The Natural Stories Corpus

Richard Futrell[1], Edward Gibson[1], Harry J. Tily[2], Idan Blank[1], Anastasia Vishnevetsky[1], Steven T. Piantadosi[3], and Evelina Fedorenko[4,5]
[1]MIT Department of Brain and Cognitive Sciences [2]Netflix, Inc.
[3]University of Rochester Department of Brain and Cognitive Sciences
[4]Massachusetts General Hospital Department of Psychiatry
[5]Harvard Medical School Department of Psychiatry
{futrell, egibson, iblank, evelina9}@mit.edu,
hal.tily@gmail.com, staseyvi@mail.med.upenn.edu

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. Here we give an overview of the content of 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 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 relative clauses. Corpora of naturally-occurring text are unlikely to contain these constructions. More generally, models of human language comprehension are more likely to make distinct predictions for sentences that cause difficulty for humans, rather than for sentences that are easy to process. For instance, models of comprehension difficulty based on memory integration cost during parsing (Gibson, 2000; Lewis and Vasishth, 2005) will predict effects when the memory spans required for parsing are large, but most syntactic dependencies in naturally-occurring text are short (Temperley, 2007; Futrell et al., 2015). In general, situations that cause high processing difficulty might be rare in naturally-occurring text, because text written and edited in order to be easily understood.

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

[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.

51,501 word tokens in 2,368 sentences from British newspaper editorials, along with eyetracking data from 10 experimental participants. A dependency parse of the corpus is released in 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 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., 2008; Boston et al., 2011) 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 A.V., E.F., E.G. and S.P. by taking existing publicly available texts and editing them to use many 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. The texts and their sources are listed in Table 1.

The mean number of lexical words per sentence is 21.1, around the same as the Dundee corpus (21.7). 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.

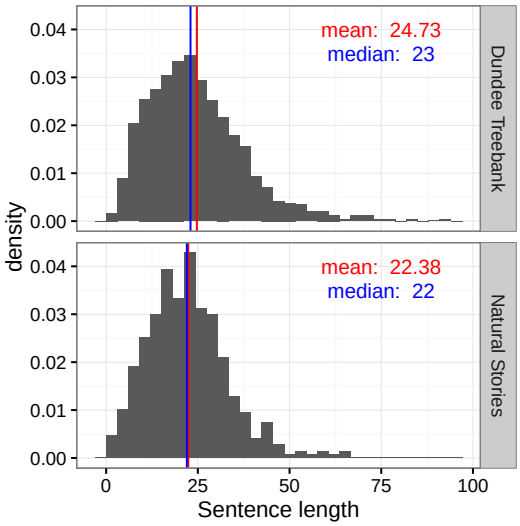

Figure 1: Histograms of sentence length (in tokens, including punctuation) in Natural Stories and the Dundee corpus.

> 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.

Figure 2: Sample text from the first story.

In Figure 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 Universal Dependencies-style parses (Nivre, 2015) automatically converted from the corrected parse trees using the Stanford Parser.

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

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

Table 1: Stories with titles and sources.

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

Table 2: Summary of stories by length.

moving window display; spaces were masked. Each participant read 5 stories per HIT. 19 participants read all 10 stories, and 3 participants stopped after one story. Each story was accompanied by 6 comprehension questions. We discarded SPR data from a participant's pass through a story if the participant got less than 5 questions correct (89 passes through stories excluded). We also excluded RTs less than 100 ms or greater than 3000 ms. Figure 3 shows histograms of RTs per story.

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 correlated the Spearman correlation of that subject's RTs on a story with average RTs from all other subjects on that story. Thus for each story we get one ISC statistic per subject. Figure 4 shows histograms of these statistics per story.

3.3.2. Psycholinguistic Sanity Checks

As a sanity check for our RT data, we checked that basic psycholinguistic effects obtain in it. Some of the most robust predictors of reading time 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-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. Specifically, we fit a model predicting reading time from log frequency, one predicting reading time from word length (measured in orthographic characters), and one predicting reading time from log probability under a trigram model. Word and trigram counts are collected from the Google Books n-grams corpus, summing over years from 1990 to 2013; we make these counts available with the corpus. Each regression is a mixed-effects regression with subject and story as random intercepts (models with random slopes did not converge), meaning that we control for by-subject and by-story variability.

