Profiling neural grammar induction on morphemically tokenised child-directed speech

Mila Marcheva[1], Theresa Biberauer[2,3,4], and Weiwei Sun[1]

[1]Department of Computer Science & Technology, University of Cambridge, UK
[2]Department of Theoretical and Applied Linguistics, University of Cambridge, UK
[3]General Linguistics Department, Stellenbosch University, South Africa
[4]Linguistics Department, University of the Western Cape, South Africa
{mmm67,mtb23,ws390}@cam.ac.uk

Abstract

We investigate the performance of state-of-the-art (SotA) neural grammar induction (GI) models on a morphemically tokenised English dataset based on the CHILDES treebank (Pearl and Sprouse, 2013). Using implementations from Yang et al. (2021b), we train models and evaluate them with the standard F1 score. We introduce novel evaluation metrics—depth-of-morpheme and sibling-of-morpheme—which measure phenomena around bound morpheme attachment. Our results reveal that models with the highest F1 scores do not necessarily induce linguistically plausible structures for bound morpheme attachment, highlighting a key challenge for cognitively plausible GI.

1 Introduction

Functional morphemes are a key focus of current generative research in First Language Acquisition (FLA) due to their role in shaping the overall structure of language (Guasti, 2017; Dye et al., 2018; Biberauer, 2019). The computational task of grammar induction (GI) takes as input a corpus of unlabelled sentences and outputs the predicted hierarchical structure for these sentences based purely on the latent statistics of the corpus; see §2.2 for an overview of GI and Figure 3 for an example of induced structures. GI thus provides a lower bound on the types of grammatical structures that can be inferred from linguistic signal alone, particularly when appropriate acquisitional metrics are employed, and recent advances in GI (Kim et al., 2019) necessitate a reevaluation of its performance in the context of FLA. This paper is concerned with bridging the gap between the state-of-the-art (SotA) in GI and in FLA by evaluating the performance of neural GI models on morphemically tokenised English child-directed speech (CDS).

To provide a more cognitively realistic setup (see §3) we propose a modification to the input of SotA neural GI systems: we only use CDS, which we morphemically tokenise (see §3.2 and §3.3) in order to reflect the salience of functional morphemes in FLA (Shi, 2013). We select SotA neural grammar induction models: Compound Probabilistic Context-Free Grammar (C-PCFG; Kim et al., 2019), Neural PCFG (N-PCFG; Kim et al., 2019), and Tensor Decomposition PCFG (TN-PCFG; Yang et al., 2021b).

We evaluate the models using the standard measure – F1 score. Furthermore, we propose two original evaluation metrics—depth-of-morpheme and sibling-of-morpheme—specific to evaluating the attachment of functional morphemes (see §4.1). Our original evaluation metrics reveal that the models with highest F1 do not necessarily induce the most linguistically plausible structures.

2 Background

2.1 Functional morphemes

The distinction between lexical and functional items is fundamental in the study of human language structure (Dye et al., 2018). Functional items encode grammatically salient information and serve as the locus for the grammatical organisation of language, as per the *Borer-Chomsky Conjecture* (Borer, 1984; Baker, 2008). During the initial focus on lexical items exhibited in FLA (Brown, 1973; Shi and Werker, 2003), functional items serve as high-frequency "edge elements", which aid in segmentation of language input and in identifying the category of the lexical item they occupy predictable positions in relation to (Mintz, 2013; Biberauer, 2019). Thus, for example, *the* consistently signals the left-edge of a (definite) noun phrase while *-ed* consistently signals the right edge of a (past-tense) lexical verb. The edge significance approach is considered SotA in FLA (see i.a. Christophe et al. (2008) and Dye et al. (2018) for further discussion). By tokenising bound functional morphemes (see §3.3), we reflect their salience in FLA.

