UD-CHILDES-BG: a dependency treebank of Bulgarian child and child-directed speech

Mila Marcheva-Nash[1] **Yasena Chantova**[2,3] **Tsvetina Kirilova**[2] **Ivelina Pavlova**[2]
Tsvetelina Stefanova[2] **Yoana Vasileva**[2] **Weiwei Sun**[1]

[1]Department of Computer Science & Technology, University of Cambridge
[2]Faculty of Slavic Studies, Sofia University "St. Kliment Ohridski"
[3]University of Library Studies and Information Technologies, Bulgaria

Abstract

This paper presents (i) UD-CHILDES-BG, a manually corrected Universal Dependencies treebank of Bulgarian child and child-directed speech, (ii) a quantitative and phenomenon-based evaluation of inter-annotator agreement on developmental data, and (iii) a systematic analysis of parser errors in this underrepresented domain. We manually correct 4,338 dependency parses (10% of the CHILDES-BG corpus), of which 14% are double-annotated. Inter-annotator agreement on UAS/LAS is 91.71/86.12 for child-directed speech (CDS) and 88.14/81.40 for child speech (CS). Parser performance on the manually corrected portion is 92.70/85.54 for CDS and 90.97/81.52 for CS, compared to a reported 93.37/90.21 on the test set of adult written language. Our analyses reveal that CDS and CS pose challenges for dependency annotation and parsing, particularly in discourse-related structures, which are less common in adult written language.

1 Introduction

Linguistically annotated corpora are central to supervised NLP, evaluation of LLMs, and theory-building in linguistics. Most syntactic treebanks are based on adult written language, while developmental data remains comparatively underrepresented, despite growing interest in the role of syntactic structure for language model robustness (Güven et al., 2025). Furthermore, first language acquisition (FLA) research relies on child-directed speech (CDS) and child speech (CS) with high quality annotations to make extrapolations about the development of language (Bosch, 2025). Providing linguistically annotated CS and CDS, along with scalable guidelines, for resource-poor languages is thus a useful area of expansion to further FLA and NLP research.

The challenge of parsing CDS and/or CS was first addressed by Sagae et al. (2001), and has since been explored primarily for English (Pearl and Sprouse, 2013; Liu and Prud'hommeaux, 2023; Yang et al., 2025), with only a small number of studies on other languages, including Hebrew (Szubert et al., 2024), Japanese (Butler et al., 2022), and Dutch (Odijk et al., 2018). Universal Dependencies (UD; de Marneffe et al., 2021) provides a cross-linguistically consistent framework for syntactic annotation, which has been applied to over one hundred languages. While treebanks for CS and CDS are scarce, they are mainly within the UD framework. For Slavic and Balkan languages, no manually corrected treebank for developmental data currently exists.

In this paper, we present UD-CHILDES-BG, a treebank of manually corrected UD parses for Bulgarian. UD-CHILDES-BG and the accompanying analysis scripts are available on GitHub.[1] The annotation scheme is supplied in Appendix B. We manually correct 4,338 dependency parses, corresponding to 10% of the CHILDES-BG corpus (Popova and Popov, 2020), with a breakdown of 2,481 CDS and 1,857 CS utterances. Starting from automatically produced UD parses, we perform targeted manual correction of core layers: lemma, UPOS, morphological features, dependency arc, and dependency relation.

We assess annotation reliability through double annotation of 14% of the data: inter-annotator agreement on UAS/LAS is 91.71/86.12 for CDS and 88.14/81.40 for CS. By treating annotation disagreement as linguistically informative, we provide annotation guidance and linguistic background for a selected set of language phenomena, including clitic doubling, vocative case, and reflexive pronoun, which are prominent in CDS and CS, but less frequent in adult written language.

We further evaluate the parser on the manually

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

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 113–129

corrected parses: UAS/LAS is 92.70/85.54 for CDS and 90.97/81.52 for CS, compared to a reported[2] 93.37/90.21 on a test set of adult written language. A per-relation analysis shows that the decreased parser performance on CS and CDS is due to discourse-specific structures frequent in conversational and developmental data, but less common in adult written language on which the parser has been trained.

2 Background

Bulgarian is a South Slavic language that exhibits both core Slavic properties, such as diminutives in CDS (Kempe and Brooks, 2001), and characteristic Balkan Sprachbund features, including an atrophied nominal case system and clitic doubling (Tomić, 2011). Bulgarian is *pro*-drop and displays relatively flexible word order, with agreement morphology on the verb marking person and number. These properties interact with discourse structure and clitic placement, making attachment decisions surrounding the phenomena less straightforward in dependency annotation. In CS and CDS such phenomena are further amplified by fragmentary utterances, vocative forms, and discourse particles.

Several syntactic resources exist for Bulgarian. BulTreeBank (BTB; Simov et al., 2002b,a; Simov and Osenova, 2003; Simov et al., 2004; Osenova and Simov, 2004; Simov et al., 2005) is a constituency treebank based on head-driven phrase structure grammar (HPSG; Pollard and Sag, 1994). UD-BTB (Osenova and Simov, 2017) is the manually corrected dependency resource based on BTB. The Bulgarian CHILDES corpus (CHILDES-BG) provides longitudinal developmental data but does not come with syntactic annotation (Popova and Popov, 2020). Parsing tools for Bulgarian include a Berkeley constituency parser trained on BTB (Petrov et al., 2006), as well as dependency parsers such as Stanza (Qi et al., 2020) and CLASSLA-Stanza (Terčon and Ljubešić, 2023) trained on UD-BTB. With regards to resources focusing on child Bulgarian, the existing literature covers a version of the MacArthur–Bates Communicative Development Inventory for Bulgarian (Andonova, 2015), as well as studies on the development of grammar in Bulgarian (Popova and Filipov, 2022; Popova, 2023).

UD resources for CS and CDS remain limited cross-linguistically, starting with the foundational dependency treebank for only CDS by Sagae et al. (2001, 2010). More recently, dependency treebanks covering both CS and CDS have been developed for English (Liu and Prud'hommeaux, 2023; Yang et al., 2025), Japanese (Butler et al., 2022), and Hebrew (Gretz et al., 2013; Szubert et al., 2024), with some cross-linguistic work on English–Hebrew corpora (Szubert et al., 2024). Constituency resources are even scarcer, including the English CHILDES-TB (Pearl and Sprouse, 2013) only for CDS and the Dutch AnnCor Treebank (Odijk et al., 2018) for both CS and CDS. Finally, there are efforts for providing automatic UD annotation to all available CHILDES transcripts (MacWhinney, 2012; Liu, 2024) via `batchalign`.[3] However, at the time of release of `batchalign`, the Bulgarian CHILDES corpus was published only with Latin transcription, and the UD parsing tools require Cyrillic input. The authors of the Bulgarian CHILDES corpus have since uploaded a Cyrillic transcription, and we independently transliterated the utterances for this project. Still, no manually corrected treebank currently exists for Bulgarian CS or CDS, or any other Balkan or Slavic language.

