PS-CHILDES-BG: a constituency treebank of Bulgarian morphemically tokenised child-directed speech

Mila Marcheva-Nash
Department of Computer Science & Technology,
University of Cambridge, UK
mmm67@cam.ac.uk

Weiwei Sun
Department of Computer Science & Technology,
University of Cambridge, UK
ws390@cam.ac.uk

Abstract

We present PS-CHILDES-BG, a constituency treebank of 2,434 Bulgarian morphemically tokenised child-directed speech sentences. We supply a morphemic tokenisation algorithm for Bulgarian, which uses UD parses as input. Additionally, we introduce a linguistically informed algorithm for Bulgarian which converts UD parses into constituency parses. These resources are intended to further the language acquisition research of Bulgarian via computational methods.

Keywords: constituency treebank, morphemic tokenisation, child-directed speech

1 Introduction

A central question in first language acquisition (FLA) is how children acquire functional categories (Dye et al., 2018). Functional categories are expressed by functional items and encode grammatical content (Baker, 2008), in contrast with lexical categories, which encode semantic content. Traditionally, FLA relies on labour-intensive methods, such as manually annotating child-directed speech (CDS) with syntactic features and descriptively analysing the resulting data. Computational methods can complement traditional approaches. For example, grammar and category induction experiments can be used to approximate the corresponding category or grammar acquisition processes. Seminal computational experiments for FLA include Carroll and Charniak (1992); Cartwright and Brent (1997); Buttery (2006); Parisien et al. (2008). Computational experiments allow for confounding factors to be controlled for and can provide a lower bound on the types of grammatical information that may be recoverable from linguistic signal alone.

In computational modelling of FLA, standard tokenisation only captures the free functional items, such as the article *the*, but ignores bound functional morphemes, such as the progressive aspect suffix *-ing*. Morphemic tokenisation allows both bound and free functional items to be represented, leading to a more complete account of the grammatical material available in the input. The motivation for morphemic tokenisation of CDS is also grounded in FLA literature. Functional items have a higher frequency than lexical items and occur in predictable positions (Lieven, 2010; Biberauer, 2019). The salience of functional items for distributional analysis makes morphemic tokenisation compatible with both generativist (Guasti, 2017) and usage-based (Lieven, 2016) accounts of FLA. Morphemic tokenisation of English CDS has recently been explored for functional category induction and grammar induction (Marcheva et al., 2025a,b; Marcheva-Nash et al., 2026b). The need for morphemic tokenisation is even stronger in morphologically richer languages, such as Bulgarian, where a larger number of functional items manifest as bound morphemes.

Caregivers are known to use a special register when communicating with infants, known as child-directed speech (CDS), which differs from adult-directed speech in prosody (higher pitch, exaggerated contours), lexical choice, and utterance length (Fernald, 1989; Soderstrom, 2007; Kuhl, 2006). The usage of CDS in computational language modelling is recommended in order to make the computational experiments more cognitively plausible, i.e. to more closely resemble the real-world learning problem. CDS corpora exist for very few languages: English (Pearl and Sprouse, 2013; Yang et al., 2025); Dutch (Odijk et al., 2018); Hebrew (Szubert et al., 2024); Japanese (Butler et al., 2022); Bulgarian UD-CHILDES-BG (Marcheva-Nash et al., 2026a). CDS corpora predominantly follow Universal Dependencies (UD; de Marneffe et al., 2021). However, constituency (phrase structure) parses can be more informative for models which operate over hierarchical phrase structure,

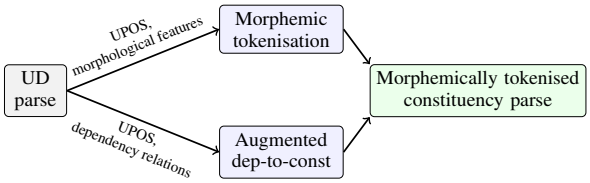

Figure 1: Pipeline for deriving a morphemically tokenised constituency parse from a UD parse.

such as grammar induction, because they represent the hierarchical structure of language directly (Carnie, 2021). A morphemically tokenised constituency parse resource currently exists only for English CDS (Pearl and Sprouse, 2013; Marcheva et al., 2025b). The algorithms and the treebank we contribute in this paper would allow for the current research based on morphemically tokenised CDS to be extended to Bulgarian.

