Deep Syntactic Annotations for Broad-Coverage Psycholinguistic Modeling

Cory Shain, Marten van Schijndel, William Schuler
Ohio State, Johns Hopkins, Ohio State
Columbus, Baltimore, Columbus
shain.3@osu.edu, mvansch2@jhu.edu, schuler.77@osu.edu

Abstract

This paper presents new hand-corrected deep syntactic annotations for the sentences in two broad-coverage psycholinguistic datasets: the Dundee eye-tracking corpus (Kennedy et al., 2003) and the Natural Stories self-paced reading corpus (Futrell et al., 2017). These texts are more ecologically valid than experiment-specific constructed stimuli, allowing researchers to probe the sentence comprehension process in a naturalistic setting. Deep syntactic annotations such as categorial grammars allow direct access to phenomena like non-local or conjoined semantic argument dependencies which are relevant to many questions about sentence processing but are difficult to compute from common markup frameworks such as Penn Treebank or Universal Dependencies. Previously no gold-standard deep syntactic markups have been available for either Dundee or Natural Stories. The deep syntactic representation used for the proposed annotations (Nguyen et al., 2012) has been shown to (1) facilitate direct extraction of long-distance dependencies as well as many other syntactic constructions of interest, (2) support accurate automatic parsing, and (3) generate surprisal estimates that correlate with measures of processing difficulty (van Schijndel and Schuler, 2015). These annotations can be used for any psycholinguistic inquiry in which predictors must be computed from latent syntax trees.

Keywords: psycholinguistics, broad-coverage, treebank, categorial grammar, incremental processing

1. Introduction

Recent developments in probabilistic parsing and statistical analysis of large heterogeneous datasets have facilitated a growing interest in "broad-coverage" studies of human sentence processing in which the linguistic stimuli are rich and naturalistic rather than carefully constructed for a particular experimental purpose (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013). Instead of manipulating linguistic variables experimentally, such studies estimate measures of main effect and control variables from corpora, often using hierarchical statistical models like linear mixed effects regression (LME) (Demberg and Keller, 2008; van Schijndel et al., 2013b) or generalized additive models (GAM) (Smith and Levy, 2013) to introduce statistical rather than experimental controls.

While some linguistic variables (e.g. incremental surprisal) are best estimated in an automatic fashion using appropriate tools (van Schijndel et al., 2013a, for example), others (e.g. non-local dependency length, incremental parser operations, syntactic categories, etc.) might benefit from the use of expert syntactic annotations, which can be less noisy than automatic parses. This work presents hand-corrected deep syntactic annotations for two large broad-coverage English-language corpora: Dundee (Kennedy et al., 2003) and Natural Stories (Futrell et al., 2017). The syntactic markup used has been shown to support accurate recovery of long-distance dependencies (Nguyen et al., 2012) and to correlate with human behavior (van Schijndel and Schuler, 2015).

2. Background

2.1. Broad-coverage sentence processing research

Research into human sentence processing is concerned with understanding the mechanisms and computational procedures used by the brain to decode the linguistic signal and construct a mental representation of meaning. An important source of evidence about the structure of the human sentence processing mechanism is incremental processing effort, which can be studied using behavioral (e.g. self-paced reading, eye-tracking) or neuro-cognitive (e.g. electro/magnetoencephalography, functional magnetic resonance imaging) measures. Many theories of human sentence processing make predictions about the expected processing difficulty at a given point in an utterance as a function of syntactic features of the utterance. For example, Dependency Locality Theory (Gibson, 2000) predicts processing difficulty proportional to the length of syntactic dependencies to preceding words in the utterance. By contrast, associative memory models of sentence processing (Lewis and Vasishth, 2005; Rasmussen and Schuler, 2017) predict processing effort as a function of cue decay, which can be indexed by distance between certain decisions of a left-corner parser (Johnson-Laird, 1983).