Predictor	$\hat{\beta}$	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

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. We predict and find negative effects of log frequency and log probability and a positive effect of word length. All p values are < 0.001.

The results of the regressions are shown in Table 3; we report results from the maximal converging models. 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. 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 and explained in Appendix A. Figure 5 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 nonlocal VP conjunction, nonrestrictive SRCs, idioms, adjective conjunction, noncanonical ORCs, local NP/S ambiguities, and it-clefts.

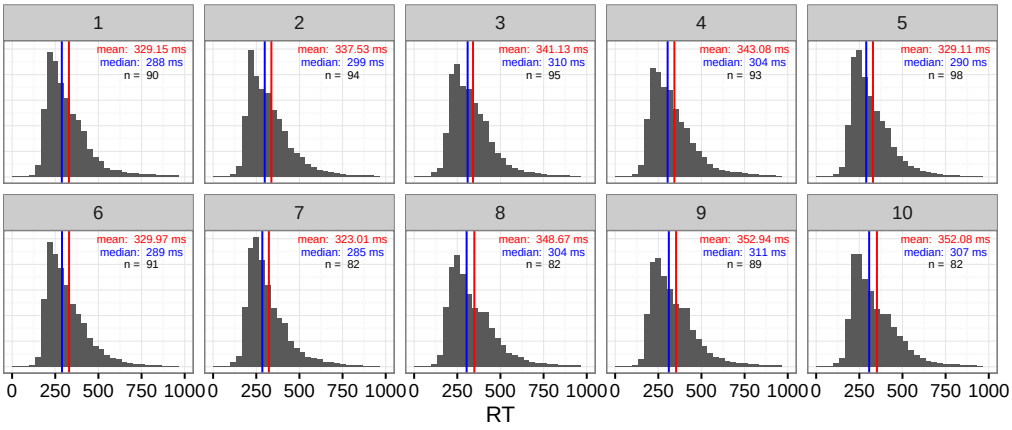

Figure 3: Histograms of SPR RTs per story, after data exclusion.

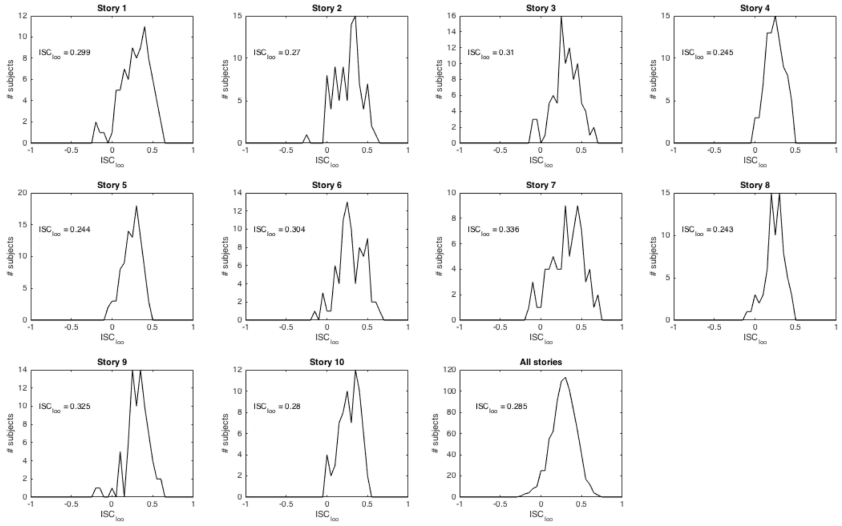

Figure 4: Leave-one-out Inter-Subject Correlations (ISCs) of RTs per story. In the panels, ISC_{loo} gives the average leave-one-out ISC for that story.