We present a morphemic tokeniser for Bulgarian inflectional morphology, which we apply to Bulgarian CDS. We also develop a linguistically informed automatic conversion system from dependency to constituency parses. Both algorithms rely on UD parses as input, which in our case are manually corrected, but automatic parses could also be used. While the two algorithms can be used independently, we combine them, as seen in Figure 1, to produce a linguistically informative resource, PS-CHILDES-BG, a constituency treebank of 2,434 Bulgarian morphemically tokenised CDS sentences, of which 15% are manually verified. The existing resources are presented in section 2, the algorithms are described in more detail in section 3, an evaluation of the algorithms is presented in section 4. The morphemic tokeniser for Bulgarian, the modified dependency to constituency algorithm, as well as the resulting treebank are available on GitHub.[1]

2 Existing resources

Bulgarian NLP resources include the constituency treebank BulTreeBank (BTB; Simov et al., 2002; Simov and Osenova, 2003) and its dependency counterpart, UD-BTB (Osenova and Simov, 2017). Bulgarian also has a CHILDES (MacWhinney, 1992) resource: CHILDES-BG provides longitudinal developmental data (Popova and Popov, 2020). Parsing tools for Bulgarian include the Berkeley constituency parser (Petrov et al., 2006), the dependency parsers Stanza (Qi et al., 2020) and CLASSLA-Stanza (Terčon and Ljubešić, 2023; Ljubešić et al., 2024).

At the subword level, several resources are also relevant. BulStem (Nakov, 2003) is an inflectional stemmer for Bulgarian. Iliev et al. (2015) present a dictionary-based Bulgarian lemmatiser. However, neither of these specifically tokenises functional morphemes. MorphScore (Arnett et al., 2025; Arnett and Bergen, 2025) is a tool for evaluating morphemic tokenisation, which includes a Bulgarian dataset based on UD-BTB. More generally, work on morpheme segmentation is represented in SIGMORPHON, a shared task on morpheme segmentation (Batsuren et al., 2022), but it excludes Bulgarian. The most closely related resources for functional morpheme tokenisation are morphemic tokenisation of English CDS (Marcheva et al., 2025a) and the morphemically tokenised English CHILDES-TB (Marcheva et al., 2025b).

3 Pipeline

In this section we give an overview of the data, the morphemic tokenisation algorithm, the augmented dependency to constituency (dep-to-const) conversion algorithm, including detail on the morphemically tokenised version. The workflow is illustrated in Figure 1.

3.1 Data

The UD parses we use as input are the manually verified CDS portion of UD-CHILDES-BG (Marcheva-Nash et al., 2026a), a UD corpus based on CHILDES-BG (Popova and Popov, 2020), but any automatic UD parses can be used as input. The current system only supports projective parses, and there are 2,434 such CDS parses in UD-CHILDES-BG. The manually verified section of UD-CHILDES-BG includes sentences from the input of five children, stratified by age. The age range of the five children is between one and three years old, which encompasses the key stages for the emergence of functional categories (Brown, 1973; Popova, 2023).

3.2 Morphemic tokenisation

There are several challenges specific to morphemic tokenisation of Bulgarian, which are not encountered in the same task for English. Several competing segmentation choices are often possible. In such cases, it is important to select a single lin-