We adopt the UD framework for three reasons. First, UD provides cross-linguistically consistent guidelines that enable direct comparison with existing CS and CDS resources. Second, the primary constituency resource for Bulgarian, BTB, employs relatively flat phrase-structure representations (Osenova and Simov, 2004), making UD's dependency representation comparable in structural granularity while avoiding commitments to language-specific phrase-structural conventions. Third, UD annotation is directly compatible with widely used parsing toolkits facilitating error analysis and automatic pre-annotation. In choosing UD, we prioritise reproducibility and alignment with existing developmental treebanks.

Beyond resource creation, recent work in annotation research has emphasized the importance of distinguishing between annotator disagreement and objectively incorrect annotation (Klie et al., 2023; Weber-Genzel et al., 2024). However, systematic analyses of annotation error in developmental and non-canonical spoken data remain rare. Our work contributes to this line of research by examining parser error and annotation contention specifically in Bulgarian CS and CDS, a typolo-

[2] https://stanfordnlp.github.io/stanza/performance.html

[3] https://github.com/TalkBank/batchalign2

gically and discourse-rich setting that challenges standard UD assumptions.

3 Data and procedure

3.1 Source Data

We use the longitudinal section of the CHILDES-BG corpus (Popova and Popov, 2020), which follows five children between the ages of one and three years and covers both the child-directed speech and the first productions of the target children. In total there are 45,000 utterances, but after removing utterances with empty transcription, 43,915 remain, see Table 1 for a breakdown by child.

	CS		CDS		
Child	M	T	M	T	%
ALE	379	3682	430	4144	10.3
BOG	97	961	148	1465	10.1
ELI	121	1208	706	6825	10.3
SIM	924	9540	805	8347	9.7
TEF	336	3586	392	4157	9.4
Total	1857	18977	2481	24938	9.9

Table 1: Breakdown of valid (non-empty) sentences in the CHILDES-BG corpus by target child. Columns show the number of manually corrected sentences (M) and the total number of valid sentences (T) for child speech (CS) and child-directed speech (CDS). % denotes what fraction of total (CS+CDS) sentences for a given child were manually corrected.

A stratified sampling strategy was adopted across children and speech types, CS and CDS. Sampling was stratified by child and by age, and sentences were drawn from across the full set of available utterances, in order to ensure broad coverage of speakers and developmental stages. The primary annotator corrected slices across all children and both registers, while five additional annotators were each assigned specific non-overlapping subsets (see Appendix A for further detail). To ensure quality control, portions of the data were deliberately double-annotated. The exact numbers of sentences annotated for each target child are presented in Table 1. In total, 1,857 CS utterances and 2,481 CDS utterances were manually corrected, corresponding to 10% of all valid CS and CDS utterances in the corpus. Double annotation was performed on 14% of these.

3.2 Preprocessing of CHILDES annotation

CHILDES corpora are transcribed and annotated following the CHAT guidelines[4] (MacWhinney, 1992). The CHAT guidelines cover child-language-specific cases such as phonological variation, e.g. *popo [: hippopotamus]*, where the bracketed expression is the standard form of the child form preceding it, or special coding for unintelligible words, which is standardly *xxx*. Sentences consisting entirely of *xxx* are dropped and do not appear in the final dataset. However, when *xxx* is part of a sentence, we retain such sentences and annotate them following Odijk et al. (2018). Because we are using automatic parses for the initial preprocessing of the utterances, we perform normalisation where phonological variation annotation is provided: we replace the child form with the adult form in order to allow for better automatic parsing.

The CHILDES annotation is not always consistently applied, which poses a challenge to automatic cleaning of the annotation. The main focus of the manual data normalisation is the phonological variation annotation. Sometimes it is applied consistently as in Figure 1 (a), where automatic normalisation is sufficient.

Дугата [: другата] ана [: страна]
dugata [: drugata] ana [: strana]
other.DEF side
'the other side'
Normalised: drugata strana

(a) Consistent use of CHAT annotation.

икам [: искам] гая [: да играя]
ikam [: iskam] gaja [: da igraja]
want.1SG play.1SG
'I want to play'
Problem: inserted *da*

(b) Inconsistent use of CHAT annotation.

Figure 1: Use of CHILDES correction brackets: (a) consistent phonological normalisation; (b) inconsistent annotation introducing additional syntactic material

However, in other cases the bracketed notation is only applied to correct one of several repeated forms, or maps one surface form to several adult forms, as is the case in Figure 1 (b), where the notation is used inappropriately to introduce the functional token *da*. Although *da* is implied in the meaning, it should not be added to the child utterance as this utterance is representative of a specific

[4] https://talkbank.org/0info/manuals/CHAT.html

stage in FLA development, where function words are omitted. During the manual normalisation of the data we ensure that child utterances are corrected where possible with the word forms to facilitate automatic parsing, but we do not allow for extra syntactic material to be added.

3.3 Annotation procedure

We use a custom-made project on the INCEpTION platform (Klie et al., 2018). INCEpTION has built-in support for data in CoNLL format, which is the standard UD format. The annotation correction includes: correction of lemma, UPOS, morphological features, dependency arcs, and dependency relations. The UD-BTB also has a very fine-grained POS tag, XPOS, with encoded morphological information (Simov et al., 2004). For UD-CHILDES-BG, the XPOS tag is not corrected or retained, as the UPOS and morphological features overlap with it. The valid (non-empty) sentences from Table 1 were all parsed using the Stanza parser (Qi et al., 2020) trained on UD-BTB. The parses of the sentences selected via stratified sampling were uploaded to INCEpTION for annotation.

The data is annotated by six annotators, who are all native speakers of Bulgarian. A1 is the main annotator who created the annotation schemes, and annotated a total of 2,500 of the 4338 annotated sentences. The annotation agreement scores are calculated between A1 and the other annotators. A1 has a background in computational linguistics including a completed undergraduate course on *Formal Models of Language* and graduate courses *Natural Language Processing (NLP)*, *Introduction to Computational Semantics*, *Introduction to Natural Language Syntax and Parsing*, as well as teaching *Formal Models of Language* to undergraduate students. A2-A6 all have bachelor degrees in various philologies, and are currently enrolled in a Computational Linguistics Master's programme. All annotators were previously familiar with the UD framework, and received a 2-hour training and overview of UD examples specific for this project, additional to previous experience.

The annotators first annotated 5% of the sentences, 1,204 CDS and 907 CS, with 238 CDS and 185 CS double annotated (all by A1, and non-overlapping subsets by the other annotators). Afterwards the annotations were analysed to identify the most common sources of error in annotation, and a 2-hour discussion was held among the annotators to decide on a unified approach for the edge cases. Following the conclusions of the discussion, the annotation scheme was refined, and the annotators revised their initial annotations and annotated a further 5% of the data following the unified principles. The precise breakdown of sentences by annotator can be found in Appendix A.

4 Annotator agreement

Below we present the annotator agreement with standard quantitative metrics and further discuss specific phenomena worth mentioning due to their pertinence in CS and CDS. Sentences that deviate from written adult language, due to exhibiting features of spoken or developmental language, are not fully captured by the existing gold standard, UD-BTB. We provide further detail on some prominent examples below, while a more comprehensive list of challenging phenomena is covered in Appendix B.