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 47–54
May 3, 2025 ©2025 Association for Computational Linguistics

Inflectional morphology, the productive combination of lexical and functional items, starts to emerge in child-produced speech in stages related to the overall vocabulary size and mean length of utternace (MLU) (Brown, 1973; Devescovi et al.; Ravid et al., 2020). English is a quite strongly isolating language, so most functional items appear as free morphemes (separate words), and there are few bound functional morphemes, which appear as affixes (see §3.3). In GI systems the bound functional morphemes are ignored because tokens are treated as atomic units. Our approach—morphemic tokenisation—addresses the loss of "edge" information that follows from this practice by splitting, for example, *runs* into the lemma *run* and the bound functional morpheme *-s* before training. This allows the system to learn the grammatical rules governing bound morphemes, which play a crucial role in syntax.

2.2 Neural grammar induction

Grammar induction (GI) is the task of finding the latent structure of a natural language, a grammar, based on a set of raw sentences from the language, a corpus. Most statistical attempts at GI rely on a sequence of POS tags as input (Carroll and Charniak, 1992; Klein, 2005; Perfors et al., 2011), and attempts to use raw text underperform (Klein and Manning, 2004). Using POS tags (or other derivatives of raw text) is unrealistic from an FLA point of view because it postulates the existence of a standalone POS induction system. Neural systems do not require such modification of the input and achieve SotA results (Kim et al., 2019).

The general principle in neural grammar induction systems is to parametrise probabilities of (phrase-structure) rules with neural networks. Dyer et al. (2016) lay the foundations for neural GI with the Recurrent Neural Network Grammar (RNNG), and more recent works include Neural PCFG (N-PCFG; Kim et al., 2019), Compound PCFG (C-PCFG; Kim et al., 2019), Neural Lexicalised PCFG (NL-PCFG; Zhu et al., 2020), Neural Bi-Lexicalised PCFG (NBL-PCFG; Yang et al., 2021a), Tensor Decomposition PCFG (TN-PCFG; Yang et al., 2021b), SimplePCFG (Liu et al., 2023). Character-based PCFG (Jin et al., 2021) has a similar motivation to ours: to utilise the information inside a word. However, we specifically target the smallest standalone linguistic unit, morphemes, instead of naively placing equal importance on all alphanumeric characters. Tsarfaty et al. (2020) provide preliminary support for the marriage of morphological information with neural unsupervised approaches.

3 Experimental setup

3.1 Systems

We perform experiments using C-PCFG, N-PCFG, and TN-PCFG. To optimise the computational resource requirements, we use the implementations of Yang et al. (2021a), and the C-PCFG and N-PCFG experiments rely on SimplePCFG (Liu et al., 2023). All of the systems work with a preset number of non-terminals (nt) and terminals (t). The number of nt and t in our experiments follows the previous experimental setup of Yang et al. (2021a).

3.2 Data

We use the CHILDES Treebank (CHITB; Pearl and Sprouse, 2013), which consists of child-directed speech (CDS) sentences with phrase structure annotations. We use all of the Brown-Adam data for testing because its annotations are most widely verified. The remaining sentences are randomly split between training and validation. Table 1 displays the number of sentences in each split. CDS differs from adult speech, and especially the Penn Treebank (PTB; Marcus et al., 1993), as shown in previous works (Gelderloos et al., 2020; Jones et al., 2023). In this specific instance, it is worth noting that: sentences of length one are common in CDS, but constitute trivial examples for the GI task, so we eliminate them; CHITB consists of a smaller vocabulary and shorter sentences than PTB; CHITB is not canonical (e.g. includes unfinished sentences).

	PTB			CHITB		
	№	S	T	S	T	MT
Train	2-21	39	912	140	643	676
Valid	22	1.7	40	24	129	136
Test	23	2.4	56	16	82	86

Table 1: Count in thousands of sentences (S), standard tokens (T) and morphemic tokens (MT) in PTB WSJ sections (№) and in CHITB.