[1] https://github.com/milamarcheva/UD-CHILDES-BG

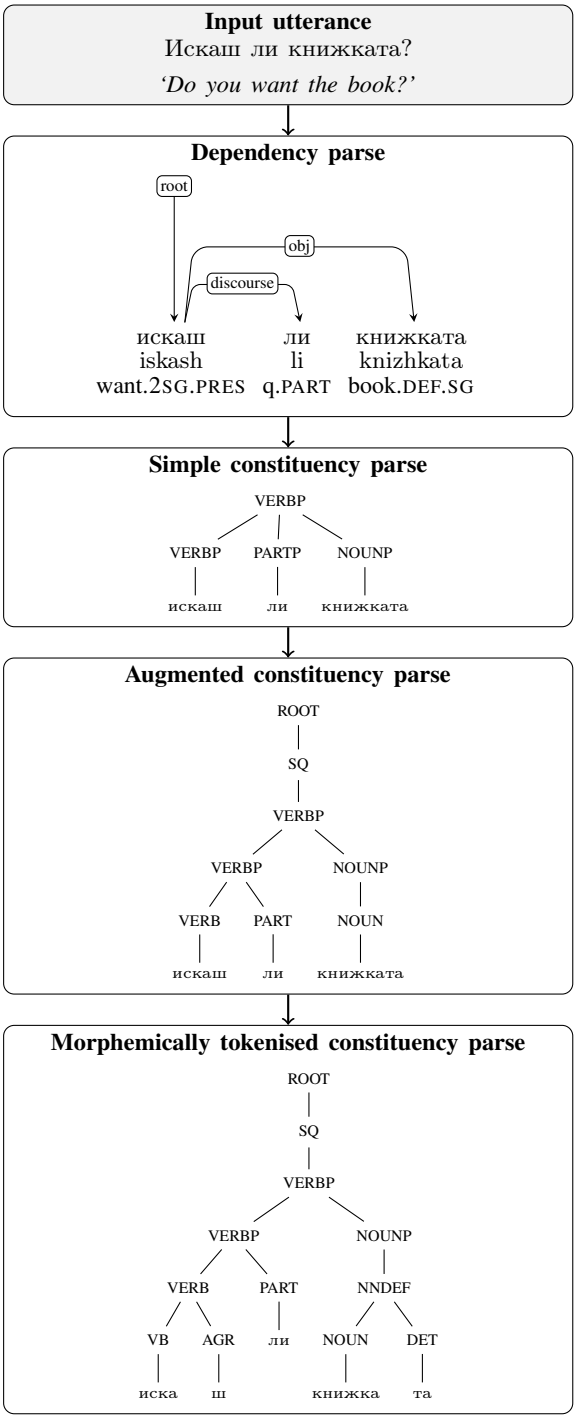

Figure 2: Different types of parses. The augmented and morphemically tokenised constituency parses are the result of the algorithms presented here.

guistically defensible implementation and apply it consistently throughout the corpus.

3.2.1 Stem vs. lemma

When the functional morpheme is identified and segmented, the remaining stem does not always correspond to the dictionary form. In some cases, the stem and the lemma are the same, such as in въпроси → въпрос и *(questions → question* + PL). In other cases, such as правим → прав им *(we do → do.STEM* + 1PL*)* and правят → прав ят *(they do → do.STEM* + 3PL*)*, the remaining stem is not a lemma (the lemma would be the first person form правя). In such cases, for PS-CHILDES-BG we choose to use the stem so that the different conjugations of the verb can still have the same base, as this is important for symbolic downstream applications. However, our morphemic tokeniser also provides functionality to instead use the lemma (which is part of the UD parse), if that is appropriate for another user's application.

3.2.2 Multiple functional morphemes

Sometimes in Bulgarian several functional morphemes can be strung together. Here, тротинетките → тротинетк и те *(the scooters → scooter* + PL + DEF.PL*)*, the functional morpheme for plural и and the functional morpheme for definite marking of a plural nominal те are both present. As the morphemes are separable, for PS-CHILDES-BG we segment them, because this would enable the morphemes to be learned separately in category and grammar induction experiments. However, the morphemic tokeniser includes an option to keep all functional morphemes as one suffix, if appropriate for other users' applications.

3.2.3 Gender

In Bulgarian, gender is one of the grammatical axes of agreement: a noun has a fixed grammatical gender and a nominal modifier of the noun, such as an adjective, needs to agree with its gender. In the regular case segmenting the suffix marking gender is easy as it is an additive suffix added to the masculine form: красив, красива, красиво *(beautiful.*M.SG*, beautiful.*F.SG*, beautiful.*N.SG*)*. However, often the stem of an adjective differs between masculine and neuter and feminine, e.g. малък, малка, малко *(small.*M.SG*, small.*F.SG*, small.*N.SG*)* or странен, странна, странно *(strange.*M.SG*, strange.*F.SG*, strange.*N.SG*)*. As there is neither a common lemma nor a common stem for all three genders in these cases, we decide to not tokenise gender in the current iteration of the system.