4.1 Quantitative evaluation

We use several metrics to quantify inter-annotator agreement: UPOS and lemma accuracy measure exact agreement between annotators on universal part-of-speech tags and lemma assignment respectively; Cohen's κ measures agreement beyond chance for UPOS annotation; Unlabelled Attachment Score (UAS) measures agreement on dependency head attachment regardless of relation label; and Labelled Attachment Score (LAS) measures agreement on both dependency head attachment and dependency relation labels. In the context of inter-annotator agreement, "accuracy" refers to pairwise agreement between annotators and does not imply comparison against an external gold-standard annotation.

Table 2 displays inter-annotator agreement on the double-annotated subsets of CDS and CS. Overall agreement is high across domains. For CDS, UPOS accuracy reaches 92.48 ($\kappa = .9143$), while CS yields slightly lower but comparable values 90.57 ($\kappa = .8899$). Lemma agreement is lower in both domains (86.50 for CDS; 79.25 for CS), reflecting increased ambiguity in lemmatisation of non-canonical and morphologically reduced forms. Syntactic attachment agreement remains strong in both domains, with UAS/LAS of 91.71/86.12 for CDS and 88.14/81.40 for CS. The similarity of LAS across CDS and CS suggests that, despite structural irregularities in child speech, annotators converge on dependency labels at comparable rates

once head attachment is established. An additional explanation of the convergence of UAS/LAS scores for CS and CDS is that CS is naturally comprised of shorter utterances, hence, there are fewer arcs and relations the annotators can disagree on.

	$UPOS_a$	$UPOS_\kappa$	Lemma	UAS	LAS
CDS	92.48	0.9143	86.50	91.71	86.12
CS	90.57	0.8899	79.25	88.14	81.40

Table 2: Combined inter-annotator agreement: UPOS (accuracy and Cohen's κ); Lemma (accuracy); Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS).

Table 3 breaks down agreement by dependency relation. IAA (inter-annotator agreement) is calculated using the F1 formula. Structural relations such as *case*, *root*, and *aux* exhibit consistently high IAA in both domains, indicating stable annotation of core grammatical structure. In contrast, discourse-sensitive relations show greater variability. In CDS, *discourse* (IAA=84.82) and *iobj* (IAA=80.43) show reduced agreement, reflecting ambiguity in clitic doubling and particle attachment. In CS, agreement for *discourse* drops substantially (IAA=55.56), alongside lower scores for *conj* and *obj*, suggesting that developmental constructions and fragmentary utterances increase annotation difficulty.

These patterns indicate that disagreement is concentrated not in core syntactic relations but in constructions that interact with discourse structure, clitic systems, and child-specific production phenomena. Rather than treating such disagreement as annotation noise, we interpret it as evidence of structural tension between standard UD guidelines and non-canonical spoken Bulgarian.

4.2 Phenomenon-based evaluation

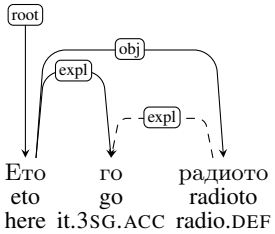
Below we discuss annotators' interpretations of several phenomena illustrative of CDS and CS in Bulgarian, using linguistic background to justify the final annotation decision. For all cases we follow the logic of annotation of UD-BTB (Osenova and Simov, 2017). In the interest of space we limit the discussion to the annotation of dependency arcs and relations; see the annotation scheme in Appendix B for comments on lemma, UPOS, and morphological features.

CDS		CS	
Relation	IAA	Relation	IAA
root	93.66	root	95.88
obj	91.03	obj	70.83
advmod	93.77	conj	62.22
nsubj	86.19	advmod	85.71
discourse	84.82	discourse	55.56
case	99.42	nsubj	75.00
aux	91.76	case	94.74
cop	97.14	nmod	77.78
vocative	90.38	aux	94.12
iobj	80.43	det	76.92
expl	78.95	vocative	50.00
cc	90.41	cc	100.00
conj	82.35	amod	85.71
obl	75.86	expl	33.33

Table 3: IAA (F1 calculations used as an inter-annotator agreement metric) for selected high-frequency relations, sorted by frequency separately for CDS and for CS; punctuation excluded.

4.2.1 Clitic doubling

Clitic doubling is exhibited by several languages in the Balkan Sprachbund (Tomić, 2011). It is the phenomenon where a clitic pronoun and a full nominal argument co-occur, which makes it a contentious case for UD annotation, as there are two forms with the same function (e.g. `obj`). There are limited examples of clitic doubling in UD-BTB, while the phenomenon is a lot more common in spoken Bulgarian, and therefore in CHILDES-BG. In UD-BTB the annotation of clitic doubling is resolved with the full nominal argument tagged with relation `obj` and the clitic pronominal with relation `expl`, sharing the same head.



'Here it is, the radio'

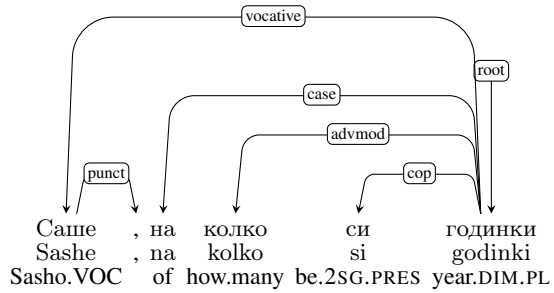
Figure 2: Clitic doubling: the adopted annotation approach. The dashed arc shows an annotation which was decided against.

The guidance all annotators agreed on is in com-

pliance[5] with UD-BTB, and is as follows: the clitic should be attached to the root of the verb (or clause), i.e. to the same head as the nominal argument and be tagged with relation **expl**, see Figure 2. An alternative analysis was proposed (dashed line in Figure 2) where the clitic would have as head the nominal argument itself, because the clitic refers to the nominal argument. This suggestion was ultimately rejected as the nominal argument and the clitic do not always form a constituent.

4.2.2 Vocative case

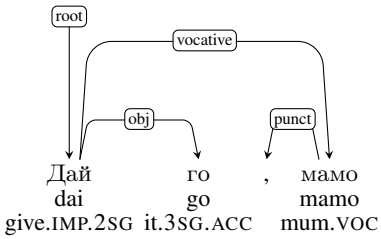
The case system in Bulgarian has been replaced mostly with prepositions, however the vocative case is still active. The vocative case is used to address someone directly, which makes it especially common in spoken language, and less so in written form, as it requires direct speech between interlocutors. UD has a relation `vocative` especially for this case and this is what we use for annotation (see Figure 3). Sometimes the vocative noun also has the semantic role of a subject. Following a discussion of whether the vocative noun should be introduced via `nsubj` or `vocative` in such cases, it was agreed upon that `vocative` has precedence. This is justified, as Bulgarian is *pro*-drop, so the more likely role of the vocative noun is the vocative role.



'Sasho, how old are you?'

Figure 3: Vocative case.

