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

Mila Marcheva-Nash mmm67@cam.ac.uk
Weiwei Sun ws390@cam.ac.uk
Department of Computer Science & Technology, University of Cambridge, UK

At a glance

- **PS-CHILDES-BG: 2,434** Bulgarian child-directed speech (CDS) sentences represented as morphemically tokenised constituency trees.
- A **Bulgarian morphemic tokeniser** exposes functional inflectional morphology from UD parses while retaining useful lexical bases.
- A linguistically augmented **dependency-to-constituency converter** adds Bulgarian-specific hierarchical phrase structure.
- The components work **independently or together**; combined, they convert a UD parse into a morphemically tokenised constituency parse.
- The tokeniser reaches **0.814 MorphScore accuracy**, and the augmented trees are substantially less flat than the simple baseline.

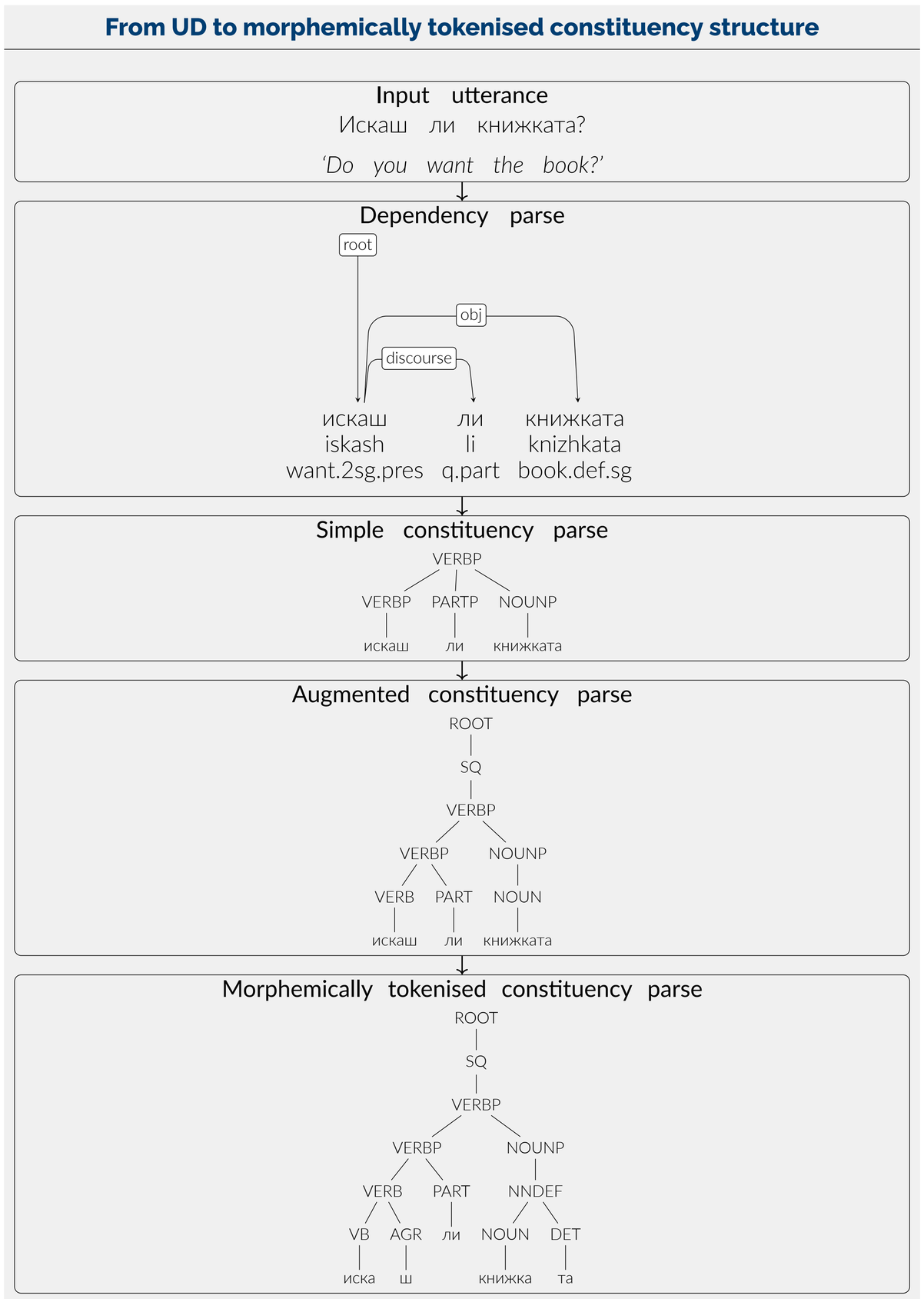
Motivation

- A central first-language acquisition question is how children acquire **functional categories** from linguistic input.
- Standard word tokenisation exposes free functional items but hides **bound functional morphemes**; morphemic tokenisation represents both.
- This distinction matters especially for morphologically rich languages such as Bulgarian.
- CDS is the cognitively relevant input register, while constituency trees directly encode the hierarchy used by phrase-structure models.

Existing resources

- Comparable morphemically tokenised CDS constituency resources are available only for English [9, 5].
- Relevant Bulgarian resources include **BulTreeBank / UD-BTB** [13, 8], **CHILDES-BG** [11], and constituency and dependency parsers such as Berkeley [10], Stanza [12] and CLASSLA-Stanza [4].
- BulStem [7] and Bulgarian lemmatisation resources operate below the word level but do not specifically tokenise functional morphemes.
- MorphScore [1] supplies Bulgarian evaluation data.
- The SIGMORPHON morphemic segmentation shared task excludes Bulgarian.
- PS-CHILDES-BG fills this gap by combining Bulgarian CDS, constituency structure and explicit functional-morpheme tokenisation.

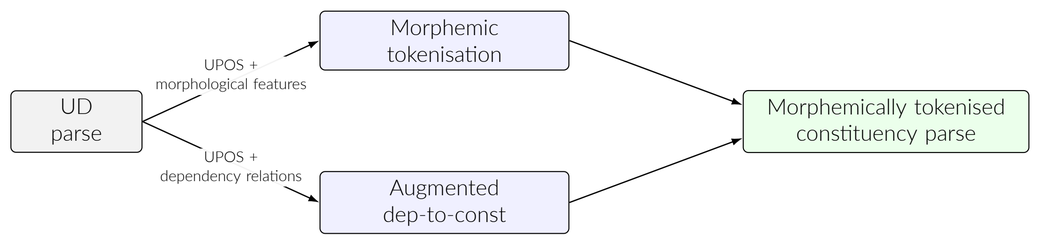