3.2.4 Verbal conjugation classes

Bulgarian verbal inflection is organised around conjugation-specific bases and endings. The three

conjugation classes are distinguished by the vowel of the third person singular present tense form: first-conjugation verbs have е, second-conjugation verbs have и, and third-conjugation verbs have an а/я-final base.[2] The morphemic tokeniser therefore does not apply a single uniform suffix rule to all verbs. Instead, it selects a segmentation rule on the basis of the lemma, UD morphological features, surface form, and, where relevant, conjugation class and tense. This allows the tokeniser to distinguish cases where a vowel belongs to the inflectional ending from cases where it belongs to the verbal base, while still keeping the lexical base stable across forms whenever possible.

3.3 Dependency to constituency conversion

Dependency to constituency (dep-to-const) conversion is commonly performed using the method of Collins et al. (1999). Here we use the implementation of Kando et al. (2022) as a starting point: the output of their algorithm is the "Simple constituency parse" in Figure 2. We perform a number of Bulgarian-specific modifications, some of which are described below, which aim to introduce more hierarchical structure through linguistic grounding to yield parses of the calibre of the "Augmented constituency parse" in Figure 2. More parses, both standardly and morphemically tokenised constituency parses are presented in Appendix A.

3.4 Hierarchical restructuring

To every parse we add a ROOT wrapper. We distinguish between pre-terminals, the UPOS tagset (e.g. NOUN) and non-terminals, the UPOS tagset + P (e.g. NOUNP), and where necessary introduce PTB-style non-terminals (e.g. SBAR for subordinate clauses). Discourse particles, vocatives, and interjections attach at the highest possible level to the clause as a whole, instead of being inserted inside the core predicate structure, such as in Figure 3. Within verbal projections, the verb should combine with its object first, such as in Figure 4. However, due to the relatively free word order of Bulgarian, an adjective may intervene in the surface order between verb and object, such as in Figure 5. Due to the restrictions of constituency grammar, the verb needs to form a constituent with the adverb and then with the object through a recursive VERBP rule.

For subordinated clauses introduced by the dependency relations `xcomp`, `ccomp`, and `advcl`, we use SBAR and subordinate them to the verb that is the head of the subordinate clause in the dependency parse, as seen in Figure 6. To distinguish interrogative clauses from non-question clauses, we use the PTB-style label SQ to introduce questions, as seen in Figure 7. Since the dependency parse does not reliably encode interrogative force, we use the presence of the question mark as the necessary condition for assigning SQ. Furthermore, within such clauses, Bulgarian interrogative *k*-words are projected as WH-type phrases according to their POS category, for example WHNP for pronominal and determiner-like forms such as кой (*who*) and WHADVP for adverbial forms such as къде (*where*).

3.5 Morphemically tokenised constituency parses

The morphemically tokenised constituency parses follow the same rules as above and additional rules are introduced to replace word-level terminals with subtrees comprising a word base and functional morphemes. For example, we use the pre-terminal AGR to mark verbal agreement morphology, DET to mark nominal definiteness morphology, and DIV to mark nominal count or plural number. Some examples are presented in Appendix A. We manually verify 15% of the total 2,434 parses in PS-CHILDES-BG.

4 Evaluation

We evaluate the morphemic tokeniser using MorphScore. In the original paper (Table 3; Arnett and Bergen, 2025), the reported accuracy score for Bulgarian is 0.584, whereas our morphemic tokeniser achieves 0.814. MorphScore is not designed specifically for functional morphology and also includes derivational morphology. MorphScore tokenises gender which we do not as explained in subsubsection 3.2.3, and this accounts for 55% of the mismatches with the MorphScore gold tokenisation, and some disagreements around verbal endings, which we tokenise as explained in subsubsection 3.2.4, account for 26.7% of the mismatches. As MorphScore is not primarily aimed at functional morpheme tokenisation, not achieving an accuracy of 1.0 is not necessarily a bad result.

[2] See the overview of Bulgarian verbal conjugation at https://www.slav.uni-sofia.bg/grammar/morf25.html and the overview of Bulgarian tense formation at https://www.slav.uni-sofia.bg/grammar/morf30.html.

Tokenisation →	STANDARD		MORPHEMIC
Metric ↓	Simple	Augm.	Augm.
Branching	2.74	1.51	1.54
Arity≥3	52.2%	0.24%	0.21%
Depth	2.35	4.94	5.18

Table 1: Tree structure evaluation using mean branching, percentage of nodes with arity≥3, and mean tree depth, for the simple dep-to-const conversion (standard tokenisation) and for the augmented dep-to-const conversion (standard and morphemic tokenisation).