A rich psycholinguistic literature explores theories such as these using stimuli constructed by the experimenters in order to manipulate variables of interest. For example, Grodner and Gibson (2005) manipulated dependency length by presenting subjects with sentences like

(1) The reporter who sent the photographer to the editor hoped for a story.

The use of constructed stimuli affords direct experimental control over the variable of interest as well as minimization of possible linguistic confounds. For many designs, there is also no need to model the response to every word in the utterance, only to those words that participate in critical regions as defined by the experiment (e.g. where long dependencies are resolved).

However, this experimental control of linguistic properties of the stimulus may come at the cost of introducing other confounds that might affect participants' responses. For example, the task of comprehending sentences like (1) pre-

primitive types				type-combining operators	
V	finite verb clause	**S**	top-level utterance	**-a**	argument expected ahead
I	infinitive clause	**Q**	subject-auxiliary inverted	**-b**	argument expected behind
B	base-form clause	**C**	complementized finite verb	**-c**	conjunct expected ahead
L	participial clause	**F**	complementized infinitive	**-d**	conjunct expected behind
A	adjectival/predicative clause	**E**	complementized base-form	**-g**	gap-filler
R	adverbial clause	**N**	nominal clause / noun phrase	**-h**	heavy-shift / extraposition
G	gerund clause	**D**	determiner / possessive	**-i**	interrogative pronoun
P	particle	**O**	non-possessive genitive	**-r**	relative pronoun
				-v	passive

Table 1: Nguyen et al. (2012) primitive types and type-combining operators.

sented in isolation is distinct in many ways from the usual conditions of human sentence processing. First, the words and constructions that appear in the stimuli rarely reflect the distributional characteristics of typical language use — in fact, constructed stimuli intentionally deviate from these distributional characteristics in order to test the hypothesis in question. Responses to unnatural utterances may not generalize to sentence processing in more typical cases. Second, constructed stimuli are usually presented in isolation, possibly introducing an inflated burden of pragmatic inference. For example, (1) contains three definite noun phrases, but participants are given no linguistic or situational context against which to interpret them. Third, the fact of presenting unusual sentences in isolation may signal to subjects that the implicit use of language for communication is being temporarily suspended. If it is not clear to subjects that the experimenters are trying to communicate a substantive message about the reporter, photographer, and editor, they may abandon their usual sentence processing routines and instead use task-specific heuristics. Added to the foregoing concerns about ecological validity is the fact that data collected in this way are at best difficult to reappropriate in order to study questions outside the purview of the original experimental design.

Broad-coverage studies are therefore an important complement to constructed-stimulus studies. By relaxing the requirement for direct manipulation and bringing linguistic confounds under statistical rather than experimental control, sentence processing researchers can mitigate the aforementioned problems by exposing subjects to context-rich connected texts and performing word-by-word modeling of responses to linguistic predictors computed from the stimuli. Such paradigms have been used to explore the sensitivity of the human sentence processing apparatus to variables like surprisal (Frank and Bod, 2011; Fossum and Levy, 2012; Demberg et al., 2013; van Schijndel and Schuler, 2015) and dependency locality (Demberg and Keller, 2008; Shain et al., 2016). By providing hand-corrected deep syntactic annotations for two large broad-coverage corpora, the current work aims to support further research along these lines.

2.2. Deep Syntactic Annotations

Explorations of memory effects in sentence processing typically require some indicator of precisely when during sentence processing certain syntactic arguments are attached and which semantic argument dependencies are associated with those syntactic arguments. This linguistic precision requires a deep syntactic annotation of the stimulus sentences that are the source of the modeled psycholinguistic phenomena.