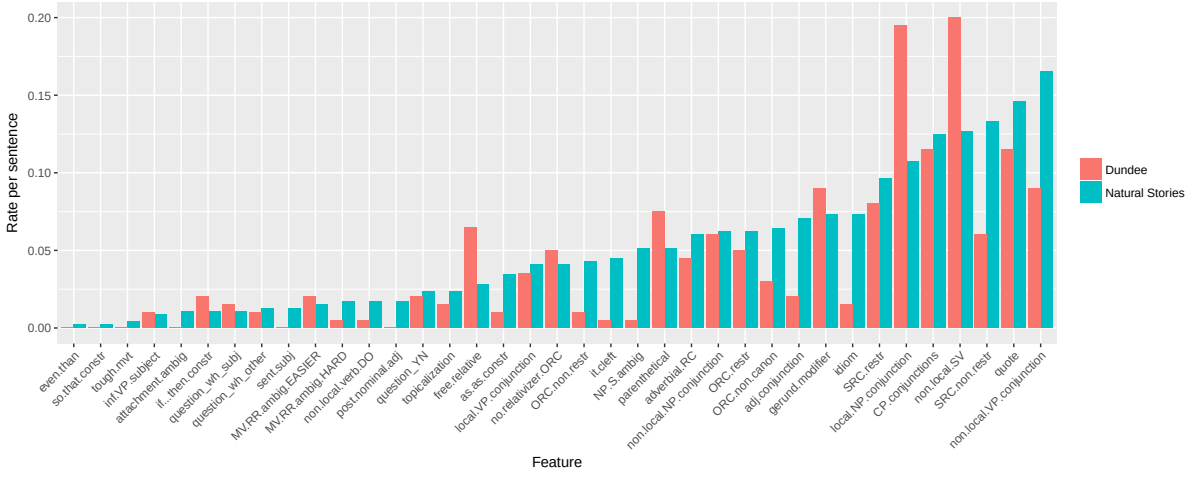

Figure 5: Rates of marked constructions in the Natural Stories corpus and in 200 randomly sampled sentences from the Dundee corpus.

4. 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.

Acknowledgments

This work was supported by NSF DDRI grant #1551543 to R.F., NSF grants #0844472 and #1534318 to E.G., 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.

5. Bibliographical References

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

Barrett, M., Agić, Ž., and Søgaard, A. (2015). The Dundee treebank. In *The 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 242–248.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., and 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., and Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

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

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

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

Frank, S. L. and 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., and Vigliocco, 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., and 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., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, et al., editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126.

Husain, S., Vasishth, S., and 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., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2):228.

Kennedy, A., Pynte, J., Murray, W. S., and 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. PMID: 22643118.

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

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430.

Kliegl, R., Grabner, E., Rolfs, M., and 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., and 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.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

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

Mitchell, J., Lapata, M., Demberg, V., and 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*, pages 196–206.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Roark, B., Bachrach, A., Cardenas, C., and 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*, pages 324–333. Association for Computational Linguistics.

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

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

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

Wu, S., Bachrach, A., Cardenas, C., and Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1189–1198.

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

A Syntactic features coded in Section 3.4.

Here we describe the syntactic features of the corpus which were reported in Section 3.4.. 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.

Conjunction

- Local/nonlocal 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. 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.*
- Sentential conjunction. Conjunction of sentences. Example: *I sang and you danced*.
- CP conjunction. 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. 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. Object-extracted relative clauses with either restrictive or nonrestrictive semantics. Example: *The man that Bob knows*.
- No-relativizer ORC. An object-extracted relative clause without an explicit relativizer, e.g. *The man Bob knows*.
- Noncanonical ORC. 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. An relative clause with an extracted adverbial. Example: *the valley where you would find the city of Bradford*.
- Free relative clause. A relative clause not modifying a noun. Example: *What I know is that Bob is a doctor.*

Ambiguity

- NP/S ambiguity. 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). 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. 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. 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. 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. 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. Questions with *wh*-movement of the subject. Example: *Who walked into the room?*
- Question with other *wh* word. Questions with *wh*-movement of anything other than the subject. Example: *Who did Bob see?*

Miscellaneous

- Nonlocal 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. The appearance of any material between a verb and its direct object. Example: *The man ate quickly the sandwich.*
- Gerund modifiers. A case of a verb phrase modifying a noun. Example: *The man walking down the street is tall*.
- Sentential subject. A sentence where the subject is an embedded clase. Example: *The fact that Bob is a doctor is interesting*.
- Postnominal adjectives. Adjectives which follow their nouns. Example: *The moon, full and bright*.
- Idiom. Any idiomatic expression, such as *busy as beavers*.
- Quotation. Any directly-reported speech. Example: *The woman said "I am here"*.
- It-clefts. Example: *It was Mary that Bob saw*.
- even...than construction. Example: *Even taller than Mary.*
- if...then construction. Example: *If you go, then I go.*
- as...as construction. Example: *Bob was as angry as Mary.*
- so...that construction. Example: *Bob was so angry that he was shaking.*
- Yes-no Question. Example: *Is Mary here?*