From UD to morphemically tokenised constituency structure



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

Pipeline overview

- The tokeniser uses **UPOS and morphological features**; the converter uses **UPOS and dependency relations**.
- Their outputs combine into a morphemically tokenised constituency parse.



Pipeline for deriving a morphemically tokenised constituency parse from a UD parse.

Data & morphemic tokenisation

- **Data:** input is the manually verified CDS portion of UD-CHILDES-BG [6]: 2,434 projective parses from speech addressed to five children aged 1–3 years. Automatic UD parses may also be used.
- **Consistent analysis:** where multiple segmentations are possible, the tokeniser applies one linguistically defensible analysis throughout the corpus.
- **Stem vs. lemma:** the user of the algorithm can select to retain stem (e.g. правим → прав + им) or lemma (правим → правя + им) as the base lexeme after morphemic tokenisation. The default is stem.
- **Multiple morphemes:** segment independently learnable suffixes, e.g. тротинетките → тротинетк + и + те (stem + pl + def.pl).
- **Gender:** not tokenised in the current system because irregular alternations often prevent a stable stem across masculine, feminine and neuter forms.
- **Verbs:** rules use the lemma, UD features, surface form, conjugation class and tense instead of applying one suffix rule to every verb.

Dependency-to-constituency conversion

- Start with the Collins conversion [2, 3], then add **Bulgarian-specific linguistic restructuring**.
- Add a **ROOT**; distinguish UPOS pre-terminals (NOUN) from phrasal non-terminals (NOUNP); introduce PTB-style labels where needed.
- Attach discourse particles, vocatives and interjections at clause level, and build recursive **VERBP** projections instead of broad flat branching.
- Map `xcomp`, `ccomp` and `advcl` clauses to **SBAR**. Use **SQ** when a question mark is present, since UD does not reliably encode interrogative force; project Bulgarian *k*-words as POS-sensitive WH phrases.
- Replace word terminals with base+morpheme subtrees: **AGR** marks verbal agreement, **DET** definiteness and **DIV** nominal plural/count morphology.