3.3 Morphemic tokenisation

The data in CHITB comes standardly tokenised, and we additionally render it lowercase and remove punctuation. The procedure for morphemic tokenisation is as follows: 1) identify words with bound functional morpheme endings; 2) ensure the word is not an exception; 3) split the original word into a word lemma and a functional morpheme, using en_core_web_lg (Montani et al., 2020) and regular expressions; 4) save the lemma and functional morpheme.

The **bound functional morphemes** in English of interest in this work are listed below, followed by the percentage that they represent of the training tokens:

- present progressive *-ing*, 2.7%
- regular plural *-s*, 1.64%
- regular past tense *-ed*, 0.82%
- regular third person present tense *-s*, 0.49%

After morphemic tokenisation, the structure of the parse trees also needs to be appropriately modified to reflect the presence of the new tokens. We attach the bound functional morpheme as illustrated in Figure 1. A complete list of cases illustrated with syntactic trees is provided in Appendix A.

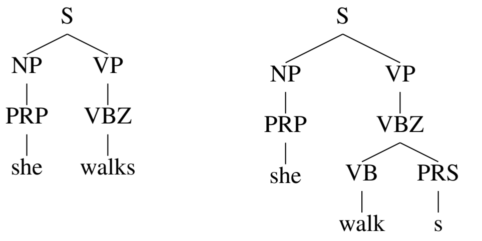

Figure 1: Regular third person present *-s*. (L) Original tree. (R) Rewritten tree post morphemic tokenisation.

The **exceptions** where morphemic tokenisation should not be applied depend on the language, and for English include: plural-only nouns (e.g. *trousers*), as these forms are monomorphemic; irregular forms of third person singular verbs (*is*, *has*, *goes*, *does*); and words which coincidentally end in *-ed* or *-ing* (e.g. *bed* or *sting*).

4 Results and analysis

The standard method of assessing GI is to use a sentence-level F1 score, which is calculated based on the gold annotations of the test set. We present the F1 scores for the different models in Table 2.

From Table 2 it is apparent that the morphemically tokenised data performs on par or better than the standardly tokenised data when using a large number of non-terminals and terminals. The highest F1 is achieved by N-PCFG (nt8192 t16384), where the standard tokenisation slightly outperforms the morphemic tokenisation. Overall systems with a higher number of non-terminals and terminals, which can capture more subtle variation in the data, perform better. The right-branching baseline achieves an F1 score comparable and even higher than for some neural models. This trend is apparent for both standard and morphemic tokenisation because English has a right-branching pattern (Greenberg, 1963). The high performance of right-branching baselines for English is reported for C-PCFG (Kim et al., 2019, Table 1) and for TN-PCFG (Yang et al., 2021b, Table 1).

Model	Morphemic	Standard
Left-branching	14.17	14.83
Right-branching	71.94	73.77
Random trees	36.45	36.61
TN-PCFG (nt9000 t4500)	73.81	45.75
C-PCFG (nt2048 t4096)	68.79	59.86
C-PCFG (nt512 t1024)	41.99	72.95
N-PCFG (nt4096 t8192)	69.19	60.83
N-PCFG (nt8192 t16384)	**78.56**	**79.01**

Table 2: Sentence-level F1 for constituency parses for morphemic and standard tokenisation.

4.1 Functional morpheme evaluation

F1 is reliant on annotations, which for natural languages are prone to ambiguity because the target grammar may not necessarily be known. We devise annotation-independent evaluation metrics focused on the structure of attachment of functional morphemes.

4.1.1 Depth-of-morpheme

We assume that the nodes for bound functional morphemes are sibling nodes for the lexeme they combine with (see §3.3 and Figure 1). To establish whether a bound functional morpheme is correctly attached in the predicted tree, we check whether it is found at the same depth as the lexeme it forms a word with. If the depth differs, then the predicted subtree is incorrect in describing the functional morpheme attachment. We perform depth-of-morpheme evaluation on the models with highest F1: TN-PCFG (nt9000 t4500), N-PCFG (nt8192 t16384), and the right-branching baseline; the results are displayed in Table 3.

