Dependency Parsing Evaluation for Low-resource Spontaneous Speech

Zoey Liu
Department of Computer Science
Boston College
ying.liu.5@bc.edu

Emily Prud'hommeaux
Department of Computer Science
Boston College
prudhome@bc.edu

Abstract

How well can a state-of-the-art parsing system, developed for the written domain, perform when applied to spontaneous speech data involving different interlocutors? This study addresses this question in a low-resource setting using child-parent conversations from the CHILDES databse. Specifically, we focus on dependency parsing evaluation for utterances of one specific child (18 - 27 months) and her parents. We first present a semi-automatic adaption of the dependency annotation scheme in CHILDES to that of the Universal Dependencies project, an annotation style that is more commonly applied in dependency parsing. Our evaluation demonstrates that an *out-of-domain* biaffine parser trained only on written texts performs well with parent speech. There is, however, much room for improvement on child utterances, particularly at 18 and 21 months, due to cases of omission and repetition that are prevalent in child speech. By contrast, parsers trained or fine-tuned with *in-domain* spoken data on a much smaller scale can achieve comparable results for parent speech and improve the weak parsing performance for child speech at these earlier ages.

1 Introduction

While the task of dependency parsing has been studied extensively, in both monolingual (Eisner, 1996; Sun et al., 2019) and crosslinguistic contexts (McDonald et al., 2013; Agić et al., 2016), much of the effort thus far has been devoted to parsing data from the written domain. By contrast, dependency parsing of spontaneous speech in general has received little attention, despite the notable and sometimes quantifiable differences of syntactic variations between written and spoken registers (Biber, 1991; O'Donnell, 1974).

One obstacle to the development of dependency parsing for speech is that, in comparison to the written domain, which offers numerous corpora with gold-standard (morpho)syntactic annotations in a large variety of languages (Zeman et al., 2020), data sets annotated with comparable levels of structural details in the spoken genre are relatively rare, regardless of the particular annotation frameworks adopted. One of the most widely used and largest spoken corpora is the Switchboard corpus (Godfrey et al., 1992), which provides constituency structures for around one million tokens from transcriptions of conversational speech in English.

Other available spoken corpora are of much smaller scale. For instance, Dobrovoljc and Nivre (2016) developed a dependency treebank for spoken Slovenian also using transcripts of spontaneous speech, which contains 29,488 tokens in total. Spence et al. (2018) built a dependency treebank for Hupa, an indigenous language residing in Northern California; the treebank used mainly transcriptions of story telling from the only master speaker of the Hupa language and contained around 6,561 mannualy annotated tokens. A few treebanks in the Universal Dependencies project v2.7 (Zeman et al., 2020) (hereafter UD) include transcriptions of spoken data, but most such data is mixed in with written texts and the register is not specified for individual sentences.

The limited availability of spoken corpora leaves open the questions of how well the current state-of-the-art dependency parser, developed mainly for written texts, performs on low-resource naturalistic speech and to what extent the performance of the parser differs depending on the specific role of the speaker. It is possible that recent advances in the parsing of written texts (Kondratyuk and Straka, 2019; Bouma et al., 2020; Qi et al., 2020) will extend to speech and that errors will be restricted to cases that are unique in the spoken genre, such as omissions of frequent or short words (Ferreira and Anes, 1994).

Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 156–165
April 20, 2021.

Alternatively, it is also reasonable to speculate that such parsers would not align well with speech given that they are being applied to a domain that is different from their original training data. Nevertheless, the potential differences between the dependency structure of an utterance assigned by a parser and one assigned manually might not necessarily indicate that there are *errors* but rather that there is more than one way of interpreting the utterance when the discourse context is not considered. Understanding whether and where parsers for written texts fall short in low-resource spontaneous speech has the potential to contribute to developing parsing systems for the spoken domain, as well as expanding the availability of spoken data.

This study makes a step towards the aforementioned directions. In order to investigate the parsing performance for spoken data in a low-resource setting (Vania et al., 2019; Agić et al., 2016) , especially spontaneous speech that involves different interlocutors, we focus on child-parent conversational interactions as our targeted evaluation data. Specifically, the evaluation sets contain utterance samples of one child, "Eve", and her parents from the Brown Corpus (Brown, 1973) in the CHILDES database (MacWhinney, 2000).