Special case: address inversion is a case of nominal address, where the speaker (addresser) uses their role in the dyad to address the addressee (Beyrer and Kostov, 1978; Braun, 1988; Pavlova, 2015). Most commonly in CHILDES-BG this is exhibited with the vocative of mum: мамо (*mamo* mum.VOC), however it is also possible with other kinship terms and names. In such cases, the consensus among the annotators is to use the `vocative` relation (see Figure 4).



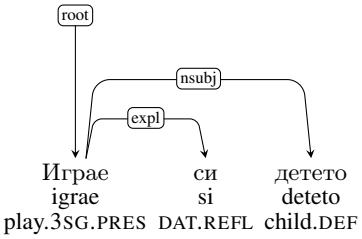
'Give it [to me], child.'

Figure 4: Address inversion.

4.2.3 Reflexive pronoun се/си (se/si refl)

In Bulgarian the form се/си (se/si refl) can manifest as three types of reflexive pronoun:[6] dative reflexive, accusative reflexive, and possessive reflexive. Below are examples of these three cases and the appropriate use of the UD relations to annotate each case. Refer to Penčev (1996) for a detailed grammatical account of these forms and to Slavcheva (2003) for the implications of these forms on morphosyntactic annotation.

Dative reflexive clitic When си (*si* refl.DAT) is a reflexive dative clitic (dativus ethicus), it is introduced with `expl` relation and headed by the verb it relates to (see Figure 5). A dativus ethicus is an optional use of the clitic which contributes emotional context (in this context, comfort or indulgence).



'The child is playing.'

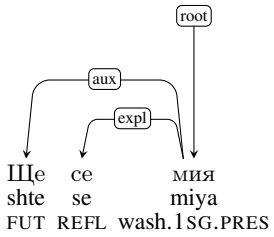
Figure 5: Dative reflexive clitic.

Accusative reflexive clitic When си (*si* refl) is an accusative reflexive clitic (true reflexive), it is introduced with `expl` relation and headed by the verb it is part of (see Figure 6). The role of the accusative reflexive clitic is to refer to the object, when it is identical to the subject. Verbs which permit or require the use of a true reflexive have that information encoded in their lemmas (not pictured). While the accusative reflexive refers to the semantic

[5] One reviewer suggested that the `dislocated` relation might be appropriate to use here, however, it is not used at all for Bulgarian. UD-BTB has existing albeit few examples of clitic doubling, hence we stick with the convention.

[6] Additionally, it is also the 2SG.PRES form of the verb 'be' as in Figure 3.

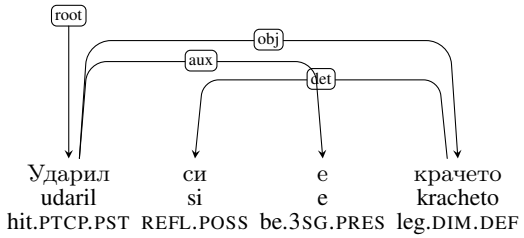
object of the sentence predicate, the convention of UD-BTB is to use the `expl` relation rather than `obj`. Using `expl` for reflexive pronouns is an established UD practice for Slavic languages.[7]



I will wash (myself).

Figure 6: Accusative reflexive clitic.

Short form of possessive pronoun If си (*si*) can be replaced with a full form possessive pronoun, its relation is `det` with head the nominal which it specifies (see Figure 7), following the case of non-contracted possessive forms.[8] These cases can be ambiguous as the use of си (*si*) can resemble a dativus ethicus. The annotator agreement is to deterministically apply the rule above when annotating to ensure consistency. Note that this may result in a non-projective parse.



'He has hit his leg.'

Figure 7: Possessive reflexive clitic.

5 Evaluation of the automatic UD parses of CS and CDS

We evaluate the performance of the Stanza parser (Qi et al., 2020) on the manually corrected portion of UD-CHILDES-BG and present key metrics in Table 4. Additionally, we provide existing benchmarks for Bulgarian UD parsing, all based on the test set of UD-BTB: the state-of-the-art Bulgarian UD parser results are from Hromei et al. (2024) and the Stanza results are from the StanfordNLP website.[9] For further review of UD parsers for Bulgarian see Atanasov (2024). As the existing Stanza parser is trained on adult written Bulgarian, UD-BTB (Osenova and Simov, 2017), we expect that it will underperform on relations more prominent in developmental Bulgarian.

System	UPOS	Lemma	UAS	LAS
Benchmark results on UD-BTB from the literature				
U-DepPLLaMA	–	–	96.37	93.77
UDPipe 2.0++	–	–	95.34	92.62
Stanza	98.68	97.29	93.37	90.21
Evaluation of Stanza on Bulgarian CS and CDS				
CDS	92.26	89.52	92.70	85.54
CS	85.60	83.54	90.97	81.52
Overall	89.93	87.43	92.10	84.14

Table 4: Dependency parsing performance for Bulgarian. Benchmark scores are based on training on the UD-BTB: UDPipe 2.0++ and U-DepPLLaMA are from Hromei et al. (2024) and the Stanza scores are from the StanfordNLP website. In this work we evaluate the Stanza parser against manually annotated Bulgarian CDS and CS.

A per-relation breakdown of the Stanza parser is reported in Table 5. As expected, the parser performs worse on child speech for the majority of the relations. Core syntactic relations such as `case`, `aux`, and `cop` achieve very reliable F1. Relations more common in spoken language such as `discourse` and `vocative` are relatively frequent in the corpus (5.9% and 3.3% of all dependencies respectively), yet they exhibit the lowest parsing accuracy, which highlights that CS and CDS exhibit phenomena less prominent in written adult language, of which UD-BTB is comprised.

5.1 Inability to handle vocative case

Although the use of the vocative case is a relatively frequent phenomenon across both CS and CDS (3.3% of all dependencies), the Stanza parser has lowest UAS and LAS on the `vocative` relation. A case study based on the most common vocative in CHILDES-BG, мамо (*mamo* mum.VOC), shows that only 7/1,848 of its occurrences in the CDS part of CHILDES-BG are correctly labelled with relation `vocative`. The rest of the time мамо (*mamo* mum.VOC) is tagged as an interjection (INTJ) and introduced with relation `discourse` instead. Most surprisingly, 598 times мамо (*mamo*

[7] https://universaldependencies.org/u/dep/expl.html

[8] https://universaldependencies.org/bg/dep/det.html

[9] https://stanfordnlp.github.io/stanza/performance.html

Relation	% of total	CDS F1	CS F1	Overall F1
case	4.03%	99.50	94.01	97.89
aux	5.16%	97.20	92.15	96.19
root	30.22%	94.07	94.33	94.19
cop	3.32%	95.24	94.79	95.14
cc	2.18%	94.62	85.29	91.78
advmod	7.34%	95.03	87.85	93.10
Stanza UD-BTB mean LAS=90.21				
obj	9.68%	89.07	80.04	86.22
amod	1.18%	89.12	86.79	88.07
ccomp	1.98%	81.48	73.76	79.67
nsubj	7.20%	85.83	78.25	83.47
det	1.91%	85.71	83.24	84.87
conj	5.45%	76.90	83.52	81.19
expl	2.16%	78.73	76.47	78.10
iobj	2.38%	75.89	73.08	75.23
discourse	6.08%	78.80	54.83	73.02
xcomp	0.71%	71.95	39.02	65.37
obl	1.86%	65.02	55.17	61.68
advcl	0.67%	69.57	44.83	63.01
nmod	1.15%	64.42	49.16	57.36
vocative	3.16%	28.09	24.88	26.93

Table 5: Dependency relation performance (F1) for Bulgarian child-directed speech (CDS), child speech (CS), and the combined dataset. Relations are grouped into high- and low-performing categories based on overall F1. Percentages indicate the proportion of all gold dependency relations across the CS and CDS annotated dataset.

mum.VOC), which is not a homonymic form, is mistakenly labelled as a verb with lemma мам-(се) (*mam-(se)*), which cannot be found in dictionaries for Bulgarian or in the UD-BTB corpus. Interpreting мамо (*mamo* mum.VOC) as a verb is ungrounded linguistically.