The deep syntactic annotation used in this resource is a generalized categorial grammar (GCG) of English (Nguyen et al., 2012).[1] This representation both (1) defines a small set of licensed syntactic compositions, like e.g. Combinatory Categorial Grammar (Steedman, 2000), and (2) restricts the inventory of types to those needed to enforce grammatical constraints, like e.g. Head-driven Phrase Structure Grammar (Pollard and Sag, 1994). This markup assigns a category or sign type to each meaningful sequence of words in each sentence, consisting of a *primitive clausal type* (e.g. a verb-headed clause **V**, base-form clause **B**, noun phrase or nominal clause **N**, etc.) lacking one or more dependent types, each delimited by a *type-combining operator* (e.g. **-a** to define a missing argument expected immediately ahead in the utterance, **-b** to define a missing argument expected immediately behind in the utterance, **-c** or **-d** to missing conjuncts ahead or behind, and so on). Table 1 lists the complete set of primitive types and type-combining operators in this markup. The marked up categories are constrained by a set of grammatical inference rules which assign semantic dependencies in cases of argument and modifier attachment, and keep track of these dependencies through phenomena like passive alternations, conjunctions, filler-gap constructions, extrapositions, and subject-auxiliary inversions. Table 2 lists a set of grammatical inference rules that constrain possible annotations, and Figures 1 and 2 show some example marked up sentences. This rich markup can be reliably automatically reannotated from Penn Treebank markup (Marcus et al., 1993) if available, or automatically suggested by a robust PCFG parser (Petrov and Klein, 2007, for example) trained on existing reannotated markup, with or without hand correction.

Unlike Penn Treebank markup (Marcus et al., 1993) or syntactic dependency markup (de Marneffe et al., 2006; Nivre et al., 2016, for example), unbounded dependencies are represented locally in the Nguyen et al. (2012) markup, permitting access to a store of incomplete non-local dependencies at any point in parsing. The syntactic composition rules

[1] Further in-depth details of the GCG specification are available here: `http://go.osu.edu/gcg`.

grammatical inference rules	
A	argument attachment ahead or behind
C	conjunct attachment ahead or behind
E	extraction
G	gap-filler attachment ahead or behind
H	heavy-shift / extraposition
I	interrogative clause attachment
M	modifier attachment ahead or behind
Q	subject-auxiliary inversion
R	relative clause attachment
T	type conversion / argument elision
U	auxiliary attachment ahead or behind
V	passive
X	it-extraposition
Z	zero-head introduction

Table 2: Grammatical inference rules, adapted from Nguyen et al. (2012).

map deterministically to semantic composition operations, allowing certain incremental semantic processing decisions to be recovered from syntactic annotations. The advantage of this markup for psycholinguistic modeling is the direct access it affords to incremental non-local dependency features and semantic composition operations, both of which may play a role in human sentence processing. In this respect, this markup is similar to HPSG, LFG, and various instantiations of categorial grammar such as CCG. In fact, with appropriate reannotation scripts, the present markup can in principle be used to generate these other markups automatically. GCG was selected for the present annotation because of previous work showing evidence that it has several psycholinguistically desirable properties: better automatic recovery of filler-gap and other non-local dependencies than parsers trained on dependency representations (Nguyen et al., 2012), better control over syntactic frequency confounds in psycholinguistic data than controls based on Penn Treebank annotations (van Schijndel et al., 2014), and correlation between human response times and surprisal estimates computed by an incremental parser trained on this representation (van Schijndel and Schuler, 2015).

2.3. Corpora annotated

This work presents annotations for the Dundee (Kennedy et al., 2003) and Natural Stories (Futrell et al., 2017) reading time corpora. Dundee contains eye-tracking measures from 10 subjects reading 20 editorials from *The Independent* newspaper. The stimulus set contains a total of 51,502 tokens and 2,368 sentences (Kennedy et al., 2003), with a total of 260,065 fixation events across all subjects. Dundee has been in existence for some time and has been used for psycholinguistic hypothesis testing in a variety of studies (Demberg and Keller, 2008; Frank, 2009; Frank and Bod, 2011; Fossum and Levy, 2012; Smith and Levy, 2013; Demberg et al., 2013). A treebank exists for Dundee (Barrett et al., 2015) using syntactic dependencies (Nivre et al., 2016), but syntactic dependencies are optimized for efficient parsing and as a result are not as closely related to