Evaluation

- **Morphemic tokenisation:** MorphScore accuracy is **0.814**, compared with the Bulgarian result of **0.584** from the MorphScore paper [1].
- **Remaining mismatches:** 55% come from gender tokenisation, which this system deliberately omits; 26.7% concern different analyses of verbal endings.
- **Tree structure:** augmentation lowers branching and high-arity nodes while increasing depth, producing substantially more hierarchy.

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

Mean branching, percentage of nodes with arity≥3, and mean tree depth for the simple conversion and the augmented conversion with standard and morphemic tokenisation.

Conclusion

- PS-CHILDES-BG provides **2,434** Bulgarian CDS constituency trees with explicit functional-morpheme structure; **15%** are manually verified. Manual verification remains limited because constituency-tree checking is time-consuming.
- The morphemic tokeniser and augmented converter can be used independently or in a pipeline.

References

[1] Catherine Arnett and Benjamin Bergen. Why do language models perform worse for morphologically complex languages? In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.441/.

[2] Michael Collins, Jan Hajic, Lance Ramshaw, and Christoph Tillmann. 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, June 1999. Association for Computational Linguistics. doi: 10.3115/1034678.1034754. URL https://aclanthology.org/P99-1065/.

[3] Shunsuke Kando, Hiroshi Noji, and Yusuke Miyao. Multilingual syntax-aware language modeling through dependency tree conversion. In Andreas Vlachos, Priyanka Agrawal, André Martins, Gerasimos Lampouras, and Chunchuan Lyu, editors, *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 1–10, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.spnlp-1.1. URL https://aclanthology.org/2022.spnlp-1.1/.

[4] Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 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, 2024. Institute of Contemporary History. doi: 10.5281/zenodo.13936406. URL https://doi.org/10.5281/zenodo.13936406.

[5] Mila Marcheva, Theresa Biberauer, and Weiwei Sun. Profiling neural grammar induction on morphemically tokenised child-directed speech. In Tatsuki Kuribayashi, Giulia Rambelli, Ece Takmaz, Philipp Wicke, Jixing Li, and Byung-Doh Oh, editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 47–54, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-227-5. doi: 10.18653/v1/2025.cmcl-1.7. URL https://aclanthology.org/2025.cmcl-1.7/.

[6] Mila Marcheva-Nash, Yasena Chantova, Tsvetina Kirilova, Ivelina Pavlova, Tsvetelina Stefanova, Yoana Vasileva, and Weiwei Sun. 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, 2026. To appear.

[7] Preslav Nakov. BulStem: Design and evaluation of inflectional stemmer for Bulgarian. In *Proceedings of the Workshop on Balkan Language Resources and Tools*, Thessaloniki, Greece, November 2003. URL https://www.researchgate.net/publication/250443777_Design_and_Evaluation_of_Inflectional_Stemmer_for_{Bulgarian}. 1st Balkan Conference in Informatics.

[8] Petya Osenova and Kiril Simov. Recent developments within BulTreeBank. In Jan Hajič, editor, *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 129–137, Prague, Czech Republic, 2017. URL https://aclanthology.org/W17-7617/.

[9] Lisa S. Pearl and Jon Sprouse. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 2013. URL https://ling.auf.net/lingbuzz/001493.

[10] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In Nicoletta Calzolari, Claire Cardie, and Pierre Isabelle, editors, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, July 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220230. URL https://aclanthology.org/P06-1055/.

[11] Velka Popova and Dimitar Popov. CHILDES Bulgarian LabLing Corpus, 2020. URL https://childes.talkbank.org/access/Slavic/{Bulgarian}/LabLing.html.

[12] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 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*, 2020.

[13] Kiril Simov, Gergana Popova, and Petya Osenova. HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In Paul Rayson Andrew Wilson and Tony McEnery, editors, *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, pages 135–142. Lincom-Europa, 2002.

Code & resource: https://github.com/milamarcheva/UD-CHILDES-BG