	TN-PCFG	N-PCFG	Right-br.
-ed	100	55.11	30.49
-ing	100	41.11	21.37
-s	95.11	95.95	40.88

Table 3: Percentage of bound functional morphemes attached at the correct depth. *-s* has two uses (§3.3).

N-PCFG has the highest F1, but appears not to be expressive enough to encode the examined linguistic phenomena: this is likely because it captures a higher frequency of simple cases. TN-PCFG makes no errors on *-ed* and on *-ing*. The right-branching baseline, although comparable in F1 score with the neural models, underperforms on the task of correctly attaching the bound functional morpheme. These insights highlight the importance of acquisitionally-focused evaluation, because standard NLP measures, such as F1, may obscure task-specific errors.

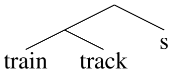

Figure 2: N-PCFG prediction for the structure of the morphemically tokenised phrase "train tracks". The induced structure implies the functional morpheme *-s* combines with the whole noun compound.

Note however that the binary nature of depth-of-morpheme also obscures patterns which may be of linguistic interest. For example, N-PCFG predicts that the plural noun morpheme *-s* attaches to the whole noun compound as displayed in Figure 2, and the depth-of-morpheme of *-s* is therefore incorrect (the expected pattern is for *-s* to attach to the single noun preceding it, as displayed in Figure 5). However, the induced structure might be of linguistic interest because the compounding of the nouns is not implausible. To gain deeper insight, depth-of-morpheme should be used in combination with sibling-of-morpheme, the metric introduced in the following section.

4.1.2 Sibling-of-morpheme evaluation

We next analyse the sibling of the bound functional morpheme in the predicted tree. The sibling is the span of the smallest tree immediately dominating the tree where the functional morpheme node appears; in linguistics, this notion is also referred to as a sister. For example, the sibling of *-ed* in Figure 3 is *knock* (as predicted by TN-PCFG) and the subtree spanning *one down* (as predicted by N-PCFG). The sibling predicted by TN-PCFG is linguistically plausible, whereas the one predicted by N-PCFG is not – not only does it group two words in a grammatically unlikely constituent, but it implies the functional morpheme does not combine with the verb.

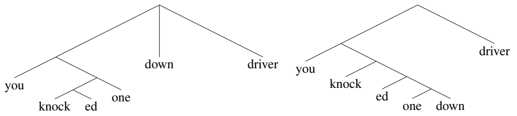

Figure 3: Predicted trees for the sentence "You knocked one down, driver."; TN-PCGF (L) and N-PCFG (R).

To systematically look for patterns in the siblings of morphemes, we look at the siblings' semantic role labels (SRL)[1], obtained via SRL BERT[2] (Shi and Lin, 2019), and whether there is any relation to the *depth-of-morpheme* measure. For an overview of SRLs please consult Jurafsky and Martin (Chapter 21; 2025). Beyond standard SRLs we introduce two more labels: the "straddles boundary" category signifies that the sibling of the morpheme spans more than one semantic role and this kind of attachment is always incorrect, because it poses a grammatically incoherent constituency. The "all O" category, where all of the leaves in the sibling are labelled as (O)utside of a semantic role, applies to cases which may include a constituent boundary, or more rarely where a constituent was missed by the SRL model.

SRL	TN-PCFG		N-PCFG	
	Count	% Correct	Count	%Correct
Overall	1796	95.43	1796	78.12
ARG1	780	95.38	790	*91.65
ARG2	211	*91.94	210	*89.52
V	398	*99.75	118	*98.31
ARGM	63	95.24	72	77.78
all O	32	96.88	40	75.0
ARG0	27	92.59	26	*100
ARG3	3	100	3	100
ARG4	2	100	14	*14.29
strad. b.	6	*0	249	*0

Table 4: Comparison of SRL Tag Performance: TN-PCFG vs. N-PCFG. Statistically significant ($p < 0.05$) difference from the **Overall** marked with *.

