Text Genre and Training Data Size in Human-Like Parsing

John T. Hale and **Adhiguna Kuncoro**
DeepMind / London
{jthale,akuncoro}@google.com

Keith B. Hall
Google Research / New York
kbhall@google.com

Chris Dyer
DeepMind / London
cdyer@google.com

Jonathan R. Brennan
University of Michigan / Ann Arbor
jobrenn@umich.edu

Abstract

Domain-specific training typically makes NLP systems work better. We show that this extends to cognitive modeling as well by relating the states of a neural phrase-structure parser to electrophysiological measures from human participants. These measures were recorded as participants listened to a spoken recitation of the same literary text that was supplied as input to the neural parser. Given more training data, the system derives a better cognitive model — but only when the training examples come from the same textual genre. This finding is consistent with the idea that humans adapt syntactic expectations to particular genres during language comprehension (Kaan and Chun, 2018; Branigan and Pickering, 2017).

1 Introduction

Natural language processing (NLP) systems based on deep neural networks are sensitive to the amount and type of training data that they receive. A "data hungry" method may not work well until it is supplied with sufficient examples (e.g. Yogatama et al., 2019). Likewise, transfer to a different textual genre may be poor (Petrov and McDonald, 2012). This is the classic problem of domain adaptation[1] which arises in many areas of NLP.

This paper revisits domain adaptation in the context of human-like parsing. With this human-like aspect in mind, we consider models that use linguistically-plausible trees (see Frank, 2011 for a review) and operate incrementally from left to right (e.g. Steedman, 2000). We quantify the fit to human language performance using freely-available electrophysiological data (henceforth: EEG) that was elicited by a pre-existing literary text (Brennan and Hale, 2019).

These EEG data come from a naturalistic stimulus, and in virtue of their higher temporal resolution, are more detailed than reading times or plausibility judgments. Hale et al. (2018) were the first to model them using a neural parser. In that study, textual training data came from the same book as did the human participants' stimuli. While this yielded a model that was quite well-matched to the EEG modeling task, its training data was confined to just 1543 sentences (24K words). In contrast, recent studies suggest that neural language models require substantially more data to achieve human-like linguistic competence (Gulordava et al., 2018; Futrell et al., 2019; Frank and Hoeks, 2019).

We investigate this question of data size together with a contrast between textual domains or "genres". Modeling human neural signals, we find that in-domain training leads to a better and better fit as more examples are added to the training set whereas with out-of-domain data, more examples do not help. This is interesting given the consistent reductions in language modeling perplexity with more data, which we observe across both domains. We further find that, across all amounts of training data, models that incorporate linguistically-plausible phrase structure achieve a better fit to human EEG data than models that do not. This suggests that phrase structure should play an important role in human-like models of language comprehension, even in models that benefit from large training data.

2 Methodology

We proceed by comparing parsing systems that are based on Recurrent Neural Network Grammars (Dyer et al., 2016; Wilcox et al., 2019, henceforth; RNNG) and trained according to fourteen

[1] It remains quite difficult to reconcile human-like incremental parsing with high performance out-of-domain; many researchers take a nonincremental whole-sentence approach (Gildea, 2001; Baucom et al., 2013; Joshi et al., 2018).

complexity metric:	surprisal of hypotheses in beam (Hale, 2001; Roark et al., 2009)
amount of training data:	39832,$\{100, 250, 500, 750\}$K, 1M and 1437575 sentences
genre:	newspaper text (Graff et al., 2005) and lexically-similar literature (Gutenberg)

Table 1: Training regimes

different regimes. These training regimes cross seven different amounts of training data with two different genres of writing. The question across all regimes is: how well does a complexity metric derived from these parsers improve the modeling of EEG data from human participants engaged in language comprehension of the same text? Note that the dependent variable here is not parsing performance or language modeling perplexity of the RNNG, but rather the helpfulness, as regards cognitive modeling, of a derived measure based on the intermediate states that the RNNG visits during decoding of the stimulus text that the human participants heard. Table 1 summarizes these training regimes.