Additionally, to demonstrate the quantitative difference between the parses resulting from the Collins algorithm (Collins et al., 1999; Kando et al., 2022) and the ones from our augmented algorithm we quantify several measures of tree flatness in Table 1. Our parses have a decreased branching factor, very few nodes with high arity, and an increased tree depth. As expected, the morphemically tokenised trees are even deeper, due to the way we introduce functional morphemes with subtrees. Overall, our augmented algorithm introduces hierarchy in the parses, making them more linguistically informed.

5 Conclusion

We present PS-CHILDES-BG, a constituency treebank of 2,434 Bulgarian CDS sentences, of which 15% are manually verified. The constituency parses are converted from UD parses, using a linguistically augmented dep-to-const algorithm. We also supply a morphemic tokenisation algorithm for Bulgarian, which can be used independently or in conjunction with the augmented dep-to-const algorithm to yield morphemically tokenised constituency parses for Bulgarian. In future work we will use PS-CHILDES-BG for FLA research via computational methods.

Limitations

A limitation of the resource, as also noted by a reviewer, is that the manual verification of the data is limited, due to how time-consuming manual verification is. We have increased it from 10% to 15% to address the reviewer's comment, but the issue is still worth considering in future iterations. Another limitation of the resource is that it relies on existing dependency parsing models which may not have the desired accuracy. However, the focus of this paper is to demonstrate how we can harness existing resources to create a silver-standard, if not gold-standard, resource which has potential to offer a different insight into computational modelling of acquisition than UD resources. Finally, the current verbal morphemic tokenisation does not label different verbal endings according to their function, such as person, number, tense, aspect, and mood, because of some fusional characteristics of Bulgarian morphology, which make the disentanglement of verbal morphology a problem for future work.

References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Catherine Arnett, Marisa Hudspeth, and Brendan O'Connor. 2025. Evaluating Morphological Alignment of Tokenizers in 70 Languages. In *Proceedings of the ICML 2025 Tokenization Workshop (TokShop)*.

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

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

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

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

Alastair Butler, Susanne Miyata, and Yumiko Kinjo. 2022. The Soyogo Treebank – a parsed corpus of child Japanese. https://soyogo.github.io.

Paula J. Buttery. 2006. Computational models for first language acquisition. Technical Report UCAM-CL-TR-675, University of Cambridge, Computer Laboratory.

Andrew Carnie. 2021. *Syntax*, 4 edition. Introducing Linguistics. Wiley-Blackwell, Hoboken, NJ.

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.

Timothy A. Cartwright and Michael R. Brent. 1997. Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, 63(2):121–170.

Michael Collins, Jan Hajic, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, College Park, Maryland, USA. Association for Computational Linguistics.

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.

Anne Fernald. 1989. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, 60(6):1497.

Maria Teresa Guasti. 2017. *Language acquisition*, 2 edition, chapter 1: Basic Concepts. Bradford Books, Cambridge, MA.

Grigor Iliev, Nadezhda Borisova, Elena Karashtranova, and Dafina Kostadinova. 2015. A publicly available cross-platform lemmatizer for Bulgarian.

Shunsuke Kando, Hiroshi Noji, and Yusuke Miyao. 2022. Multilingual syntax-aware language modeling through dependency tree conversion. In *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 1–10, Dublin, Ireland. Association for Computational Linguistics.

Patricia K. Kuhl. 2006. Is speech learning 'gated' by the social brain? *Developmental Science*, 10(1):110–120.

Elena Lieven. 2010. Input and first language acquisition: Evaluating the role of frequency. *Lingua*, 120(11):2546–2556.

Elena Lieven. 2016. Usage-based approaches to language development: Where do we go from here? *Language and Cognition*, 8(3):346–368.

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. In *Conference on Language Technologies and Digital Humanities (JT-DH-2024)*, Ljubljana, Slovenia. Institute of Contemporary History.

Brian MacWhinney. 1992. The CHILDES project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8(2):217–218.