The remainder of the paper is structured as follows: we review related work in section 2; section 3 points out some general issues with parsing speech; section 4 describes our procedures of adapting the current dependency annotation scheme in CHILDES to that of UD; section 5 presents our experiments[1]; and we conclude with section 6.

2 Related Work

Although most work in the dependency parsing and annotation literature has examined the written domain, there have been a number of notable exceptions. Looking at the domain of human-computer interactions, Adams (2017) adapted the UD annotation scheme to a data set of 882 sentences that are mostly transcripts generated from automatic speech recognition systems. Their work showed that a parser ensemble approach (two parsers in this case) yielded the best results. Davidson et al. (2019) also investigated dependency parsing with transcripts from dialogue systems; using a data set of 1,500 sentences annotated with UD labels, they found that the best parsing accuracy was achieved when an out-of-domain parser was fine-tuned with more in-domain data.

Other work such as Kong et al. (2014) and Liu et al. (2018) has focused on dependency parsing of tweets, which can be considered a domain that lies between the spoken genre and other written contexts. The data set from Liu et al. (2018) contains a total of 3,550 tweets, four times the number used in Kong et al. (2014). Their annotation standards followed the UD style while allowing multiple roots in tweets that have more than one sentences indicated mainly by punctuation. They reported that an ensemble approach with 20 different parsers was the most effective.

3 Issues with Speech Orthography

One major challenge with parsing speech lies in finding a standard way of transcribing and formatting spoken language. In particular, syntactic parses should be developed for what the speakers have actually said, without further added facilitation during the transcription process. This general issue has not been addressed consistently, at least in the current format of CHILDES. As a demonstration, consider the following annotated examples (adapted from their initial forms in CHILDES).

(1) *that green bean white* **.**

(2) *you* ***v*** *more cookies*

(3) *well* ***(tagmarker)*** *we already did that*

Cases such as (1) contain punctuation, yet punctuation is not explicitly articulated during spontaneous speech. With transcriptions like (2), in this particular instance *v* represents an omitted verb (e.g. *want*). In example (3), a tagmarker was added in order to (mostly) indicate that the first token of the sentence is a discourse marker or communicator. All of these added items were annotated with syntactic dependencies in CHILDES.

The problems with including punctuation and tagmarkers for syntactic analyses could potentially be relaxed to some extent via automatically removing the added items from the dependency parse following certain heuristics. For example, a rule of thumb could be to change all tokens before the tagmarker to be a dependent of the root of the utterance, with the dependency relation *discourse*, then discard the tagmarker itself. However, this would not entirely alleviate the problem, since the

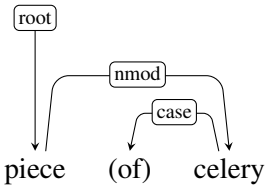
[1]Our annotations, code and full results for parser comparisons are in quarantine at https://github.com/zoeyliu18/Parsing_Speech.

parser sometimes assign dependents to punctuation or tagmarker. In our manual check of about 250 utterances of the English data from CHILDES that were not labeled as gold-standard (see section 5), this was indeed the case. While in some sentences the root can be assigned as the head of the dependents of the punctuation or tagmarker assigned by an automatic parser, in other cases, the correct syntactic heads are not easily identifiable without actually looking at the utterance.

Utterances with omitted words are even more challenging than those with punctuation or tagmarkers. Ideally, to syntactically analyze spoken data, the focus should be on the actual utterances rather than what the annotators *think* the utterance should be, especially since the inclusion of additional words in transcribed speech assumes that there were words that were omitted to begin with. This is not to deny that all manual annotations are subjective in some respects, and a less arguable dependency parse could be assigned to an utterance with potential word omissions when the preceding or following contexts are taken into account. For instance [2], in example (4) *celery* is treated as a nominal modifier of *piece*; and since the token *of*, which did not occur in the child utterance, has no further dependents, it could then be removed from the parse without causing any ambiguous issues.

(4) *Child: piece celery*

Parent: piece of celery

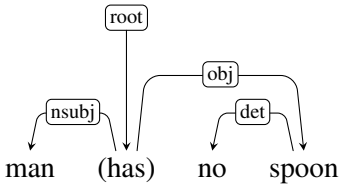


Nevertheless, that is not always the case. Consider (5), where the utterance could be interpreted in at least in three different ways , each with a different syntactic parse.