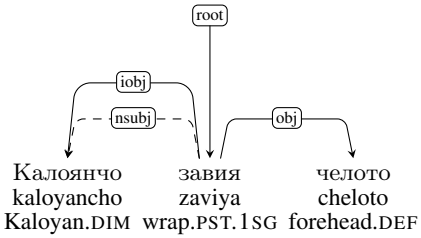
5.2 Reflexive pronoun си (si refl)

The clitic си (*si* refl) poses a challenge to the parser as it can manifest as various types of reflexive pronouns. The relevant dependency relations to си (*si* refl) are `expl` used for the accusative and dative reflexive pronoun, and `det` used for the short form of the possessive reflexive (see subsubsection 4.2.3).

In order to correctly establish the dependency relation, the parser first needs to predict the morphological information of си (*si*). Although the parser often successfully distinguishes between the accusative/dative and the possessive (which can be discerned by the correct assignment of the lemma as the long form possessive pronoun), in the possessive case it does not correctly attach it to the object with the `det` relation, despite having the capability to create non-projective parses (see Figure 7).

5.3 Child speech

As expected, the parser performs worse on CS than on CDS (see Table 4 and Table 5). CS may contain utterances which are ungrammatical to adult speakers, due to atypical word order and/or omitted function words such as case markers, which the parser has no way of recovering. The below CS sentence, Figure 8, exhibits a non-standard word order as well as an omitted function word. Although the typical order in Bulgarian is SVO, relatively free word order is allowed and subject drop is allowed. The order in Figure 8 is IO (indirect object) - V (verb with implied subject) - O (object), which is not grammatical for adults, but it is still recoverable. The parser wrongly assigns the verb to be in third person, and due to the positioning of the indirect object at what would typically be the overt subject position, wrongly assigns the indirect object as a subject. The example is made more challenging as there is an omitted case marker на (*na* 'on'), which the annotators can recover, but the parser, which relies purely on the surface form, cannot.



'I am (going) to wrap the forehead of Kaloyan.'

Figure 8: Child speech with non-standard word order. Incorrect relation assignment by the parser on child speech, `nsubj`, is illustrated with a dashed arc. The correct relation is `iobj`, as assigned by an annotator.

6 Conversion of dependency to constituency parses

Dependency parses can be converted to constituency parses following the Collins method (Collins et al., 1999), if constituency parses are required or preferred. The dependency to constituency conversion based on the Collins method does not provide support for non-projective dependency parses. Non-projective parses comprise 1.45% of the manually corrected sentences. We tested the conversion algorithm from Kando et al. (2022), which is available on GitHub,[10] and augmented it to

[10] https://github.com/gifdog97/dep-to-const

ensure the constituency grammar was more linguistically meaningful. Specifically, we added a distinction between pre-terminals and non-terminals and added a start symbol `ROOT` for all utterances. This facilitates the use of the resulting constituency parses with common NLP libraries such as `NLTK`. The automatically converted constituency parses will be made available alongside the publication of the rest of the treebank and the augmentation of the conversion algorithm will be made available on GitHub upon publication.

7 Conclusion

This paper presents (i) UD-CHILDES-BG, a manually corrected Universal Dependencies (UD) treebank of Bulgarian child and child-directed speech, (ii) a quantitative evaluation of inter-annotator agreement on developmental data, and (iii) a systematic analysis of parser errors on developmental Bulgarian. UD-CHILDES-BG is comprised of 10% manually corrected parses, of which 14% are independently double-annotated. We find that parsing errors arise mainly from properties of Bulgarian discourse and child-specific constructions. The resulting resource constitutes the first manually corrected CS and CDS treebank for a Balkan or Slavic language and offers a methodological account of annotation correction and error analysis in linguistically challenging data.

Limitations

The data from CHILDES-BG contains some inconsistencies: words are not always canonically transcribed and the CHILDES CHAT annotation conventions are not always followed consistently. For the manually corrected part of UD-CHILDES-BG, after automatic preprocessing of the CHAT annotations, the utterances were further manually normalised. INCEpTION is a suitable tool for this project as it requires no installation and can be accessed online by all annotators. However, it has limited support for the morphological features layer, which could have negatively affected the morphological feature annotation.

An additional limitation of this project and a direction for future work is the retraining of Bulgarian UD parsers using the manually corrected portion of UD-CHILDES-BG as supplementary training data. Given the performance decrease observed on CS and CDS relative to adult written Bulgarian, fine-tuning on developmental and spoken language may improve parsing accuracy for discourse-related and child-specific constructions. Extending the manually corrected portion of the corpus beyond the current 10% sample would also support further investigation of parser adaptation at larger scale.

Acknowledgments

We would like to thank Professor Svetla Koeva for connecting the annotators with each other.

References

Elena Andonova. 2015. Parental report evidence for toddlers' grammar and vocabulary in bulgarian. *First Language*, 35(2):126–136.

Atanas Atanasov. 2024. Dependency parser for Bulgarian. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 98–105, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

A. Beyrer and K. Kostov. 1978. "Umgekehrte Anrede" im Bulgarischen und Rumänischen? ("Address inversion" in Bulgarian and Rumanian?). *Balkansko Ezikoznanie*, 21(4):41–53.

Núria Bosch. 2025. Categorial granularity in syntactic acquisition: A multilingual corpus study on the left periphery. *Glossa*, 10.

Friederike Braun. 1988. *Terms of Address: Problems of Patterns and Usage in Various Languages and Cultures*, volume 50 of *Contributions to the Sociology of Language*. Mouton de Gruyter, Berlin and New York and Amsterdam.

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

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.

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

Shai Gretz, Alon Itai, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2013. Parsing hebrew childes transcripts. *Language Resources and Evaluation*, 49(1):107–145.

Arzu Burcu Güven, Anna Rogers, and Rob Van Der Goot. 2025. Do syntactic categories help in developmentally motivated curriculum learning for language

models? In *Proceedings of the First BabyLM Workshop*, pages 288–300, Suzhou, China. Association for Computational Linguistics.