Verbal instances are ones where the bound morpheme is attached to a verb identified by the SRL model (Figure 3 illustrates a verbal instance). In the verbal instance, the only correct label for the sibling of the morpheme is V. The TN-PCFG system correctly attaches all instances of *-ed* and *-ing* to a single lexeme with SRL V (also see Table 3), but N-PCFG makes errors where the bound morpheme is attached to lexemes tagged as direct object (ARG1), indirect object (ARG2), adjuncts (ARGM), and others.

[1] Other annotations (e.g. dependencies) may be used.

[2] `https://paperswithcode.com/lib/allennlp/srl-bert`

In the **non-verbal** instance, the morpheme attaches to a lexeme which is not identified as a verb by the SRL tagger, but we nonetheless look to find the functional role of the sibling in the sentence. Here there is no one correct SRL (see Appendix A for the full range of cases). Table 4 displays the percentage of morphemes which are found at the correct depth, grouped by the SRL of their sibling. We perform Fisher's Exact Test (Fisher, 1922) to identify SRLs for which the percentage of correct depth-of-morpheme differs significantly from the overall rate of correct depth-of-morpheme for that system. For TN-PCFG, indirect object (ARG2) siblings of functional morphemes co-occur with a significantly lower depth-of-morpheme correct percentage, especially in comparison with direct object (ARG1) and adjunct (ARGM) siblings, which follow the same as the overall rate and appear to pose less of a challenge for the model. The N-PCFG system has a very high number of siblings of morphemes which include a boundary, which lower the **Overall** depth-of-morpheme correctness for the system. This result again highlights that the system with the highest F1 does not necessitate the correct attachment of functional morphemes: N-PCFG (highest F1) often predicts that the functional morphemes attach to an implausible constituent.

5 Conclusion

We explore how morphemic tokenisation, an insight inspired by FLA, influences neural GI systems. We evaluate the GI systems with F1 score, and conduct further error analysis on the attachment of bound morphemes. Our findings reveal that high F1 scores do not always correspond to linguistically meaningful structures for functional morpheme attachment. In the future, we will apply this methodology to CDS from morphologically rich languages, such as the ones in SPMRL (Goldberg et al., 2014).

6 Limitations

Morphemic tokenisation follows a generativist perspective rather than a theory-neutral approach, so it may not align with non-generativist frameworks. The limitations of the novel evaluation metrics —depth-of-morpheme and sibling-of-morpheme— mainly stem from the fact that their utility depends on morphemically tokenised text. Additionally, there are cases where the binary result of depth-of-morpheme may not be informative enough (e.g. Figure 2), which is why the depth-of-morpheme metric should be used in combination with the sibling-of-morpheme metric.

English is currently the only language with an annotated CDS treebank of suitable magnitude, but our focus on English unfortunately further reinforces the dominance of English in NLP research. Since English is a largely isolating/weakly inflecting language with minimal inflectional morphology, a morphologically complex language would provide a more rigorous test for morphemic tokenisation, with greater potential benefits, but potentially also increased challenges. Future work will expand both the linguistic scope and the experimental design.

References

Mark C. Baker. 2008. The macroparameter in a microparametric world. In *The Limits of Syntactic Variation*, page 351–373. John Benjamins Publishing Company.

Theresa Biberauer. 2019. Children always go beyond the input: The maximise minimal means perspective. *Theoretical Linguistics*, 45(3–4):211–224.

Hagit Borer. 1984. *Parametric syntax : case studies in Semitic and Romance languages / Hagit Borer.* Studies in generative grammar ; 13.

Roger Brown. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Glenn Carroll and Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. Technical Report CS-92-16, Dept. of Computer Science, Brown University, Providence, RI.

Anne Christophe, Séverine Millotte, Savita Bernal, and Jeffrey Lidz. 2008. Bootstrapping lexical and syntactic acquisition. *Language and Speech*, 51(1–2):61–75.

