UCLA
UCLA Previously Published Works

Title

The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions.

Permalink
https://escholarship.org/uc/item/8q75w6fc

Journal
Language Resources and Evaluation, 55(1)

ISSN
1574-020X

Authors
Futrell, Richard
Gibson, Edward
Tily, Harry
et al.

Publication Date
2021

DOI
10.1007/s10579-020-09503-7

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Lang Resources & Evaluation (2021) 55:63–77
https://doi.org/10.1007/s10579-020-09503-7

ORIGINAL PAPER

The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions

Richard Futrell[1] · Edward Gibson[2] · Harry J. Tily[3] · Idan Blank[4] · Anastasia Vishnevetsky[2] · Steven T. Piantadosi[5] · Evelina Fedorenko[2]

Published online: 4 September 2020
© The Author(s) 2020

Abstract It is now a common practice to compare models of human language processing by comparing how well they predict behavioral and neural measures of processing difficulty, such as reading times, on corpora of rich naturalistic linguistic materials. However, many of these corpora, which are based on naturally-occurring text, do not contain many of the low-frequency syntactic constructions that are often required to distinguish between processing theories. Here we describe a new corpus consisting of English texts edited to contain many low-frequency syntactic constructions while still sounding fluent to native speakers. The corpus is annotated with hand-corrected Penn Treebank-style parse trees and includes self-paced reading time data and aligned audio recordings. We give an overview of the content of the corpus, review recent work using the corpus, and release the data.

Keywords Cognitive modeling · Reading time · Psycholinguistics

1 Introduction

It is becoming a standard practice to evaluate theories of human language processing by comparing their ability to predict behavioral and neural reactions to fixed standardized corpora of naturalistic text. This method has been used to study

✉ Richard Futrell
rfutrell@uci.edu

[1] University of California, Irvine, USA

[2] Massachusetts Institute of Technology, Cambridge, USA

[3] Viome, Inc., Seattle, USA

[4] University of California, Los Angeles, USA

[5] University of California, Berkeley, USA

several dependent variables which are believed to be indicative of human language processing difficulty, including word fixation time in eyetracking (Kennedy et al. 2013), word reaction time in self-paced reading (Roark et al. 2009; Frank et al. 2013), BOLD signal in fMRI data (Bachrach et al. 2009), and event-related potentials (Dambacher et al. 2006; Frank et al. 2015).

The more traditional approach to evaluating psycholinguistic models has been to collect psychometric measures on hand-crafted experimental stimuli designed to tease apart detailed model predictions. While this approach makes it easy to compare models on their accuracy for specific constructions and phenomena, it is hard to get a sense of how models compare on their coverage of a broad range of phenomena. Comparing model predictions over standardized texts makes it is easier to evaluate coverage.

Although the corpus approach has these advantages, the existing corpora currently used are based on naturally-occurring text, which is unlikely to include the kinds of sentences which can crucially distinguish between theories. Many of the most puzzling phenomena in psycholinguistics, and the phenomena which have been used to test models, have only been observed in extremely rare constructions, such as multiply nested object-extracted relative clauses (Roland et al. 2007). Corpora of naturally-occurring text are unlikely to contain these constructions.

Here we attempt to combine the strength of experimental approaches—which can test theories using targeted low-frequency structures—and corpus studies—which provide broad-coverage comparability between models. We introduce and release a new corpus, the **Natural Stories Corpus**, a series of English narrative texts designed to contain many low-frequency and psycholinguistically interesting syntactic constructions while still sounding fluent and coherent. The texts are annotated with hand-corrected Penn Treebank style phrase structure parses, and Universal Dependencies parses automatically generated from the phrase structure parses. We also release self-paced reading time data for all texts, and word-aligned audio recordings of the texts. We hope the corpus can form the basis for further annotation and become a standard test set for psycholinguistic models.[1]

2 Related work

Here we survey datasets which are commonly used to test psycholinguistic theories, and how they relate to the current release.

The most prominent psycholinguistic corpus for English is the **Dundee Corpus** (Kennedy 2003), which contains 51,501 word tokens in 2368 sentences from British newspaper editorials, along with eyetracking data from 10 experimental participants. The full corpus is not publically available. A dependency parse of the corpus was released by Barrett et al. (2015). Like in the current work, the eyetracking data in the Dundee corpus is collected for sentences in context and so reflects influences

[1] The corpus is available from http://github.com/languageMIT/naturalstories. This corpus is distributed under an Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license, allowing free modification and re-distribution of the corpus so long as derivative work is released under the same terms.

beyond the sentence level. The corpus has seen wide use (Demberg and Keller 2008; Mitchell et al. 2010; Frank and Bod 2011; Fossum and Levy 2012; Smith and Levy 2013; van Schijndel and Schuler 2015; Luong et al. 2015).