Claudiu Daniel Hromei, Danilo Croce, and Roberto Basili. 2024. U-DepPLLaMA: Universal dependency parsing via auto-regressive large language models. *Italian Journal of Computational Linguistics*, 10(1).

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.

Vera Kempe and Patricia J. Brooks. 2001. The role of diminutives in the acquisition of russian gender: Can elements of child-directed speech aid in learning morphology? *Language Learning*, 51(2):221–256.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198.

Houjun Liu. 2024. Morphosyntactic analysis for CHILDES. *Language Development Research: An Open-Science Journal*, 4.

Zoey Liu and Emily Prud'hommeaux. 2023. Data-driven parsing evaluation for child-parent interactions. *Transactions of the Association for Computational Linguistics*, 11:1734–1753.

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

Brian MacWhinney. 2012. Morphosyntactic analysis of the CHILDES and TalkBank corpora. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2375–2380, Istanbul, Turkey. European Language Resources Association (ELRA).

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. 2004. BTB-TR05: BulTreeBank Stylebook. Technical report, Bulgarian Academy of Sciences.

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.

Neda Pavlova. 2015. "Reverse addresses" in Bulgarian speech – between kinship appellatives and discursive markers. *Balkanistic Forum*, 21(1).

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*.

Jordan Penčev. 1996. Functions of the formant se/si in bulgarian. *Revue des études slaves*, 68(4):497–515.

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.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.

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 Vladimir Filipov. 2022. Acquisition of Modal Verbs in Bulgarian: Analysis of Longitudinal Data from CHILDES. *Ezikov Svyat (Orbis Linguarum)*, (ezs.swu.v20i3):332–346.

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*, pages 101–108, Online. Association for Computational Linguistics.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2010. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2001. Parsing the CHILDES database: Methodology and lessons learned. In *Proceedings of the Seventh International Workshop on Parsing Technologies*, pages 166–176, Beijing, China.

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

Kiril Simov, Petya Osenova, Alexander Simov, and Milen Kouylekov. 2005. Design and Implementation of the Bulgarian HPSG-based Treebank. *Journal of Research on Language and Computation. Special Issue*, pages 495–522.

Kiril Simov, Petya Osenova, and Milena Slavcheva. 2004. BTB-TR03: BulTreeBank Morphosyntactic Tagset. Technical report.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002a. 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).

Kiril Simov, Gergana Popova, and Petya Osenova. 2002b. 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.

Milena Slavcheva. 2003. Some aspects of the morphological processing of Bulgarian. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 71–77, Budapest, Hungary. Association for Computational Linguistics.

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. *arXiv preprint*.

Olga Mišeska Tomić. 2011. *16 Balkan Sprachbund features*, page 307–324. DE GRUYTER MOUTON.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

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 Sentence assignment to annotators

A1 (MM)

- CS: ALE_cs.conllu (sentences 66-245; 321-379)
- CDS: ALE_cds.conllu (sentences 1-215; 321-430)
- CS: BOG_cs.conllu (sentences 1-97)
- CDS: BOG_cds.conllu (sentences 1-148)
- CDS: TEF_cds.conllu (sentences 1-10)
- CS: SIM_cs.conllu (sentences 101-400; 505-924)
- CDS: SIM_cds.conllu (sentences 161-805)
- CS: ELI_cs.conllu (sentences 1-88)
- CDS: ELI_cds.conllu (sentences 101-400; 505-924)

A2 (YV)

- CS: BOG_cs.conllu (sentences 1–92)
- CDS: BOG_cds.conllu (sentences 1–142)
- CS: TEF_cs.conllu (sentences 276-336)
- CDS: TEF_cds.conllu (sentences 266-392)

A3 (YC)

- CS: ALE_cs.conllu (sentences 1–65)
- CDS: ALE_cds.conllu (sentences 1–90)
- CS: ALE_cs.conllu (sentences 246–320)
- CDS: ALE_cds.conllu (sentences 216–320)

A4 (TK)

- CS: SIM_cs.conllu (sentences 1–100)
- CDS: SIM_cds.conllu (sentences 1–160)
- CS: SIM_cs.conllu (sentences 401–515)
- CDS: SIM_cds.conllu (sentences 161-335)

A5 (IP)

- CS: TEF_cs.conllu (sentences 1–65)
- CDS: TEF_cds.conllu (sentences 1–90)
- CS: TEF_cs.conllu (sentences 65-180)
- CDS: TEF_cds.conllu (sentences 90-265)

A6 (TS)

- CS: ELI_cs.conllu (sentences 1–100)
- CDS: ELI_cds.conllu (sentences 1–160)
- CS: ELI_cs.conllu (sentences 101–121)
- CDS: ELI_cds.conllu (sentences 386–560)
- CS: TEF_cs.conllu (sentences 181–275)

B Annotation Scheme

B.1 Scope

The starting point is automatically generated Universal Dependencies (UD) parses produced using Stanza. For every sentence, annotators perform the following checks (and corrections if necessary), in order:

- Lemma
- UPOS
- Dependency head
- Dependency relation
- Morphological features (when clearly recoverable)
- XPOS tags are **not** corrected

The lemmas, UPOS tags, dependency edges, dependency labels, and morphological features should follow the guidelines of the UD Bulgarian dataset. An exhaustive list of the UPOS tags and dependency tags are explained with examples here: https://universaldependencies.org/treebanks/bg_btb/index.html

Gold labelled UD treebank for Bulgarian based on BulTreeBank: https://github.com/UniversalDependencies/UD_Bulgarian-BTB (files: bg_btb-ud-dev.conllu, bg_btb-ud-test.conllu, bg_btb-ud-train.conllu)

B.2 Step-by-Step Annotation Procedure

For each sentence:

1. Read the entire utterance.
2. Correct token-level information (lemma and UPOS).
3. Correct dependency heads.
4. Correct dependency relations.
5. Review morphological features when present.

B.3 Lemma Correction

- Lemmas must use standard Bulgarian orthography (dictionary entries) and be lowercased.
- The lemma of a proper name should also be lowercased, e.g. the lemma of Ани is ани).
- Normalise obvious child and regional variation (e.g. the lemma of икам is искам; the lemma of ши is ще; Кутийката-а-а-а – кутийка; панна [падна] – падам).
- Use the infinitive/base form for verbs (1st person singular, present).
- For verbs with reflexive clitic the lemma is скрия-(се), don't remove the dash, it is convention.
- If the surface form is diminutive, keep the lemma diminutive (e.g. книжката – книжка).
- Use nominative for the lemma of vocative nouns (e.g. the lemma for мамо is мама)
- The lemma of тати is unclear whether it should be тати or тате so for consistency we keep тати.
- For pronouns and function words, use standard dictionary lemmas
- For pronouns use 1st person singular, following the gold standard, e.g. й – аз; го – аз, си – се or си – свой.
- xxx is only used in CHILDES when the surface form is not recoverable. If the surface form is xxx, the lemma is also xxx.
- If the surface form is an unrecognisable word, its lemma is the lowercased version of it, e.g. Тпру → тпру.
- Repetitions: single lemma: Дай-дай-дай-дай → дай
- Example of frequent child words and their adult pairings, which should be used as their lemmas if recoverable: йейя – ле-ля, тетито/тефи/тефани/тефче - стефи/стефче/стефани, икам – искам, а – на,

B.4 UPOS Correction

Ensure UPOS matches syntactic behaviour. Use UPOS=X only when no category is recoverable.