Antonella Devescovi, Maria Cristina Caselli, Daniela Marchione, Patrizio Pasqualetti, Judy Reilly, and Elizabeth Bates. A crosslinguistic study of the relationship between grammar and lexical development. *Journal of Child Language*, 32(4).

Cristina Dye, Yarden Kedar, and Barbara Lust. 2018. From lexical to functional categories: New foundations for the study of language development. *First Language*, 39(1):9–32.

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.

R. A. Fisher. 1922. On the interpretation of 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87.

Lieke Gelderloos, Grzegorz Chrupała, and Afra Alishahi. 2020. Learning to understand child-directed and adult-directed speech. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1–6, Online. Association for Computational Linguistics.

Yoav Goldberg, Yuval Marton, Ines Rehbein, Yannick Versley, Özlem Çetinoğlu, and Joel Tetreault, editors. 2014. *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*. Dublin City University, Dublin, Ireland.

Joseph Harold Greenberg. 1963. *Universals of language*. MIT press.

Maria Teresa Guasti. 2017. *Language acquisition*, 2 edition. A Bradford Book. Bradford Books, Cambridge, MA.

Lifeng Jin, Byung-Doh Oh, and William Schuler. 2021. Character-based PCFG induction for modeling the syntactic acquisition of morphologically rich languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4367–4378, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gary Jones, Francesco Cabiddu, Doug J. K. Barrett, Antonio Castro, and Bethany Lee. 2023. How the characteristics of words in child-directed speech differ from adult-directed speech to influence children's productive vocabularies. *First Language*, 43(3):253–282.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition, chapter 21: Semantic Role Labeling and Argument Structure. Online manuscript released January 12, 2025.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound probabilistic context-free grammars for grammar induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385, Florence, Italy. Association for Computational Linguistics.

Dan Klein. 2005. *The unsupervised learning of natural language structure*. Stanford University.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 478–485, Barcelona, Spain.

Wei Liu, Songlin Yang, Yoon Kim, and Kewei Tu. 2023. Simple hardware-efficient PCFGs with independent left and right productions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1662–1669, Singapore. Association for Computational Linguistics.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330.

Toben H. Mintz. 2013. The segmentation of sub-lexical morphemes in english-learning 15-month-olds. *Frontiers in Psychology*, 4.

Ines Montani, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Maxim Samsonov, Jim Geovedi, Jim Regan, György Orosz, Paul O'Leary McCann, Søren Lind Kristiansen, Duygu Altinok, Roman, Leander Fiedler, Grégory Howard, Wannaphong Phatthiyaphaibun, Explosion Bot, Sam Bozek, Mark Amery, Yohei Tamura, Björn Böing, Pradeep Kumar Tippa, Leif Uwe Vogelsang, Ramanan Balakrishnan, Vadim Mazaev, GregDubbin, jeannefukumaru, Jens Dahl Møllerhøj, and Avadh Patel. 2020. explosion/spaCy: v3.0.0rc: Transformer-based pipelines, new training system, project templates, custom models, improved component API, type hints & lots more.

Lisa S. Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*.

Andrew Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

Dorit Ravid, Emmanuel Keuleers, and Wolfgang Dressler. 2020. *Emergence and early development of lexicon and morphology*, pages 593–633.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint*.

Rushen Shi. 2013. Functional morphemes and early language acquisition. *Child Development Perspectives*, 8(1):6–11.

Rushen Shi and Janet F. Werker. 2003. The basis of preference for lexical words in 6-month-old infants. *Developmental Science*, 6(5):484–488.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online. Association for Computational Linguistics.

Songlin Yang, Yanpeng Zhao, and Kewei Tu. 2021a. Neural bi-lexicalized PCFG induction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2688–2699, Online. Association for Computational Linguistics.

Songlin Yang, Yanpeng Zhao, and Kewei Tu. 2021b. PCFGs can do better: Inducing probabilistic context-free grammars with many symbols. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1487–1498, Online. Association for Computational Linguistics.