The **Potsdam Sentence Corpus** (Kliegl et al. 2006) of German provides 1138 words in 144 sentences, with cloze probabilities and eyetracking data for each word. Like the current corpus, the Potsdam Sentence Corpus was designed to contain varied syntactic structures, rather than being gathered from naturalistic text. The corpus consists of isolated sentences which do not form a narrative, and during eyetracking data collection the sentences were presented in a random order. The corpus has been used to evaluate models of sentence processing based on dependency parsing (Boston et al. 2011, 2018) and to study effects of predictability on event-related potentials (Dambacher et al. 2006).

The **MIT Corpus** introduced in Bachrach et al. (2009) has similar aims to the current work, collecting reading time and fMRI data over sentences designed to contain varied structures. This dataset consists of four narratives with a total of 2647 tokens; it has been used to evaluate models of incremental prediction in Roark et al. (2009), Wu et al. (2010), and Luong et al. (2015).

The **UCL Corpus** (Frank et al. 2013) consists of 361 English sentences drawn from amateur novels, chosen for their ability to be understood out of context, with self-paced reading and eyetracking data. The goal of the corpus is to provide a sample of typical narrative sentences, complementary to our goal of providing a corpus with low-frequency constructions. Unlike the current corpus, the UCL Corpus consists of isolated sentences, so the psychometric data do not reflect effects beyond the sentence level.

Eyetracking corpora for other languages are also available, including the **Postdam-Allahabad Hindi Eyetracking Corpus** (Husain et al. 2015) and the **Beijing Sentence Corpus of Mandarin Chinese** (Yan et al. 2010).

3 Corpus description

3.1 Text

The Natural Stories corpus consists of 10 stories of about 1000 words each, comprising a total of 10,245 lexical word tokens in 485 sentences. The stories were developed by taking existing publicly available texts and editing them to contain many rare or marked syntactic constructions, while still retaining the same meaning, and while maintaining a high degree of overall fluency and comprehensibility as judged subjectively by the editor.[2] The editors focused on including the following marked syntactic constructions: subject- and object-extracted relative clauses, clefts, topicalized structures, extraposed relative clauses, sentential subjects, sentential complements, local structural ambiguity (especially NP/Z ambiguity), idioms, and conjoined clauses with a variety of coherence relations. More details on these

[2] The editors were Anastasia Vishnevetsky, Edward Gibson, Evelina Fedorenko, and Steven T. Piantadosi.

Table 1 Stories with titles and sources

Story	Title	Source title	Source author
1	Boar	The Legend of the Bradford Boar[a]	E. H. Hopkinson
2	Aqua	Aqua, or the Water Baby[b]	Kate Douglas Wiggin
3	Matchstick	The Little Match-Seller[c]	Hans Christian Andersen
4	King of Birds	The King of the Birds[d]	Brothers Grimm
5	Elvis	Elvis Died at the Florida Barber College[e]	Roger Dean Kiser
6	Mr. Sticky	Mr. Sticky[f]	Mo McAuley
7	High School	Bullies	Sarah Cleaves
8	Roswell	Roswell UFO incident[g]	Wikipedia
9	Tulips	Tulip mania[h]	Wikipedia
10	Tourette's	Tourette Syndrome Fact Sheet[i]	NINDS

[a] http://www.make4fun.com/stories/British-short-story/3917-The-Legend-of-the-Bradford-Boar-by-E-H-Hopkinsona

[b] http://fullreads.com/literature/aqua-or-the-water-baby/

[c] http://stenzel.ucdavis.edu/180/anthology/matchgirl.html

[d] http://www.apples4theteacher.com/holidays/bird-day/short-stories/the-king-of-the-birds.html

[e] http://www.eastoftheweb.com/short-stories/UBooks/ElvDie.shtml

[f] http://www.eastoftheweb.com/short-stories/UBooks/MrStic.shtml

[g] http://en.wikipedia.org/w/index.php?title=Roswell_UFO_incident&oldid=331989741

[h] http://en.wikipedia.org/w/index.php?title=Tulip_mania&oldid=329157998

[i] http://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Tourette-Syndrome-Fact-Sheet

Fig. 1 Histograms of sentence length (in tokens, including punctuation) in Natural Stories and the Dundee corpus