3 Materials

3.1 Lexically-similar literature

In the first genre, we ranked e-books from the freely-available Project Gutenberg collection according to the CosineTop50 metric from McClosky et al. (2010). This (dis)similarity metric compares vectors whose components are attestation counts from the 50 most frequent words in a reference text. Here, the reference text is the one that human EEG study participants listened to – the first chapter of Alice in Wonderland. Table 2 shows some highly-ranked books on this scheme. It yielded largely juvenile literature from the 19$^{\text{th}}$ century aimed at girls. The average sentence length in this selection was 17 words.

dissimilarity	title	author
0.0584	The Admiral's Caravan	Charles E. Carryl
0.0620	The Secret Garden	Frances Hodgson Burnett
0.0628	The Lodger	Marie Belloc Lowndes
0.0687	The Girls and I: A Veracious History	Mary Louisa Stewart Molesworth
0.0689	What Timmy Did	Marie Adelaide Belloc
0.0724	Little Miss Peggy	Mrs. Molesworth
0.0725	The Girls of St. Olave's	Mabel Mackintosh
0.0741	The Celebrity at Home	Violet Hunt
0.0750	I've Married Marjorie	Margaret Widdemer
0.0752	The Forged Note	Oscar Micheaux
0.0755	Mary Erskine	Jacob Abbott
0.0758	The Bountiful Lady	Thomas Cobb
0.0758	Legacy	James H Schmitz
0.0763	Some Little People	George Kringle
0.0774	In the Wilderness	Robert Hichens

Table 2: Alice-like books from Project Gutenberg

3.2 Newspaper text

In the second genre, we randomly sampled news articles from the English Gigaword corpus (Graff et al., 2005). This sampling was made regardless of the particular national source, i.e. Agence France-Press, New York Times or Xinhua News Agency. Sentences in this sample were, on average, 20 words long. This out-of-domain text had a CosineTop50 dissimilarity level of 0.56.

3.3 Presumptive Trees

Both genres were parsed using a re-implementation of the Berkeley parser (Petrov et al., 2006) to yield presumptively-correct, "silver-grade" trees. This Berkeley parser was trained on a diverse set of annotated data. These include the Penn Treebank's Wall Street Journal materials (Marcus et al., 1993), the Question Treebank (Judge et al., 2006), Ontonotes (e.g. Pradhan and Ramshaw, 2017) and Parsing-the-Web Corpora (Petrov and McDonald, 2012). In a manual inspection of randomly sampled silver trees, the only obvious mistakes were tagging errors (1 newspaper, 2 literature), which are not harmful since RNNGs do not use part of speech tags. Indeed, the bracketing and phrase labels on these silver trees appeared to be fully consistent with the Penn Treebank Bracketing Guidelines (Bies et al., 1995). Before being passed to the RNNG as training data, these trees were post-processed to remove empty elements, punctuation, and function tags.

4 Training

4.1 Hyperparameters

We tuned RNNG hyperparameters to optimize development set perplexity. The development set comprises the entire Alice in Wonderland book as obtained from Project Gutenberg.[2] To control for

[2]While the first chapter of Alice in Wonderland that we use as stimuli is thus a subset of the development set, recall from section 2 that our evaluation metric is fit to the human EEG, rather than validation perplexity.

model capacity, across all training regimes we used the same hyper-parameter that consistently achieved good validation perplexities across all configurations.[3]

4.2 Vocabulary

Within each of the two genres, we facilitate comparison across different training set sizes by running the RNNGs with a shared vocabulary derived from the largest training configuration. An attestation frequency cut-off was applied to each of them to ensure broadly comparable rates of out-of-vocabulary words on the development set, Alice in Wonderland. For the lexically-similar literature, this threshold was 5 attestations, whereas for the newspaper text this threshold was 20 attestations. On the validation set, these cut-offs yielded out-of-vocabulary rates of 2.3 to 2.4 percent for newspaper text and 0.5 to 1.28 percent for lexically-similar literature.