- бе, хайде, айде, де all have UPOS tag interjection INTJ
- ще is AUX
- какво is tagged as DET in the gold
- да is AUX when used in verbal constructions, and PART when used in discourse
- Tokens marked xxx: recover UPOS if possible; otherwise use UPOS=X.
- Proper nouns: may not always be capitalised as expected, but treat them as PROPN (e.g. калоянчо, мончи)

B.5 Morphological Features

Do not guess features if the form is unclear. Leave empty.

Verbal

`Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act`

- Mood: Cnd, Imp, Ind
- Number: Plur, Sing
- Tense: Imp, Past, Pres
- Person: 1, 2, 3
- Aspect: Imp, Perf
- Voice: Act, Pass
- VerbForm: Fin, Part

Nominal

`Definite=Def|Gender=Fem|Number=Sing`

- Gender: Fem, Masc, Neut
- Case: Voc

- Definite: Def, Ind
- Number: Plur, Sing

Notes on the features:

- We only use `Case=Voc` for nouns but not for verbs.
- Mark vocative case even when the vocative form is the same as the nominal form (e.g. вуйчо, дай ми го).
- Definite: even if not with ът correctly when it is in a role of a subject, e.g. кълвача, the morphological feature should still include `Definite=Def`.

B.6 Punctuation

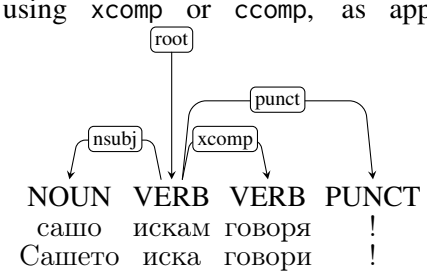
- Attach punctuation using `punct`.
- Attach to the syntactic head of the phrase that necessitates the use of the punctuation (not always the sentence root).

B.7 Dependency Head Correction

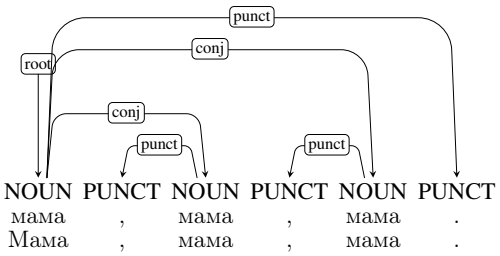
IMPORTANT: In INCEpTION dependency arcs are drawn from the UPOS box (blue) of the head to the UPOS box of the child. Hold Shift, select the UPOS of the head (From), and then select the POS of the dependent (To), let go of Shift. Edit the dependence relation.

Checklist:

- Exactly one root per sentence.
- Verify attachment directionality: the arrow should be pointing from the head to the child.
- Finite verb is typically the root.
- In copular clauses, the predicate nominal is the root (**not** the subject).
- In copula questions with k-word, the k-word is usually the root.
- In fragments, the semantic head is the root.
- Single-word utterances receive root relation.
- In cases of repetition, first token is root; others attach via `conj`.
- When no element is subordinated (e.g. Христо Смирненски the leftmost element is the root.
- In child utterances where да is omitted, attach infinitival verb as in adult speech, using `xcomp` or `ccomp`, as appropriate.



- Coordination:
  - First conjunct attaches to the head.
  - Subsequent conjuncts attach via `conj`.
  - Coordinating conjunction, if present, attaches via `cc`.



- Punctuation attach to the particle/clause that necessitates it – not always the root (see above).
- Adposition attachment to the head of the phrase rather than to interrogative modifier.

B.8 Dependency Relation Correction

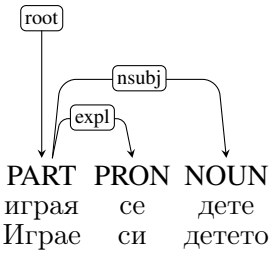
Common distinctions (review the examples file `UD_relations_example_for_each_relation` on INCEpTION). Below is a list of all relations and how to distinguish between them:

- **xcomp vs ccomp**:
  - `xcomp` = embedded clause without its own subject (the subject of the xcomp is inherited from the higher clause);
  - `ccomp` = embedded clause with its own subject.
- **parataxis vs conj** In UD-BulTreeBank, parataxis is used exclusively for direct speech in a sentence. Following the gold standards, when there are clauses which are equal (neither is subordinated) we use conj even if a cc is missing.

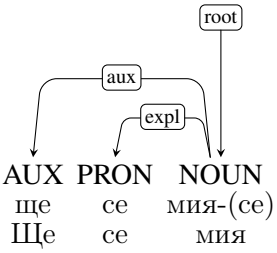
- **nmod vs obl**:
  - `nmod` nominal modifier modifies a noun or nominal phrase.
  - `obl` oblique modifies averb/verbal phrase; it introduces **optional** circumstantial information.
- **iobj vs obl**: not everything with an adposition is indirect object; consider the verb valency and whether the sentence would be complete without this phrase – if the phrase is optional, it is `obl`.
- **fixed vs flat**:
  - `fixed` = rigid multiword expressions with no internal syntax (e.g. multi-word coordinating conjunction, such as само че);
  - `flat` = names or sequences with internal structure suppressed (e.g. proper names, such as Кума Лиса).
- **nmod vs flat**: In noun+name expressions (e.g. баба Лили лили has head баба with relation `nmod` (as opposed to flat; flat is for e.g. name + surname)
- **expl (expletive)**: Reflexive pronoun e.g. се, си; as well as the clitic in clitic doubling.
- **vocative**: Direct address to interlocutor (e.g. мамо).
- **discourse**: Discourse particles or interjections (e.g. бе, хайде, де, ами).
- **root**: Main predicate of the sentence.
- **amod**: Adjectival modifier of a noun.
- **advmod**: Adverbial modifier of verb/adjective.
- **det**: Determiner modifying a noun; can also be used for possessive clitics (e.g. крачето си, неговия учител, as well as for numerical determiner един момък дойде; една жена ми каза
- **case**: Adposition introducing nominal dependents.
- **cc**: Coordinating conjunction.
- **conj**: Conjunct in coordination.
- **cop**: Copula (usually forms of съм).
- **aux / aux:pass**: Auxiliary verbs; passive auxiliaries use `aux:pass`.
- **nsubj / nsubj:pass**: Nominal subject (active / passive).
- **obj**: Direct object.
- **iobj**: Indirect object.
- **mark**: Subordinating marker (e.g. че)
- **acl / acl:relcl**: Clausal modifier of noun; for relative clauses introduced by a k-word (който/която/които), use `acl:relcl`, otherwise use `acl`
- **advcl**: Adverbial clause modifier.
- **advcl vs obl**: advcl is for clauses (with verb), whereas obl is for noun phrases (for преди да отиде use advcl, for преди обед use obl).
- **appos**: Appositional noun phrases.
- **nummod**: Numerical modifier.
- **punct**: Punctuation.