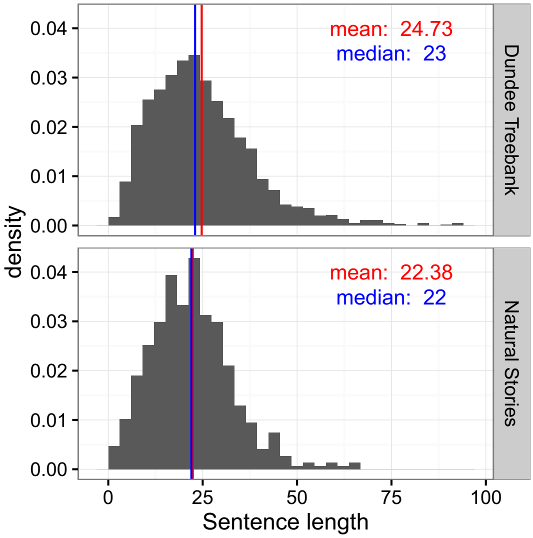
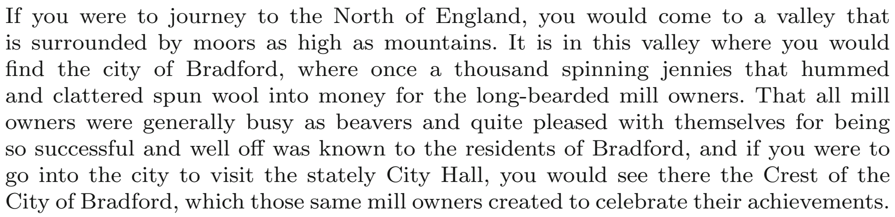


Table 2 Summary of stories by length

Story	# Words	# Sentences
1	1073	57
2	990	37
3	1040	55
4	1085	55
5	1013	45
6	1089	64
7	999	48
8	980	33
9	1038	48
10	938	43

Here, 'words' refers to lexical words



If you were to journey to the North of England, you would come to a valley that is surrounded by moors as high as mountains. It is in this valley where you would find the city of Bradford, where once a thousand spinning jennies that hummed and clattered spun wool into money for the long-bearded mill owners. That all mill owners were generally busy as beavers and quite pleased with themselves for being so successful and well off was known to the residents of Bradford, and if you were to go into the city to visit the stately City Hall, you would see there the Crest of the City of Bradford, which those same mill owners created to celebrate their achievements.

Fig. 2 Sample text from the first story. The sample text contains marked syntactic structures: (1) an it-cleft ("It is in this valley where ..."), (2) a very long sentential subject ("That all mill owners ...was well known"), (3) an object-extracted relative clause ("..., which those same mill owners created to celebrate ..."), and others

constructions are provided in Appendix. The texts and their sources are listed in Table 1. Along with the release of the texts and reading time data, we also release a document showing which marked syntactic constructions are present in which sentences.

The mean number of tokens per sentence is 22.38, around the same as the Dundee corpus (24.73). Figure 1 shows a histogram of sentence length in Natural Stories as compared to Dundee. The word and sentence counts for each story are given in Table 2. Each token has a unique code which is referenced throughout the various annotations of the corpus, defined in the file `words.tsv`.

In Fig. 2 we give a sample of text from the corpus (from the first story).

3.2 Parses

The texts were parsed automatically using the Stanford Parser (Klein and manning 2003) and hand-corrected. Trace annotations were added by hand. We provide the resulting Penn Treebank-style phrase structure parse trees. We also provide

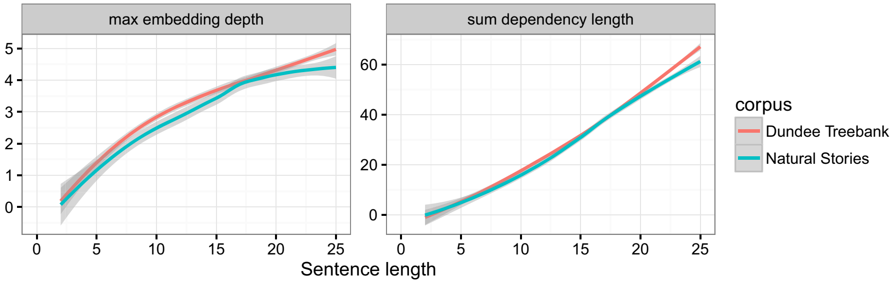


Fig. 3 Sum dependency length and maximum embedding depth per sentence in Natural Stories (with automatic conversion to Universal Dependencies) and the Dundee Treebank (Barrett et al. 2015)

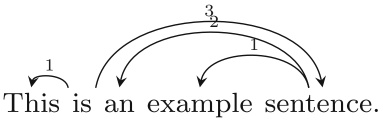


Fig. 4 Example of sum dependency length and maximum embedding depth. Arcs are drawn from syntactic heads to syntactic dependents. Numbers over dependency arcs indicate dependency length. The sum dependency length for this sentence is 1 + 3 + 2 + 1 = 7. The maximum embedding depth = 3, because there are three arcs over the word "example"

Universal Dependencies-style parses (Nivre 2015) automatically converted from the corrected parse trees using the Stanford Parser. Deep syntactic annotations following a categorial grammar are provided by Shain et al. (2018b).