Mila Marcheva, Theresa Biberauer, and Weiwei Sun. 2025a. Functional category induction with theory-neutral cognitive biases. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Mila Marcheva, Theresa Biberauer, and Weiwei Sun. 2025b. Profiling neural grammar induction on morphemically tokenised child-directed speech. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 47–54, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Mila Marcheva-Nash, Yasena Chantova, Tsvetina Kirilova, Ivelina Pavlova, Tsvetelina Stefanova, Yoana Vasileva, and Weiwei Sun. 2026a. UD-CHILDES-BG: A Dependency Treebank of Bulgarian Child and Child-Directed Speech. In *Proceedings of the 20th Linguistic Annotation Workshop*, San Diego, United States of America. To appear.

Mila Marcheva-Nash, Suchir Salhan, and Weiwei Sun. 2026b. A computational operationalisation of competing maturational theories of syntactic development via statistical grammar induction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 48. To appear.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, page 1–54.

Preslav Nakov. 2003. BulStem: Design and evaluation of inflectional stemmer for Bulgarian. In *Proceedings of the Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece. 1st Balkan Conference in Informatics.

Jan Odijk, Alexis Dimitriadis, Martijn van der Klis, Marjo van Koppen, Meie Otten, and Remco van der Veen. 2018. The AnnCor CHILDES treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Petya Osenova and Kiril Simov. 2017. Recent developments within BulTreeBank. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 129–137, Prague, Czech Republic.

Christopher Parisien, Afsaneh Fazly, and Suzanne Stevenson. 2008. An incremental Bayesian model for learning syntactic categories. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 89–96, Manchester, England. Coling 2008 Organizing Committee.

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*.

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.

Velka Popova. 2023. The emergence of word classes in early Bulgarian language ontogenesis. a pilot corpus study. *Journal of Bulgarian Language*, 70(PRIL):290–308.

Velka Popova and Dimitar Popov. 2020. CHILDES Bulgarian labling corpus.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Kiril Simov and Petya Osenova. 2003. Practical annotation scheme for an HPSG treebank of Bulgarian. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-2003)*, pages 17–24.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Ida Szubert, Omri Abend, Nathan Schneider, Samuel Gibbon, Louis Mahon, Sharon Goldwater, and Mark Steedman. 2024. Cross-linguistically consistent semantic and syntactic annotation of child-directed speech. *Language Resources and Evaluation*, 59(2):727–776.

Luka Terčon and Nikola Ljubešić. 2023. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025. UD-English-CHILDES: A collected resource of gold and silver Universal Dependencies trees for child language interactions. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 52–58, Ljubljana, Slovenia. Association for Computational Linguistics.

A Constituency parse examples

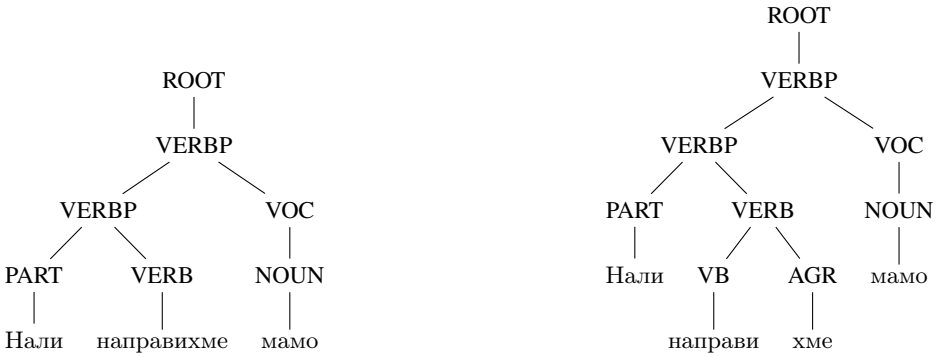

Figure 3: Нали направихме, мамо. ‘Right, we did it, *inverse-vocative-address*.’

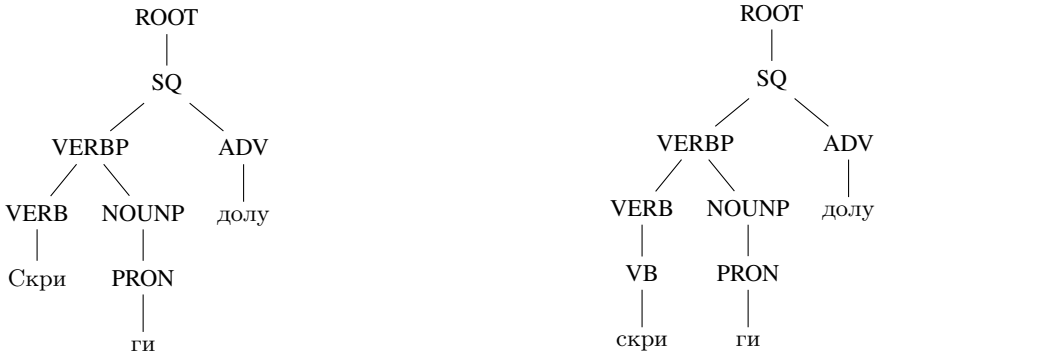

Figure 4: Скри ги долу. ‘He/She/It hid them under.’