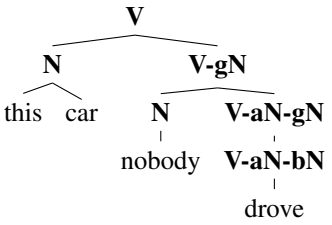

Figure 1: A simple sentence, *This car nobody drove*, annotated with Nguyen et al. (2012) markup. At the top, the noun phrase, *this car*, attaches as a gap filler (**-gN**) using inference rule G. Below that, the noun phrase, *nobody*, attaches as the first argument (**-aN**) of the verb *drove* using inference rule A. Below that, the gap filler is identified as an extracted second argument (**-bN**) of the verb *drove*, using inference rule E.

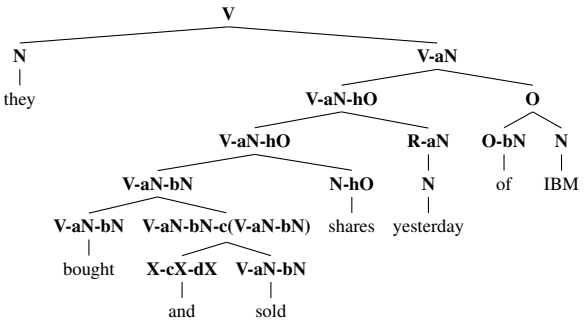

Figure 2: A more complex sentence, involving conjunction (**-c**, **-d**) of a transitive verb (**V-aN-bN**), and extraposition of a genitive complement (**-hO**) across an adverb (**R-aN**).

semantic argument structure as dependencies derived from categorial grammar markup.[2] This distinction is apparent in cases of conjunctions, extrapositions, and filler-gap extractions from embedded clauses. For example, because syntactic dependency representations typically analyze conjunctions as linked lists of conjuncts, they are not able to assign different analyses to high and low attachment readings of *old* in the conjunction *old men and women* (see Figure 3), since the word *men* serves as both the high and low site for modifier attachment. Markup based on categorial grammar or phrase structure is able to distinguish these different attachment analyses using different bracketings.

Natural Stories contains self-paced reading measures from 181 subjects reading (some subset of) 10 short stories on Amazon Mechanical Turk. The stimulus set contains a total of 10,245 tokens and 485 sentences, with a total of 848,768 reading events across all subjects. Shain et al. (2016) used Natural Stories to test hypotheses about retrieval costs during sentence processing. Natural Stories distributes with

[2]Note that there exists an enhanced deep markup for universal dependencies (Schuster and Manning, 2016) which can mitigate some of the problems with shallow dependency annotations. However, no hand-corrected deep dependency markups are available for either Dundee or Natural Stories, and many of the representational advantages of deep dependency markups are provided directly by the current GCG annotation scheme.

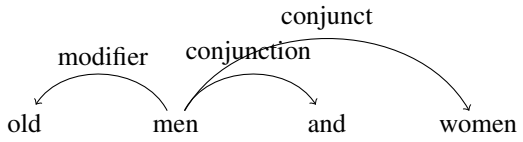

Figure 3: Syntactic dependency analysis of both high and low attachment of *old* in the conjunction *old men and women*.

hand-corrected Penn Treebank (PTB) annotations, as well as uncorrected syntactic dependency annotations automatically generated from the PTB source. The deep syntactic annotations described here complement these existing annotations by providing rich phrase-structural representations that locally encode syntactic dependency information and deterministically represent semantic composition operations, allowing researchers to more easily study the role of these features in human sentence processing.