Figure 3 shows some basic syntactic features of the corpus as compared with the Dundee corpus, computed from the dependency parses of the text. In this figure, **sum dependency length** is the total length of all dependency arcs in the dependency parse of the sentence, where length is calculated as the number of intervening words between the head and the dependent plus one, as is standard in corpus studies of dependency length (Liu 2008; Futrell et al. 2015). **Maximum embedding depth** refers to the maximum depth of a stack that would be required to parse a sentence using an incremental stack-based dependency parser such as Nivre and Scholzm (2004); it is equal to the maximum number of dependency arcs over a word at any point in a sentence. For example, the sum dependency length and maximum embedding depth of a sample sentence is calculated in Fig. 4. Although the corpus contains many low-frequency and marked constructions, its dependency length and embedding depth are not greater than the newspaper text in the Dundee corpus.

3.3 Self-paced reading data

We collected self-paced reading (SPR) data (Just et al. 1982) for the stories from 181 native English speakers over Amazon Mechanical Turk. Text was presented in a dashed moving window display, with masked spaces. For each word, we recorded its reading time (RT) as the amount of time taken by the reader to press the button to advance to the next word. Line breaks were determined by fitting the texts to a random width of the screen, so that line breaks do not occur for the same word across participants. Each story was accompanied by 6 comprehension questions, where participants chose the correct answer from a set of two. These comprehension questions are included in our data release.

Each participant read 5 stories per HIT. Participants were paid \$2.[3] 19 participants read all 10 stories, and 3 participants stopped after one story. Figure 5 shows histograms of RTs per story. For this analysis, we discarded SPR data from a participant's pass through a story if the participant got less than 5 questions correct, resulting in the exclusion of 89 passes (9% of passes excluded). We also excluded RTs less than 100 ms or greater than 3000 ms. Data exclusions of this kind and magnitude are common in psycholinguistics (see for example Boyce et al. (2020)).

3.3.1 Inter-subject correlations

In order to evaluate the reliability of the self-paced reading RTs and their robustness across experimental participants, we analyzed inter-subject correlations (ISCs). For each subject, we calculated the Spearman correlation of that subject's RTs on a story with average RTs from all other subjects on that story. In this way, for each story, we get one ISC statistic per subject. Figure 6 shows histograms of these statistics per story. High correlations indicate high inter-subject reliability.

3.3.2 Psycholinguistic validation

In order to check the integrity of our RT data, we verified that it shows some of the basic effects that have been documented in the field of psycholinguistics. Some of the most robust predictors of reading time in the psycholinguistic literature are frequency, word length, and surprisal (Kliegl et al. 2004; Smith and Levy 2013). More frequent words are read more quickly, longer words are read more slowly, and more surprising words (as determined using e.g. an *n*-gram model) are read more slowly. Here we check whether these well-known effects can be found in our SPR corpus.

To do this, we fit a regression models to predict reading time based on each of the three predictors individually. For example, in order to predict the reading time from log frequency, we fit a regression of the form:

$$\hat{y_i} = \alpha + \beta \log f_i + \epsilon_i, \tag{1}$$

where y_i is the reading time (in ms) for the ith word in the corpus, f_i is the frequency of the ith word in the corpus, and the scalars α and β are chosen to minimize the sum of the squared errors ϵ_i for each word. For this analysis, we excluded outlier data

[3] The SPR experiment was carried out in early 2011, when such wages were common on Mechanical Turk.

