures.
IsModalResponse	Whether the target word was the most commonly produced response (1) or not (0)
ModalResponse	The modal response. If IsModalResponse is 1, this is the same as Word (see above). If IsModalResponse is 0, this is whichever response was provided most frequently.
ModalResponseCount	A count of how many times the modal response was provided in the cloze procedure
Certainty	The cloze probability of the modal response. Certainty = ModalResponseCount/ResponseCount
POS_CLAWS	The part of speech tag of the target word (See http://ucrel.lancs.ac.uk/claws/ for more information on the meaning of the specific tags.)
Word_Content_Or_Function	Whether the word is a content word or a function word, based on POS_CLAWS
Word_POS	A more general grouping of parts of speech, based on POS_CLAWS, which includes the following categories: adjective, adverb, article, conjunction, determiner, existential, infinitive marker, negative, noun, number, preposition, pronoun, verb
POSMatch	The proportion of responses with the same POS as the target, using Word_POS (e.g., Target and response are both nouns).
POSMatchModel	The same as POSMatch, but instead of raw proportions this variable is the fitted values generated by a logit mixed-effects model that included only random by-word intercepts. These values correlate with POSMatch ($r = .999$) but never include 0 or 1. This variable is provided for users who desire to log transform the predictability measures.
InflectionMatch	The proportion of responses that carried the same inflection (number for nouns, tense for verbs) as the target (for nouns and verbs only)—e.g., target and response are both past-tense verbs.
InflectionMatchModel	The same as InflectionMatch, but instead of raw proportions this variable is the fitted values generated by a logit mixed-effects model that included only random by-word intercepts. These values correlate with InflectionMatch ($r = .999$) but never include 0 or 1. This variable is provided for users who desire to log transform the predictability measures.
LSA_Context_Score	A measure of the semantic association between the target word and the entire preceding passage context, obtained using latent semantic analysis (Landauer & Dumais, 1997; http://lsa.colorado.edu/). For example, the LSA score for the word "rumblings" was obtained by comparing "rumblings" to the preceding context "There are now." This score is a measure of the contextual fit of a given target word.
LSA_Response_Match_Score	The mean LSA match score between the target and all provided responses. For example, pairwise LSA was used to compare the target "carts" with the responses provided in the cloze procedure (e.g., "horses," "slower," and "the"), and the LSA scores for all responses were averaged. This measure is an estimate of the semantic predictability of a given target word (i.e., could participants have a good sense for the general meaning of the upcoming word, even if they cannot predict exactly what that word will be).
IA_ID	Identification number for each interest area in the text. Note that because of typos and text parsing errors, this number may not correspond with Word_Number.
IA_LABEL	The string of letters (w/ punctuation) contained within the interest area

Table 2 (continued)