Hao Zhu, Yonatan Bisk, and Graham Neubig. 2020. The return of lexical dependencies: Neural lexicalized PCFGs. *Transactions of the Association for Computational Linguistics*, 8:647–661.

A All cases of tree rewriting

All cases of tree rewriting are shown below in Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, Figure 9. The original trees are on the left, and the rewritten trees are on the right. The trees are constructed as explained in §3.3.

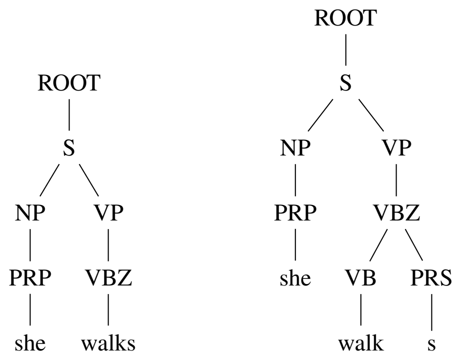

Figure 4: Regular 3rd person present *-s*.

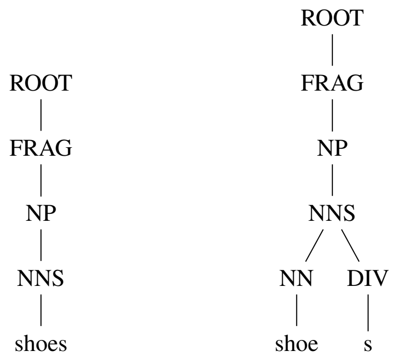

Figure 5: Regular plural -s.

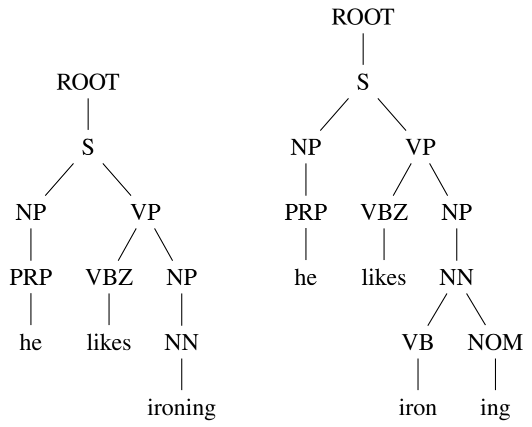

Figure 6: Nominal *-ing*

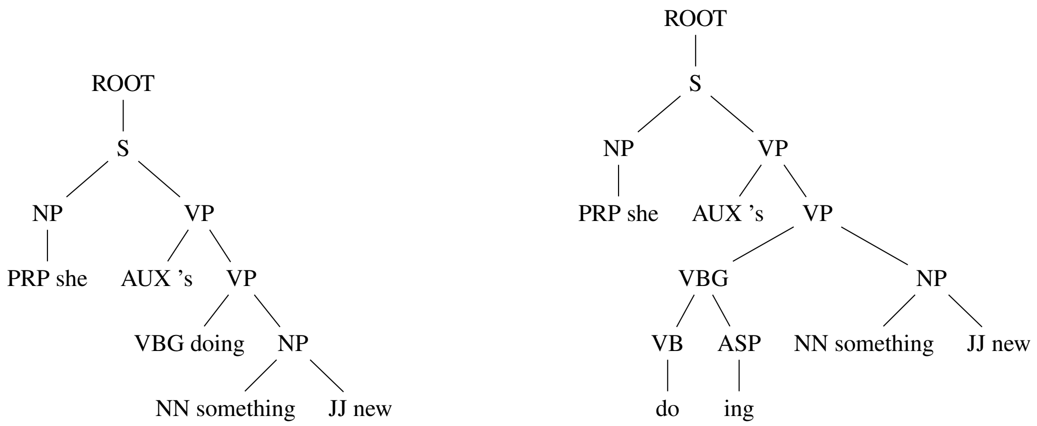

Figure 7: Progressive *-ing*.