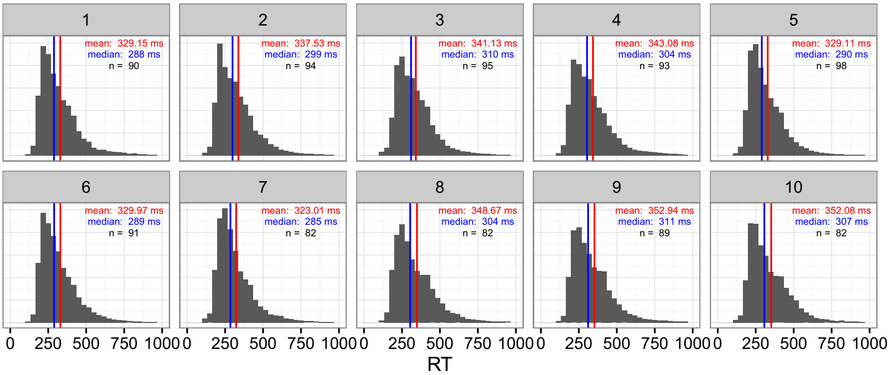


Fig. 5 Histograms of SPR reading times (RTs) per story, after data exclusion

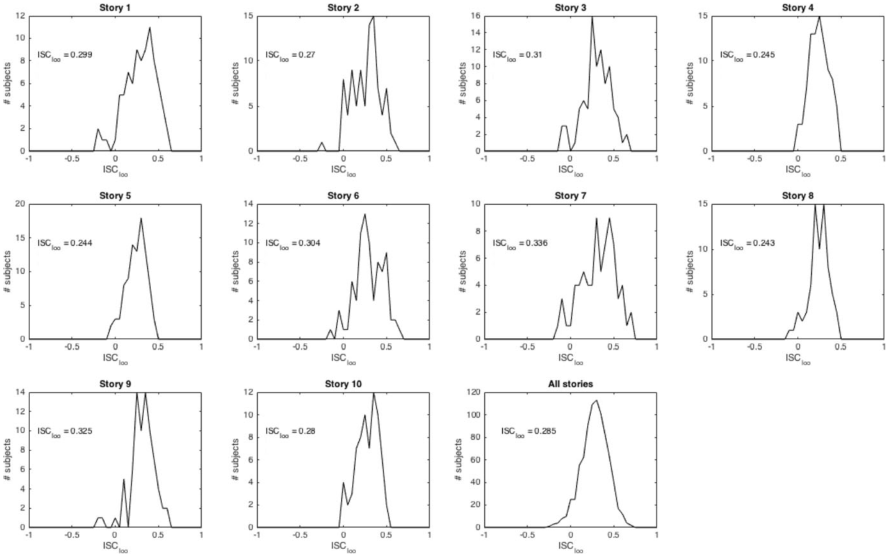


Fig. 6 Leave-one-out inter-subject correlations (ISCs) of RTs per story. In the panels, ISC_{loo} gives the average leave-one-out ISC for that story

and participants with low comprehension scores following the same criteria as in Sect. 3.3.

We fit a regression model as in Eq. 1 to predict reading time from log frequency, another model to predict reading time from word length (measured in orthographic characters), and another model to predict reading time from log probability under a trigram model. We expect to find a negative value of the coefficient β when predicting reading time from frequency and trigram probability, and a positive value of the coefficient β when predicting reading time from word length. Word and trigram counts are collected from the Google Books *n*-grams corpus, summing over

Table 3 Regression coefficients from individual mixed-effects regressions predicting RT for each of the three predictors log frequency, log trigram probability, and word length

Predictor	β	Std. error	t value
Log frequency	− 2.61	0.08	− 32.27
Log trigram probability	− 2.19	0.09	− 23.90
Word length	4.21	0.12	35.72

The first column is the predictor used in a regression predicting reading time; the column β is the fitted regression coefficient for the predictor; Std. error is the standard error on the estimate of β, and the t value is the t-statistic for the value of β as compared to 0. We predict and find negative values of β for log frequency and log probability and a positive effect of word length. All p values are <0.001

years from 1990 to 2013; we make these counts available along with the corpus. Each regression is a mixed-effects regression with subject and story as random intercepts (models with random slopes did not converge), in addition to the predictors in Eq. 1. By including random intercepts, we control for by-subject and by-story variability.

The results of the regressions are shown in Table 3. In keeping with well-known effects, increased frequency and trigram probability both lead to faster reading times, and word length leads to slower reading times. These results show that basic psycholinguistic effects are present in our SPR data.

3.4 Aligned audio

We also release audio recordings of the text. These recordings are meant to be used as auditory stimuli in settings such as fMRI experiments. Five stories were read by a male (stories 1, 2, 5, 8, and 9), and the other five by a female (stories 3, 4, 6, 7, and 10).

Along with the raw audio recordings, we release time-alignments by word. The alignments were created by initial forced alignment with a proprietary text-to-audio aligner developed by Lincoln Labs, and then hand-checked and corrected by a research assistant.

3.5 Syntactic constructions

Here we give an overview of the low-frequency or marked syntactic constructions which occur in the stories. We coded sentences in the Natural Stories corpus for presence of a number of marked constructions, and also coded 200 randomly selected sentences from the Dundee corpus for the same features. The features coded are listed by name and explained in Appendix.