Variable	Description
TRIAL_INDEX	The order that the text was presented within the experiment for a given participant
IA_LEFT	The left boundary of the interest area, in pixels from the left of the screen
IA_RIGHT	The right boundary of the interest area, in pixels from the left of the screen
IA_TOP	The top boundary of the interest area, in pixels from the top of the screen
IA_BOTTOM	The bottom boundary of the interest area, in pixels from the top of the screen
IA_AREA	The total screen area of the interest area, in pixels
IA_FIRST_FIXATION_DURATION	First Fixation Duration: The duration of the first fixation on the interest area, in milliseconds.
IA_FIRST_FIXATION_INDEX	Ordinal sequence of the first fixation that was within the current interest area
IA_FIRST_FIXATION_VISITED_IA_COUNT	The number of interest areas visited prior to first fixation on the current interest area
IA_FIRST_FIXATION_X	The X position of the first fixation event that was within the current interest area, in pixels
IA_FIRST_FIXATION_Y	The Y position of the first fixation event that was within the current interest area, in pixels
IA_FIRST_FIX_PROGRESSIVE	Checks whether later interest areas have been visited before the first fixation enters the current interest area. 1 if NO higher IA ID in earlier fixations before the first fixation in the current interest area; 0 otherwise. This measure is useful in reading to check whether the first run of fixations in this interest area is in fact first-pass fixations.
IA_FIRST_FIXATION_RUN_INDEX	This counts how many runs of fixations have occurred when a first fixation is made to an interest area. The current run is also included in the tally.
IA_FIRST_FIXATION_TIME	Start time of the first fixation to enter the current interest area
IA_FIRST_RUN_DWELL_TIME	Gaze duration: Dwell time (i.e., summation of the duration across all fixations) of the first run within the current interest area
IA_FIRST_RUN_FIXATION_COUNT	Number of all fixations in a trial falling in the first run of the current interest area
IA_FIRST_RUN_START_TIME	Start time of the first run of fixations in the current interest area
IA_FIRST_RUN_END_TIME	End time of the first run of fixations in the current interest area
IA_FIRST_RUN_FIXATION_%	Percentage of all fixations in a trial falling in the first run of the current interest area
IA_DWELL_TIME	Total Reading Time: Dwell time (i.e., summation of the duration across all fixations) on the current interest area
IA_FIXATION_COUNT	Total fixations falling in the interest area
IA_RUN_COUNT	Number of times the Interest Area was entered and left (runs)
IA_SKIP	An interest area is considered skipped (i.e.,IA_SKIP = 1) if no fixation occurred in first-pass reading.
IA_REGRESSION_IN	Whether the current interest area received at least one regression from later interest areas (e.g., later parts of the sentence). 1 if interest area was entered from a higher IA_ID (from the right in English); 0 if not.
IA_REGRESSION_IN_COUNT	Number of times interest area was entered from a higher IA_ID (from the right in English)
IA_REGRESSION_OUT	Whether regression(s) was made from the current interest area to earlier interest areas (e.g., previous parts of the sentence) prior to leaving that interest area in a forward direction. 1 if a saccade exits the current interest area to a lower IA_ID (to the left in English) before a later interest area was fixated; 0 if not.
IA_REGRESSION_OUT_COUNT	Number of times interest area was exited to a lower IA_ID (to the left in English) before a higher IA_ID was fixated in the trial
IA_REGRESSION_OUT_FULL	Whether regression(s) was made from the current interest area to earlier interest areas (e.g., previous parts of the sentence). 1 if a saccade exits the current interest area to a lower IA_ID (to the left in English); 0 if not. Note that IA_REGRESSION_OUT only considers first-pass regressions whereas IA_REGRESSION_OUT_FULL considers all regressions, regardless whether later interest areas have been visited or not.
IA_REGRESSION_OUT_FULL_COUNT	Number of times interest area was exited to a lower IA_ID (to the left in English)
IA_REGRESSION_PATH_DURATION	Go-Past Time: The summed fixation duration from when the current interest area is first fixated until the eyes enter an interest area with a higher IA_ID
IA_FIRST_SACCADE_AMPLITUDE	Amplitude (in degree of visual angle) of the first saccade entering into the current interest area NOTE: Saccade data have not been cleaned, and so include return sweeps (large eye movements from the end of one line to the beginning of the next). Excluding saccades >15 deg removes these return sweeps without impacting other reading-related saccades.

Table 2 (continued)

Variable	Description
IA_FIRST_SACCADE_ANGLE	Angle between the horizontal plane and the direction of the first saccade entering into the current interest area
IA_FIRST_SACCADE_START_TIME	Start time of the saccade that first landed within the current interest area
IA_FIRST_SACCADE_END_TIME	End time of the saccade that first landed within the current interest area

Procedure Participants were told that they would be reading short texts on a computer screen while their eye movements were recorded. These texts were the same 55 texts that had been used in the survey. Each trial involved the following sequence. The trial began with a gaze trigger, a black circle presented in the position of the first character in the text. Once a stable fixation was detected on the gaze trigger, the text was presented. The participant read the text and pressed a button when finished. Then a new gaze trigger appeared, and the next trial began. The texts were presented in a random order for each participant. Participants had no task other than to read for comprehension.

Content of eye-tracking data file Prior to the analysis of eye-tracking data, the data were cleaned, with fixations shorter than 80 ms and longer than 800 ms being removed (about 4% of the data). We note that this cleaning procedure does not guarantee that all of the measures will be outlier-free. Any saccade-based measure and any measure comprising the sum of several fixations (e.g., gaze duration or total reading time) would still contain outliers. We have left these in so that users may apply their own preferred cleaning criteria. There are also some missing data values in the file. These cells are denoted with "NA." Different reading measures were computed for predefined interest areas around each word in each passage, comprising the letters of each word and half of the white space surrounding each word, both vertically and horizontally.