4.3 Achieved perplexity

Per-action perplexity levels[4] achieved by the trained RNNG parsing models on the development set are shown in Figure 1. Validation perplexity consistently improves with more and more silver-grade training data, even when that data comes from a different domain.[5]

5 EEG Regression model

Surprisal values from a beam search parser based on RNNG are entered as predictors into a regression model of scalp voltages. Models are fit with the `brm` function in R and model fits were compared using Bayesian model comparison (Vehtari et al., 2017). These regression models include random intercepts for participant along-side predictors to account for factors of non-interest that nevertheless are known to influence sentence processing difficulty (see e.g. Goodkind and Bicknell, 2018). These are: sentence position in the stimulus text, word position within each sentence, acoustic sound power, unigram word frequency in the HAL corpus (Balota et al., 2007). Unigram predictors are included for the previous word, the current word and the next word.

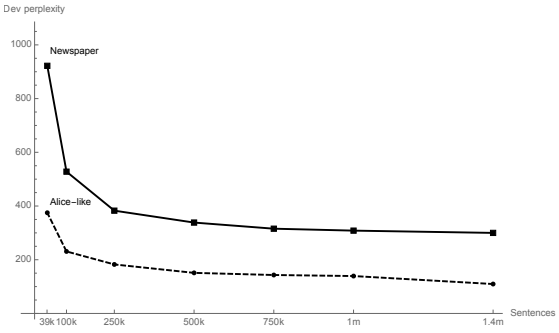

Figure 1: Language modeling perplexity (lower is better) on Alice in Wonderland across the two genres of training data.

The EEG data themselves come from 33 datasets that were collected by Brennan and Hale (2019) while participants listened to the first chapter of Alice in Wonderland.[6] Ocular artifacts and other noise sources are removed from the raw signal using ICA and visual inspection. The EEG data are reduced to a single spatio-temporal region of interest (ROI) comprising the data from anterior channels across both hemispheres between 200 and 400 ms after the onset of content-words. This anterior ROI has, uniquely, shown sensitivity to surprisal values from incremental parsers under a data-driven whole-scalp analysis in Brennan and Hale (2019). Note that this ROI is earlier and more anterior than the usual locus of the N400 component, which typically manifests on central electrodes at or around 400 ms post-stimulus (for a review see Kaan, 2007).

6 Results

Figure 2 plots the goodness-of-fit of a regression model that includes RNNG-derived complexity metrics. As the neural phrase structure parser is trained on increasingly larger corpora, the regression model of EEG amplitudes fits better and better. However, this pattern only obtains with in-domain training data that is lexically-similar to the first chapter of Alice in Wonderland. When trained on newspaper text from the Gigaword cor-

[3] The RNNG hyperparameters are: 2-layer stack LSTM (Dyer et al., 2015), 450 hidden units, and an initial SGD learning rate of 0.3, decayed exponentially with a factor of 0.1 applied after the tenth epoch.

[4] The per-action perplexity is computed based on the joint probability of strings ($\mathbf{x}$) and trees ($\mathbf{y}$), denoted as $p(\mathbf{x}, \mathbf{y})$, which therefore aggregate the perplexity of the next-word prediction with the perplexity of tree-building actions. Approximate inference methods such as important sampling can be used to derive an estimate of $p(\mathbf{x})$ (Dyer et al., 2016).

[5] While relative perplexity levels internal to a genre are comparable in virtue of a shared vocabulary, absolute perplexity values are not directly comparable across the genres since these vocabularies are different.

[6] These data are available at: https://doi.org/10.7302/Z29C6VNH.