Figure 7 shows the rates of occurrence for these marked constructions per sentence in the two corpora. From the figure, we see that the Natural Stories corpus has especially high rates of relative clauses, idioms, adjective conjunction, local NP/S ambiguities, and clefts. Although there are some marked constructions which have higher frequency in Dundee than in Natural Stories, most (27/37) of the

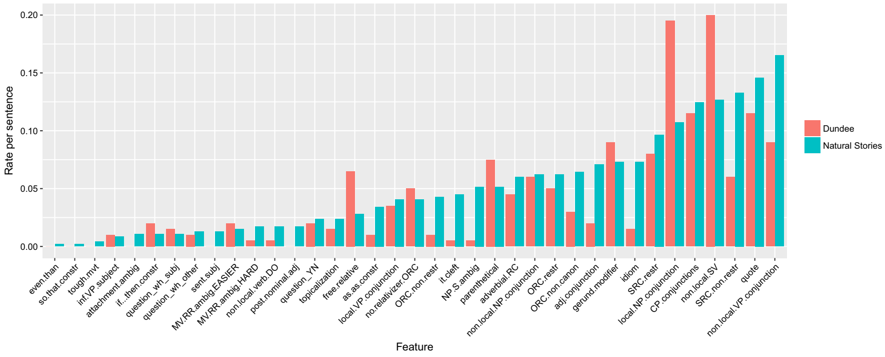


Fig. 7 Rates of marked constructions in the Natural Stories corpus and in 200 randomly sampled sentences from the Dundee corpus

constructions are more frequent in Natural Stories than in Dundeee. Furthermore, the constructions which are especially frequent in Natural Stories are some of those which have played an important role in psycholinguistics. In particular, we point out the case of object-extracted relative clauses, which have formed the basis of much theorizing about the role of expectations and memory in human sentence processing (Grodner and Gibson 2005; Levy 2008), but which are rare in naturalistic text including Dundee (Roland et al. 2007).

4 Uses of the corpus

Here we review recent work that has used the Natural Stories corpus. The corpus has primarily been used to test models of incremental difficulty in language processing: Shain et al. (2016) use it to test theories of processing slowdown due to work memory retrieval events; Shain et al. (2018a) use it to detect effects of semantic distance on language processing beyond what would be predicted by surprisal-based models; and Yan et al. (2018) use it to test whether reading times for a word in context are affected by the average diversity of contexts in which that word appears. Schijndel and Schuler (2017) and van Schijndel and Linzen (2019) have used the corpus to test explanations for a curious effect in reading times, whereby the reading time of the current word appears to be affected by the surprisal of the *following* word, and to test a neural-network model of adaptation in reading times (van Schijndel and Linzen 2018).

The corpus has also appeared in methodological studies. Shain and Schuler (2018) use the corpus to demonstrate the validity of a new methodology for controlling for "spillover" effects in self-paced reading times, where the effect of the difficulty of a word shows up in the reading times of following words.

In addition to these uses already demonstrated, the corpus has further potential applications. For example, it may be possible to use the corpus as part of a psychometric test for language comprehension, or to use the reading times from the corpus as a source of data for grammar induction models or parsers. It is our hope that, as these studies are carried out, there will be increasing interest from the computational linguistics community in reading time corpora and in psycholinguistics more generally.

5 Conclusion

We have described a new psycholinguistic corpus of English, consisting of edited naturalistic text designed to contain many rare or hard-to-process constructions while still sounding fluent. We believe this corpus will provide an important part of a suite of test sets for psycholinguistic models, exposing their behavior in uncommon constructions in a way that fully naturalistic corpora cannot. We also hope that the corpus as described here forms the basis for further data collection and annotation.

Acknowledgements This work was supported by NSF DDRI Grant #1551543 to R.F., NSF Grants #0844472 and #1534318 to E.G., NSF Grant IIS-1947307 to R.F and NIH career development Award HD057522 to E.F. The authors thank the following individuals: Laura Stearns for hand-checking and correcting the parses, Suniyya Waraich for help with syntactic coding, Cory Shain and Marten van Schijndel for hand-annotating the parses for traces, and Kyle Mahowald for help with initial exploratory analyses of the SPR data. The authors also thank Nancy Kanwisher for recording half of the stories (the other half was recorded by E.G.), Wade Shen for providing initial alignment between the audio files and the texts, and Jeanne Gallee for hand-correcting the alignment.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix: Syntactic features coded in Sect. 3.5

Here we describe the syntactic features of the corpus which were reported in Sect. 3.5. Where necessary, we give examples of each syntactic feature. We categorize the features into conjunction features, relative clause features, ambiguity features, displacement features, and miscellaneous. We also associate each construction below with its abbreviation(s) in Fig. 7.