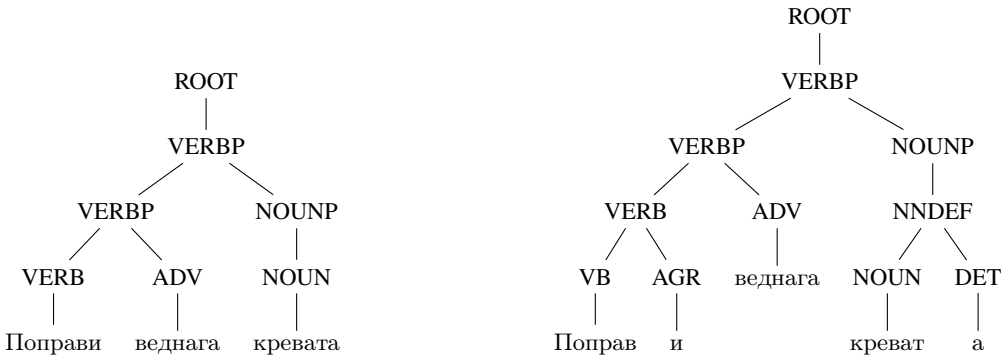

Figure 5: Поправи веднага кревата. ‘Fix the bed immediately.’

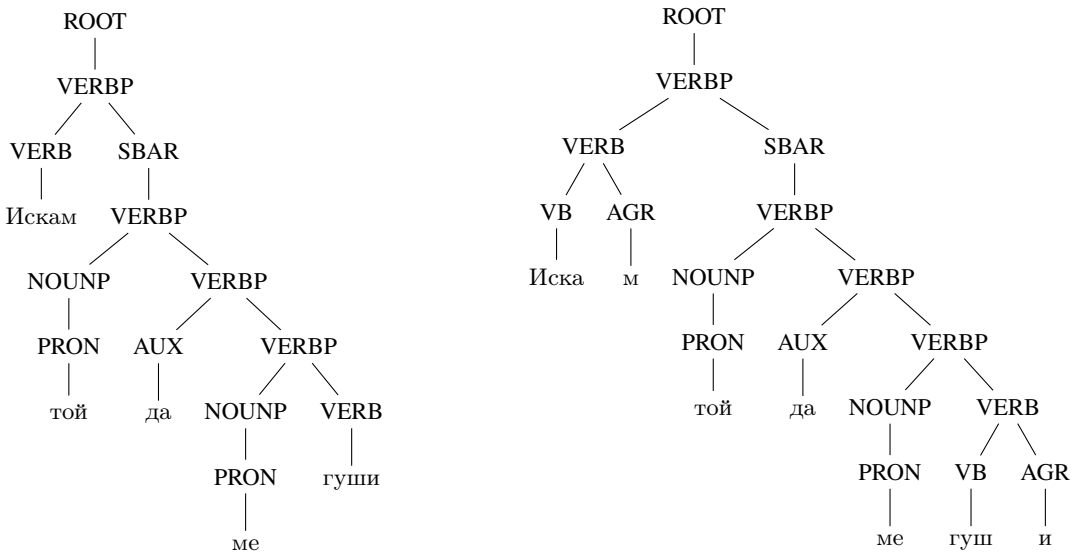

Figure 6: Искам той да ме гуши. ‘I want him to hug me.’