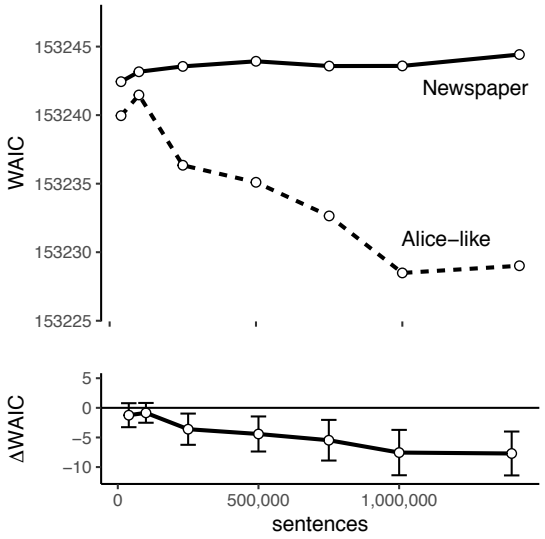

Figure 2: Goodness of fit of surprisal predictions from RNNG at beam size $k = 200$ syntactic analyses on anterior electrodes during the time window 200-400ms after the onset of a spoken word in the first chapter of Alice in Wonderland. The widely-applicable information criterion WAIC is described further in Vehtari et al. (2017). Lower WAIC indicates a better fit.

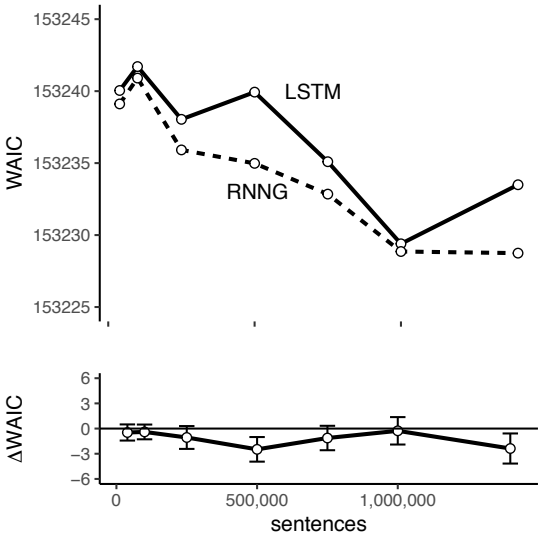

Figure 3: Comparison between RNNG with explicit hierarchical representations and LSTM without such representations. Lower WAIC indicates a better fit to human electrophysiological data.

pus, the goodness of fit to human EEG data remains about the same no matter how much training data is supplied. Tukey's test of additivity indicates an interaction between genre and training set size, $F(1, 5) = 171.3, p < 0.000001$. No effect of surprisal values obtained in electrodes and timepoints corresponding to the N400 component.

Figure 3 compares the RNNG, which explicitly uses phrase structure representations, to an LSTM sequence model which does not (Hochreiter and Schmidhuber, 1997). Here, both models receive the benefit of in-domain training on the lexically-similar literature. But regardless of how much in-domain training data is made available, the explicitly phrase-structural RNNG always offers a better account of the EEG signal.

7 Related Work

The importance of domain adaptation in NLP has been well-established in earlier work (see Daumé III, 2007 and footnote 1), including applications to parsing (Sarkar, 2001; McClosky et al., 2006; Søgaard and Rishøj, 2010; Weiss et al., 2015). Our approach to in-domain data selection is closely related to earlier work in language modeling and machine translation (Keller and Lapata, 2003; Moore and Lewis, 2010; Axelrod et al., 2011). Our focus on large training set responds to recent work in language modeling (Devlin et al., 2019; Radford et al., 2019). The research reported here differs in two ways from these earlier studies. First, we consider the interaction between domain and amount of training data, rather than examining each variable in isolation. Second, we investigate the impact of these variables on cognitive modeling, which reveals a pattern that is different from what we observe in the standard perplexity evaluation. We focus on text genre, rather than online adaptation as van Schijndel and Linzen (2018) do. Despite coming at the problem from a different direction (and using EEG rather than self-paced reading) our results agree with van Schijndel and Linzen in suggesting that some kind of adaptation must be going on in human language comprehension.