B.9 Common difficult cases with examples

- Expletive си vs determiner си
  - **expl** if it is part of the verb, as reflexive dative clitic (ethical/benefactive dative)

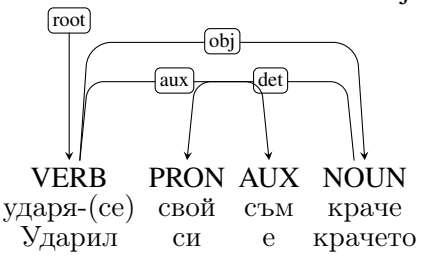


  - **expl** if it is part of the verb, as accusative reflexive clitic (true reflexive)

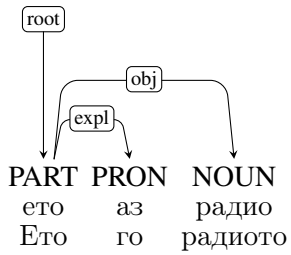


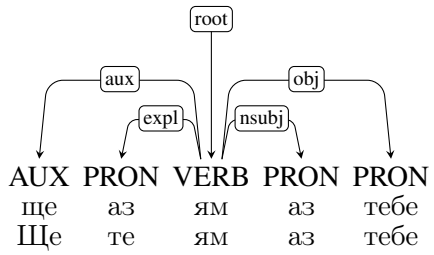
  - **det** if it can be replaced with свой (if it can be replaced with свой/своя/свои,

then its lemma is свой and the relation is det with head the object)

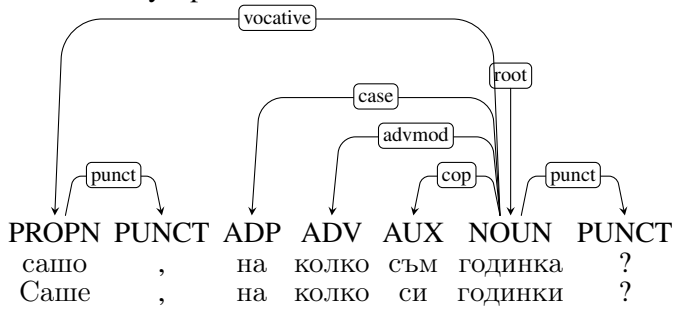


- Clitic doubling: the clitic should be attached to the root of the verb (or clause), i.e. to the same head as the object and be tagged with relation **expl**



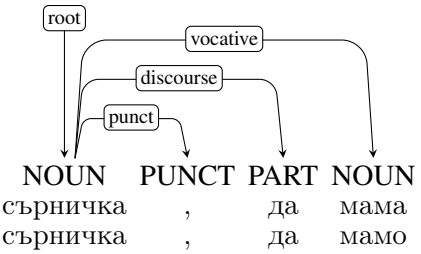


- nsubj or vocative: When a form could be analysed as both nsubj and vocative we always pick the vocative relation.



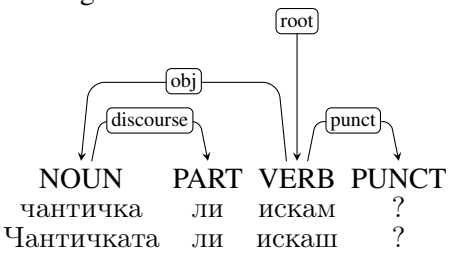
  - Attach punctuation to the head that necessitates it – e.g. in the above example, Саше, на колко си годинки? the vocative noun necessitates the comma, so it is attached to it.
  - Attach adposition to the same head as the head of the k-word.

- мамо referring to the child: treat it as a standard vocative noun (not as adjective мамин)

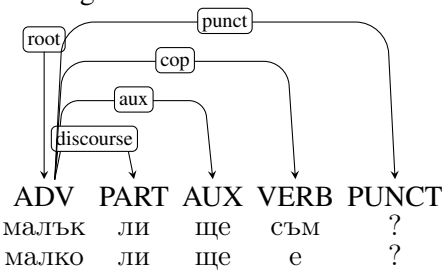


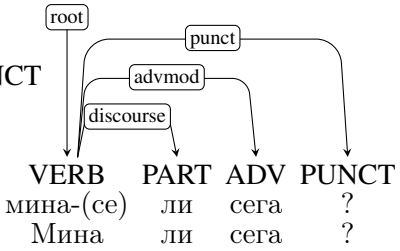
- attachment of ли: In the gold standard it is attached to the nearest item. ли is a clitic which changes the focus (topic) so attach it to the focus of the sentence, which might be the root. If it is ambiguous, then attach it to the root.

  – attaching to the focus

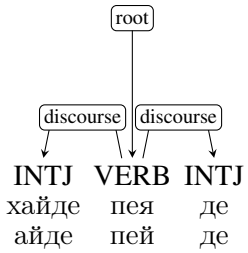


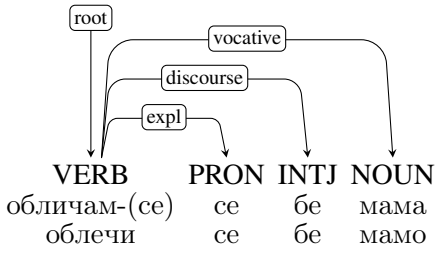
  – attaching to the root



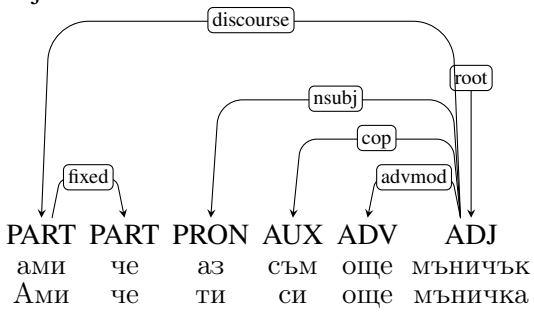


- Multiple interjections (discourse items): test by removing each interjection, and if the sentences with one interjection is still grammatical, then attach the interjections independently to the root using relation **discourse**

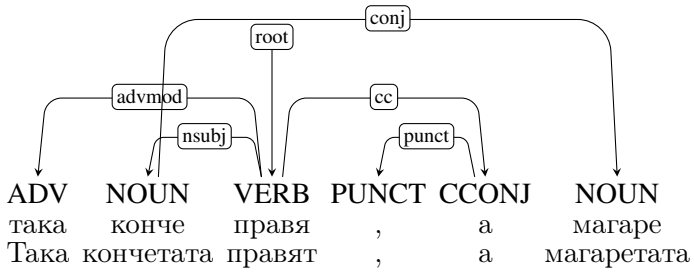




- multi-word coordinating conjunctions: use **fixed**



- ellipsis – difficult examples, excluded from the UD BulTreeBank. Attach words to the corresponding word with the same part of speech.



B.10 Final checks

- Exactly one root
- No orphan nodes

If uncertain:

- Follow the existing gold standard: https://github.com/UniversalDependencies/UD_Bulgarian-BTB
- Leave a comment in the annotation tool.