In Table 2, the columns that appear in the Provo Eye-Tracking Corpus are listed and described. First, participant and word identification variables are listed. Then come variables associated with traditional measures of predictability (cloze scores). Next are variables associated with morpho-syntactic predictability (the predictability of a word class and inflection). Variables associated with semantic relationships and predictability appear next. Finally, eye-tracking variables conclude the list. These variables are the output of the SR Research Data Viewer (SR Research Ltd., version 1.11.1), and the descriptions for these variables come from or are modifications of the descriptions found in the Data Viewer User's Manual. Means and standard deviations for these variables can be found in Luke and Christianson (2016), Table 6. Various analyses using this data are also described in Luke and Christianson (2016).

Availability

The Provo Corpus can be downloaded from the Open Science Framework at http://osf.io/sjefs. It consists of two files, which can be downloaded separately. The file Provo_Corpus-Predictability_Norms.csv is a comma-separated text file that contains the predictability norms, in the format described above. This file is for users who want to create predictability stimuli or to explore how different factors (e.g., word frequency, contextual constraint) influence the cloze task responses (see, e.g., Luke & Christianson, 2016; Staub, Grant, Astheimer, & Cohen, 2015). Users interested in the eye-tracking corpus should download the file Provo_Corpus-Eyetracking_Data.csv, another comma-separated text file, which contains the eye-tracking data. This file also contains summary predictability values (see Table 2), so that users only interested in the eye-tracking data do not need to download both files.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73,* 247–264. doi:10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57,* 502–518.

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology, 58A,* 1065–1086. doi:10.1080/02724980443000476

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17,* 364–390. doi:10.1016/0010-0285(85)90013-1

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

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*. doi:10.1017/S0140525X1500031X

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods, 49,* 602–615. doi:10.3758/s13428-016-0734-0

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B, 369,* 20120394. doi:10.1098/rstb.2012.0394

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass, 8,* 631–645.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20,* 641–655. doi:10.1016/S0022-5371(81)90220-6

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112,* 777–813. doi:10.1037/0033-295X.112.4.777

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. N. Leech, & T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102–121). London, UK: Longman.

Hamberger, M. J., Friedman, D., & Rosen, J. (1996). Completion norms collected from younger and older adults for 198 sentence contexts. *Behavior Research Methods, Instruments, & Computers, 28,* 102–108.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research, 1626,* 118–135.

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience, 31,* 19–31.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133–156. doi:10.1016/S0749-596X(03)00023-8

Kennedy, A., Hill, R., & Pynte, J. (2003). *The Dundee Corpus.* Paper presented at the 12th European Conference on Eye Movement, Dundee, Scotland.

Kennedy, A., Pynte, J., Murray, W. S., & Paul, S.-A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *Quarterly Journal of Experimental Psychology, 66,* 601–618.

Kliegl, R., & Engbert, R. (2005). Fixation durations before word skipping in reading. *Psychonomic Bulletin & Review, 12,* 132–138.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16,* 262–284. doi:10.1080/09541440340000213

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General, 135,* 12–35. doi:10.1037/0096-3445.135.1.12

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31,* 32–59.

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). New York: Oxford University Press.

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. doi:10.1037/0033-295X.104.2.211

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88,* 22–60.

McDonald, S. A., & Tamariz, M. (2002). Completion norms for 112 Spanish sentences. *Behavior Research Methods, Instruments, & Computers, 34,* 128–137.

Nuthmann, A., Engbert, R., & Kliegl, R. (2007). The IOVP effect in mindless reading: Experiment and modeling. *Vision Research, 47,* 990–1002. doi:10.1016/j.visres.2006.11.005

Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology, 52,* 1456–1469.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11,* 105–110. doi:10.1016/j.tics.2006.12.002

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36,* 329–347. doi:10.1017/S0140525X12001495

Pynte, J., New, B., & Kennedy, A. (2009). On-line contextual influences during reading normal text: The role of nouns, verbs and adjectives. *Vision Research, 49,* 544–552.

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 514–528. doi:10.1037/a0020990

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3,* 504–509. doi:10.3758/BF03214555

Schwanenflugel, P. J. (1986). Completion norms for final words of sentences using a multiple production measure. *Behavior Research Methods, Instruments, & Computers, 18,* 363–371. doi:10.3758/BF03204419

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128,* 302–319. doi:10.1016/j.cognition.2013.02.013

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass, 9,* 311–327.

Staub, A., Abbott, M., & Bogartz, R. S. (2012). Linguistically guided anticipatory eye movements in scene viewing. *Visual Cognition, 20,* 922–946.

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language, 82,* 1–17.

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

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83,* 176–190.