3. Methods

Stimuli from both corpora were syntactically annotated via single-expert hand-correction of automatically-generated deep syntactic markup. In the case of Dundee, the automatic source parses were produced by the Petrov and Klein (2007) parser trained on an automatic translation from PTB to deep syntactic trees in sections 2–21 of the Wall Street Journal corpus, using the Nguyen et al. (2012) reannotation algorithm. In the case of the Natural Stories corpus, the automatic annotations were produced by applying the Nguyen et al. (2012) reannotation algorithm directly to the gold PTB-style trees supplied by the authors of the corpus (Futrell et al., 2017). Hand-correction of the Natural Stories deep syntactic reannotation was performed by a single expert annotator. The automatic parses of the Dundee corpus were partitioned in two and each set was hand-corrected by a distinct expert annotator. Depending on the complexity of the sentence, the principal annotators consulted at times with one or more additional experts before deciding on a final annotation.

4. Access

Annotations for both corpora are distributed through the ModelBlocks repository (van Schijndel and Schuler, 2013), which can be accessed at the following URL: `https://github.com/modelblocks/modelblocks-release`.[3] ModelBlocks only includes the syntactic annotations, not the stimuli themselves. Once users are in possession of the source stimuli, ModelBlocks provides scripts to automatically generate the complete treebanks by combining the annotations with the source stimuli. The Natural Stories corpus is publicly available and can be accessed at the following URL: `https://github.com/languageMIT/naturalstories`. Because of licensing restrictions on the stimuli, the Dundee corpus is not publicly available. Interested researchers must contact the authors directly (Kennedy et al., 2003) for access.

[3]This Github URL supersedes the one in the cited paper.

5. Conclusion

Because the Dundee and Natural Stories corpora are broad-coverage rather than constructed to target a particular question, they provide a more realistic measure of subjects' typical response to language stimuli, and the data they contain can be reappropriated to test a variety of hypotheses about the human sentence processing system, some of which may not have been anticipated at the time of data collection. The deep syntactic representation used here provides access to incremental non-local dependency features and semantic composition operations, which are of potential import to a range of sentence processing questions. Thus, the hand-corrected deep syntactic annotations presented in this work should have lasting value by supporting an open set of such investigations into possible determinants of sentence processing difficulty.

6. Bibliographic References

Barrett, M., Agić, Z., and Søgaard, A., (2015). *The Dundee Treebank*, pages 242–248. Association for Computational Linguistics.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

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

Demberg, V., Keller, F., and Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

Fossum, V. and Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of CMCL 2012*. Association for Computational Linguistics.

Frank, S. and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*.

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proc. Annual Meeting of the Cognitive Science Society*, pages 1139–1144.

Futrell, R., Gibson, E., Tily, H., Vishnevetsky, A., Piantadosi, S., and Fedorenko, E. (2017). The natural stories corpus. *arXiv*, (1708.05763).

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126, Cambridge, MA. MIT Press.

Grodner, D. J. and Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, 29:261–291.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA, USA.

Kennedy, A., Pynte, J., and Hill, R. (2003). The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

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

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Nguyen, L., van Schijndel, M., and Schuler, W. (2012). Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of COLING 2012*, pages 2125–2140, Mumbai, India.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Pollard, C. and Sag, I. (1994). *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Rasmussen, N. and Schuler, W. (2017). Leftcorner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science*.

Schuster, S. and Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *LREC 2016*.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., and Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Computational Linguistics for Linguistic Complexity Workshop*, pages 49–58. Association for Computational Linguistics.

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

Steedman, M. (2000). *The syntactic process*. MIT Press/Bradford Books, Cambridge, MA.

van Schijndel, M. and Schuler, W. (2013). An analysis of frequency- and memory-based processing costs. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics.

van Schijndel, M. and Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics.

van Schijndel, M., Exley, A., and Schuler, W. (2013a). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.

van Schijndel, M., Nguyen, L., and Schuler, W. (2013b). An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In *Proc. of CMCL 2013*. Association for Computational Linguistics.

van Schijndel, M., Schuler, W., and Culicover, P. W. (2014). Frequency effects in the processing of unbounded dependencies. In *Proc. of CogSci 2014*. Cognitive Science Society.