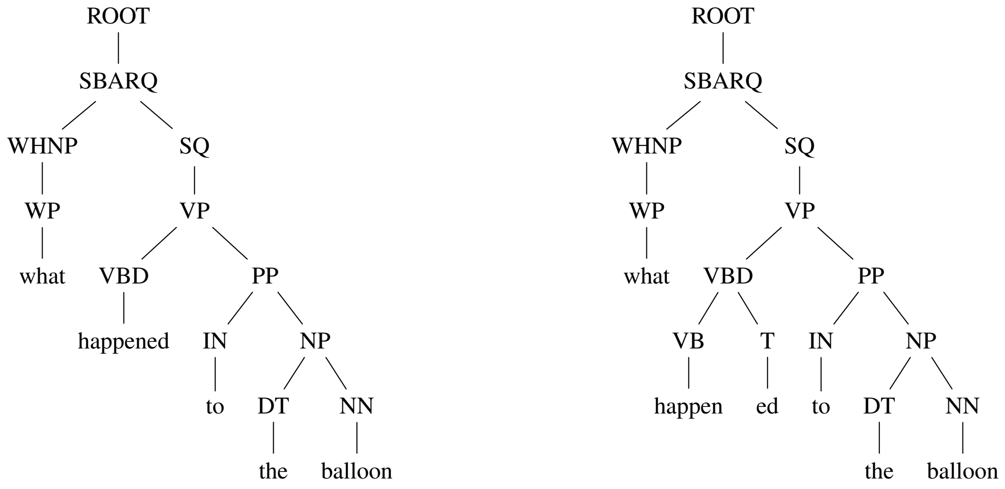

Figure 8: Regular past *-ed*.

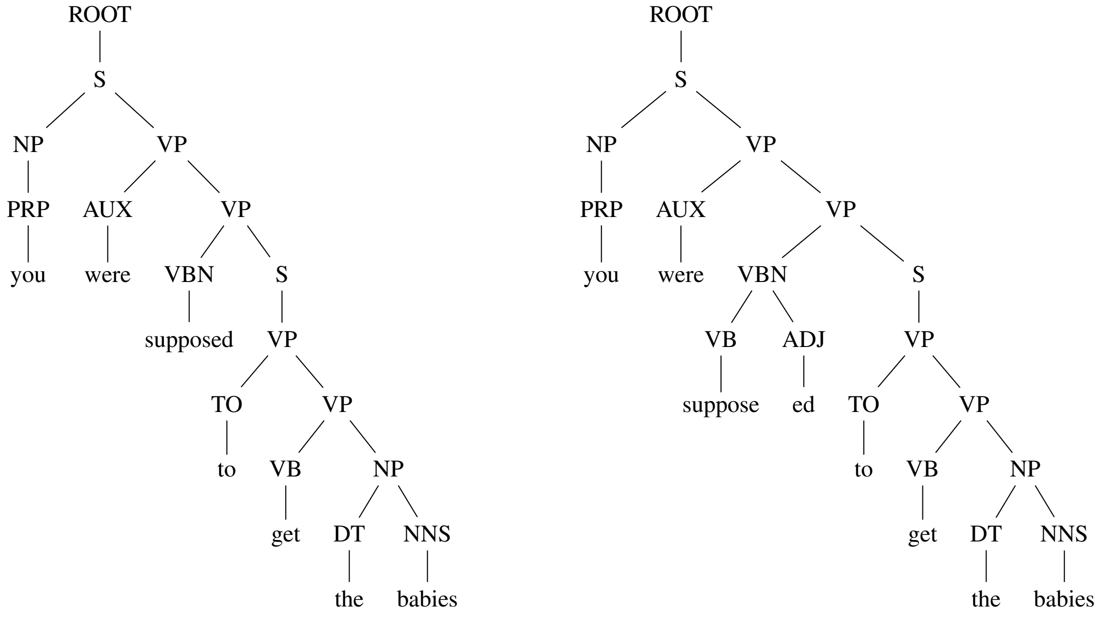

Figure 9: Adjectivial *-ed*.