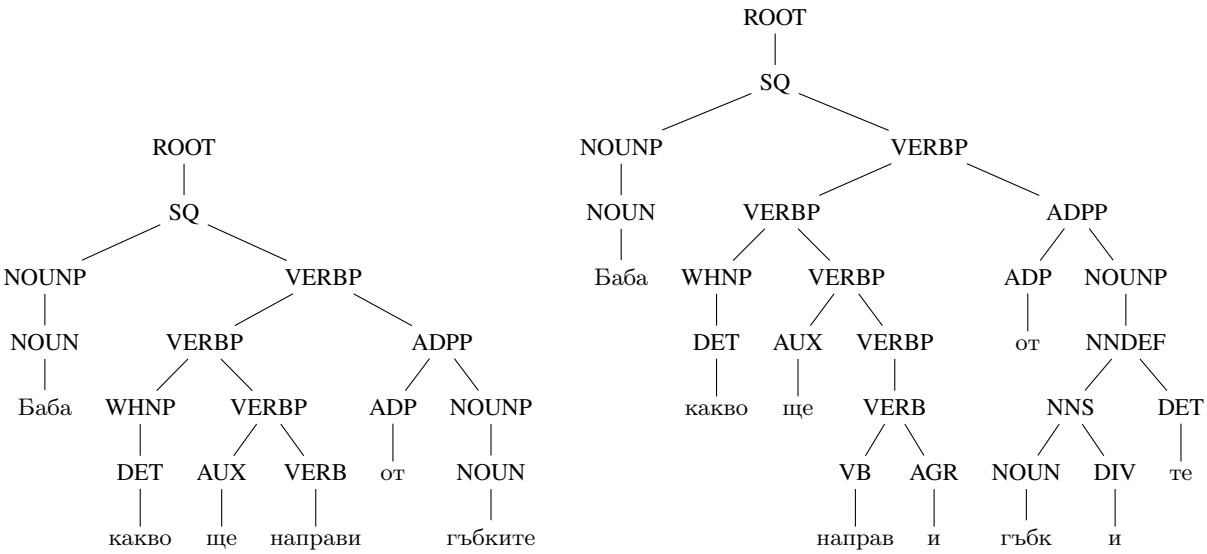

Figure 7: Баба какво ще направи от гъбките? ‘What will grandma make from the mushrooms?’

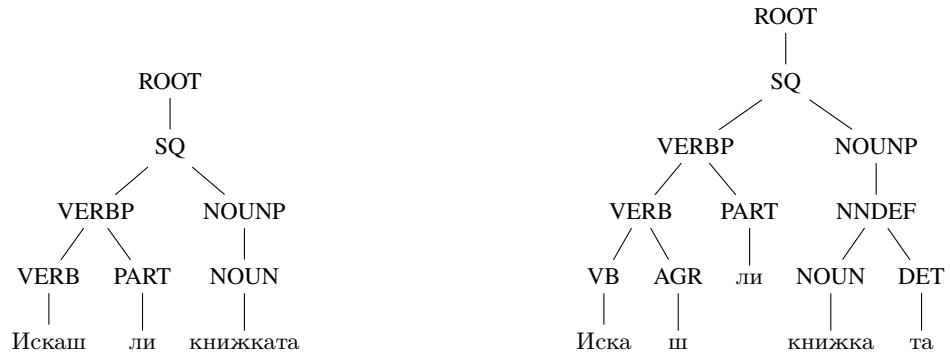

Figure 8: Искаш ли книжката? ‘Do you want the book?’

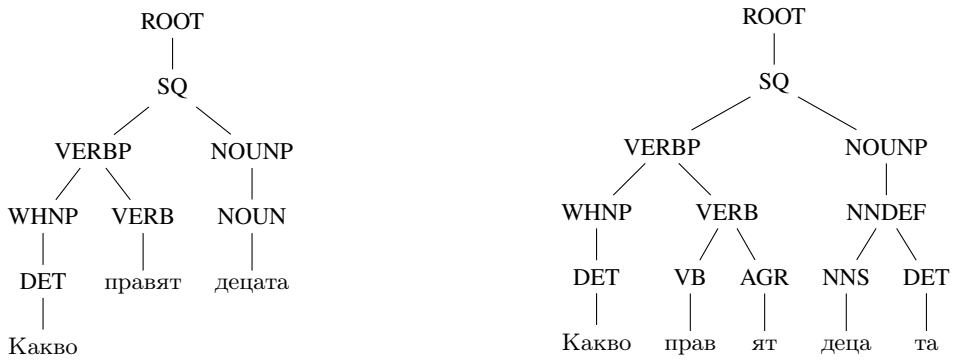

Figure 9: Какво правят децата? ‘What are the children doing?’