Conjunction

- **Local/nonlocal VP conjunction** (`local. / non.local.)VP.conjunction`. Conjunction of VPs in which the head verbs are adjacent (local) or not adjacent (nonlocal). Local example: *The man sang and danced*. Nonlocal example: *The man sang a song and danced a dance*.

- **Local/nonlocal NP conjunction** `(local. / non.local.)NP.conjunction`. Conjunction of NPs in which the head nouns are adjacent (local) or not adjacent (nonlocal). Local example: *Rewarded with land and fame*. Nonlocal example: *The people of Bradford and the people who knew them.*
- **CP conjunction** `CP.conjunctions`. Conjunction of CPs with explicit quantifiers. Example: *I know that you are a doctor and that you are a criminal*.

Relative clauses

- **Restrictive/nonrestrictive SRC** `SRC.restr / SRC.non.restr`. Subject-extracted relative clauses with either restrictive or nonrestrictive semantics. We marked relative clauses as restrictive if they served to restrict the domain of possible referents and nonrestrictive if they simply provided extra information. Restrictive example: *The man that knows Bob*. Nonrestrictive example: *The snow, which was white, fell everywhere*.
- **Restrictive/nonrestrictive ORC** `ORC.restr / ORC.non.restr`. Object-extracted relative clauses with either restrictive or nonrestrictive semantics. Example: *The man that Bob knows*.
- **No-relativizer ORC** `no.relativizer.ORC`. An object-extracted relative clause without an explicit relativizer, e.g. *The man Bob knows*.
- **Noncanonical ORC** `ORC.non.canon`. An object-extracted relative clause where the subject of the relative clause is not a pronoun. Example: *The man that the woman knows*.
- **Adverbial relative clause** `adverbial.RC`. An relative clause with an extracted adverbial. Example: *the valley where you would find the city of Bradford*.
- **Free relative clause** `free.relative`. A relative clause not modifying a noun. Example: *What I knowis that Bob is a doctor.*

Ambiguity

- **NP/S ambiguity** `NP.S.ambig`. A local ambiguity where it is unclear momentarily whether a clause is an NP or the subject of a sentence. For example, in the sentence *I know Bob is a doctor*, after reading *I know Bob* it is not clear whether Bob is an NP object of *know* or the beginning of an embedded clause.
- **Main verb/reduced relative ambiguity (easy/hard)** `MV.RR.ambig.EASIER / MV.RR.ambig.HARD`. A local ambiguity between a main verb and a reduced relative clause. For example, *The horse raced past the barn fell*. We divide these into easy and hard cases based on the annotators' judgment about how confusing the local ambiguity is in context.
- **PP attachment ambiguity** `attachment.ambig`. A global ambiguity where a PP could attach to one of two NPs. For example, in a sentence such as *The daughter of the colonel on the balcony*, it is not clear whether it is the daughter or the colonel who is on the balcony.

Displacement

- ***Tough* movement** `tough.mvt`. Cases where an adjective is modified by an infinitive verb phrase from which an object has been extracted. Example: *The point is hard to see*.
- **Parentheticals** `parenthetical`. Additional material that interrupts or lies outside the syntactic structure of the rest of the sentence; constructions that would be marked as "parataxis" in Universal Dependencies. These do not necessarily have to be marked with orthographic parentheses. Example: *There was once, legend has it, a fearful boar.*
- **Topicalization** `topicalization`. Cases where an NP is moved to the front of a sentence to serve as its topic. Example: *The history of Korea, I know nothing about.*
- **Question with *wh* subject** `question_wh_subj`. Questions with *wh*-movement of the subject. Example: *Who walked into the room?*
- **Question with other *wh* word** `question_wh_other`. Questions with *wh*-movement of anything other than the subject. Example: *Who did Bob see?*