8 Conclusion

These comparisons confirm that genre matters. If surprisal describes human linguistic expectations, then we can say that those expectations are better-modeled by a parsing system that benefits from in-domain training. This would follow if, as Kaan and Chun (2018) have suggested, people are able to very rapidly adjust their syntactic expectations to match a particular genre.

Indeed, these expectations seems to be phrase-

structural in nature. Certainly the presence of unigram nuisance predictors in the EEG regression (section 5) and the comparatively worse performance of the LSTM sequence model (Figure 3) render it unlikely that this finding is due to word frequency or superficial co-occurrence. Rather, the neural parser has learned something about 19^{th} century children's literature that can be captured at the syntactic level. Whatever these syntactic properties are, it could have been the case that they were equally learnable from newswire text or Alice-like books. But in fact Alice-like books generalize better to human EEG data.

This use of moment-by-moment processing difficulty to adjudicate between trained NLP systems offers a reminder that the quest for human-level performance in language technology should always be understood in relation to a particular kind of linguistic performance in a particular genre.

Acknowledgments

The authors would like to thank Laura Rimell for valuable discussion of this work. This material is based upon work supported by the National Science Foundation under Grant No 1607441.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

David A. Balota, Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. 2007. The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Eric Baucom, Levi King, and Sandra Kübler. 2013. Domain adaptation for parsing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 56–64, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. Linguistic Data Consortium. University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-95-06-0.

Holly P. Branigan and Martin J. Pickering. 2017. An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, 40:e282.

Jonathan R. Brennan and John T. Hale. 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PLOS ONE*, 14(1):1–17.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proc. of ACL*, pages 256–263.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*. arXiv:1810.04805.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China. Association for Computational Linguistics.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California. Association for Computational Linguistics.

Robert Frank. 2011. Phrase structure. In Patrick Colm Hogan, editor, *The Cambridge Encyclopedia of the Language Sciences*, pages 621–622. Cambridge University Press.

Stefan L. Frank and John C.J. Hoeks. 2019. The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times. In *Proceedings of the 41$^{\text{st}}$ Annual Meeting of the Cognitive Science Society*, Montreal.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2005. English Gigaword second edition. LDC2005T12.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Gutenberg. 2019. Project Gutenberg.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2727–2736, Melbourne, Australia. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1199, Melbourne, Australia. Association for Computational Linguistics.

John Judge, Aoife Cahill, and Josef van Genabith. 2006. Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 497–504, Stroudsburg, PA, USA. Association for Computational Linguistics.

Edith Kaan. 2007. Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass*, 1:571–591.

Edith Kaan and Eunjin Chun. 2018. Syntactic adaptation. In Kara D. Federmeier and Duane G. Watson, editors, *Current Topics in Language*, volume 68 of *Psychology of Learning and Motivation*, chapter 4, pages 85–116. Academic Press.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proc. of NAACL*.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. Presented at the First Workshop on Syntactic Analysis of Non-Canonical Language held in cooperation with NAACL-HLT.

Sameer Pradhan and Lance Ramshaw. 2017. Ontonotes: Large scale multi-layer, multi-lingual, distributed annotation. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 521–554. Springer Netherlands, Dordrecht.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 324–333.

Anoop Sarkar. 2001. Applying co-training methods to statistical parsing. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 1–8.

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of*

the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4704–4710, Brussels, Belgium. Association for Computational Linguistics.

Anders Søgaard and Christian Rishøj. 2010. Semi-supervised dependency parsing using generalized tri-training. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1065–1073.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

David Weiss, Christopher Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proc. of ACL*.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. In *Proc. of NAACL*.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. `arXiv:1901.11373`.