Miscellaneous

- **Nonlocal SV** `non.local.SV`. The appearance of any material between a verb and the head of its subject. Example: *The man with the hat ran away.*
- **Nonlocal verb/DO** `non.local.verb.DO`. The appearance of any material between a verb and its direct object. Example: *The man ate quickly the sandwich.*
- **Gerund modifiers** `gerund.modifier`. A case of a verb phrase modifying a noun. Example: *The man walking down the street is tall.*
- **Sentential subject** `sent.subj`. A sentence where the subject is an embedded clase. Example: *The fact that Bob is a doctor is interesting*.
- **Infinite VP subject** `inf.VP.subject`. A sentence whose subject is an infinite verb phrase. Example: *To eat sweets is forbidden.*
- **Postnominal adjectives** `post.nominal.adj`. Adjectives which follow their nouns. Example: *The moon, full and bright*.
- **Idiom** `idiom`. Any idiomatic expression, such as *busy as beavers.*
- **Quotation** `quotation`. Any directly-reported speech. Example: *The woman said "I am here"*.
- **It-clefts** `it.cleft`. Example: *It was Mary that Bob saw.*
- **even...than construction** `even.than`. Example: *Even taller than Mary.*
- **if...then construction** `if...then.constr`. Example: *If you go, then I go.*
- **as...as construction** `as.as.constr`. Example: *Bob was as angry as Mary.*
- **so...that construction** `so.that.constr`. Example: *Bob was so angry that he was shaking.*
- **Yes-no Question** `question_YN`. Example: *Is Mary here?*

References

Bachrach, A., Roark, B., Marantz, A., Whitfield-Gabrieli, S., Cardenas, C., Gabrieli. J. D. E. (2009). Incremental prediction in naturalistic langauge procesing: An fMRI study. Unpublished manuscript.

Barrett, M., Agić, Ž., Søgaard, A. (2015). The Dundee treebank. In *The 14th international workshop on treebanks and linguistic theories (TLT 14)*, pp. 242–248.

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1).

Boston, M. F., Hale, J. T., Vasishth, S., & Reinhold, K. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301–349.

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008 .

Fossum, V. L. R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics*. Association for Computational Linguistics (pp. 61–69).

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

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, *45*(4), 1182–1190.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Futrell, R., Mahowald, K., Gibson, E. (2015). Quantifying word order freedom in dependency corpora. In *Proceedings of the third international conference on dependency linguistics (Depling 2015)* (pp. 91–100). Uppsala, Sweden.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, *29*(2), 261–290.

Husain, S., Vasishth, S., & Srinivasan, N. (2015). Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, *8*(2)

Just, M. A.., Carpenter, P. A.., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228.

Kennedy, A. (2003). *The Dundee corpus [CD-ROM]*. Dundee: Psychology Department, The University of Dundee.

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*(3), 601–618. https://doi.org/10.1080/17470218.2012.676054.

Klein, D., Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting on association for computational linguistics-Volume 1* (pp. 423–30).

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.

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*(1), 12.

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

Liu, H (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.

Luong, M. -T., O'Donnell, T. J., Goodman, N. D. (2015). Evaluating models of computation and storage in human sentence processing. In *CogACLL* (p. 14).

Mitchell, J., Lapata, M., Demberg, V., Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 196–206).

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *Computational linguistics and intelligent text processing* (pp. 3–16). Springer.

Nivre, J., Scholzm, M. (2004). Deterministic dependency parsing of English text. In *Proceedings of the 20th international conference on computational linguistics* (pp. 64–70). Geneva, Switzerland, Aug 23–Aug 27 , COLING.

Roark, B., Bachrach, A., Cardenas, C., Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1-Volume 1* (pp. 324–333). Association for Computational Linguistics.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379. https://doi.org/10.1016/j.jml.2007.03.002.

Shain, C., Futrell, R., van Schijndel, M., Gibson, E., Schuler, W., Fedorenko, E. (2018a). Evidence of semantic processing difficulty in naturalistic reading. In *CUNY 2018.*

Shain, C., Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *EMNLP 2018.*

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (pp. 49–58). Osaka, Japan. Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/W16-4106.

Shain, C., van Schijndel, M., Schuler, W. (2018b). Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Proceedings of the workshop on linguistic and neuro-cognitive resources (LiNCR 2018).*

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

van Schijndel, M., Linzen, T. (2018). A neural model of adaptation in reading. In *2018 conference on empirical methods in natural language processing (EMNLP 2018).*

van Schijndel, M., Linzen, T., (2019). Can entropy explain successor surprisal effects in reading? In *Proceedings of the society for computation in linguistics (SCiL).*

van Schijndel, M., Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL.*

van Schijndel, M., Schuler, W. (2017). Approximations of predictive entropy correlate with reading times. In *Proceedings of the 39th annual meeting of the cognitive science society (CogSci 2017).*

Wu, S., Bachrach, A., Cardenas, C., Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp 1189–1198).

Yan, M., Kliegl, R., Richter, E. M., Nuthmann, A., & Shu, H. (2010). Flexible saccade-target selection in Chinese reading. *The Quarterly Journal of Experimental Psychology*, *63*(4), 705–725.

Yan, S., Mollica, F., Tanenhaus, M. (2018). A context constructivist account of contextual diversity. In *Proceedings of the 40th annual meeting of the cognitive science society (CogSci 2018).*

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.