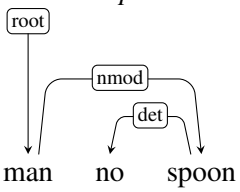
(5) *Child: man no spoon*

Parent: what

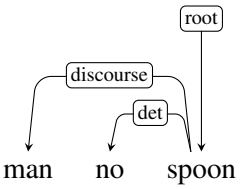
First, *man* could be the subject of the sentence an omitted verb: *man has no spoon*.



Secondly, instead of a verb, if there were indeed an omitted word, it could also be a preposition: *man with no spoon*.



Thirdly, *man* could be a discourse marker and the sentence would have a reading of *man (!) no spoon*.



To address these issues in a consistent way, in the parsing evaluation experiments here, we did not include any punctuation, tagmarkers, or omitted words in the utterances.

4 Annotation Adaptation to UD

The dependency annotations in CHILDES, which now have 37 different dependency relations, were initially developed by Sagae et al. (2004) (see also Sagae et al. (2005)). The annotation scheme was then slightly modified by Sagae et al. (2007), with adjustments to cases such as locative phrases and verb particles. Given that most of the available dependency parsers for written texts in English abide by the style of UD (Zeman et al., 2020), we first adapted the CHILDES dependency relations so that they would be comparable to those in UD.

4.1 Direct (one-to-one) conversion

Comparing the two parsing schemes, we found that there are several dependency relations that can be converted from CHILDES to UD straightforwardly, such as subjects and objects (Table 1). Other dependencies can be changed directly with the help of part-of-speech (POS) or lexical information. For instance, in the phrase *my stool*, *my* is annotated as the determiner (DET) in CHILDES and its POS is pronoun; for cases like this, the original DET relation was changed to the possessive nominal mod-

[2] All dependency graphs in this paper were annotated according to the UD annotation style.

ifier *nmod:poss* based on UD. As another example, the dependency relation LINK is used for any relativizer, complementizer, or subordinate conjunction; for such cases, the dependency can be changed to *mark* given the UD annotations if the lexical word is *if*, *because*, *so*, etc.

CHILDES	UD
SUBJ	nsubj
OBJ	obj
OBJ2	iobj
CSUBJ	csubj
APP	appos
VOC	vocative
INF	mark
DATE	nummod
ENUM (enumeration)	conj

Table 1: Examples of dependency relations in English that can be converted directly from CHILDES to UD.

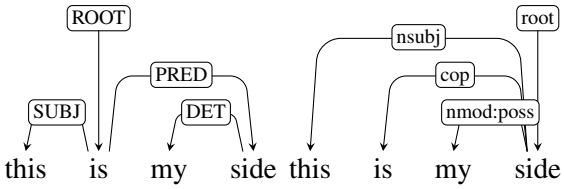
4.2 Collapsing dependency relations

Additionally, we also collapsed dependency relations in CHILDES that have very similar syntactic functions, the differences of which can be easily identified automatically if needed. For instance, BEG, COM and END all describe communicators (and sometimes vocatives). The major difference between them is only their relative position in the sentence, with BEG describing sentence-initial communicators, COM sentence-medial, and END sentence-final. Therefore we combined them all to the *discourse* relation in UD. On the other hand, CHILDES has two different relations for the root of the sentence, ROOT and INCROOT, with the former applied to utterances that are considered full sentences (e.g. *I wanna go*) while the latter specifically used for cases without a verbal or copula root (e.g. *not on the couch*). When adapting to UD we annotated both with the *root* relation.

4.3 Function head vs. content head

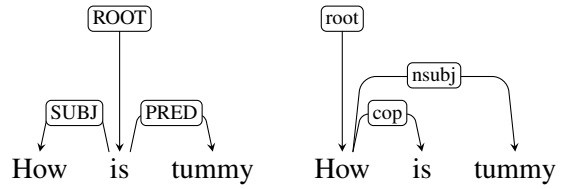
While the aforementioned conversions are comparatively trivial, other automatic conversions require more heuristic steps. The main reason is that the annotation style of UD favors content head over function head for better crosslinguistic applicability, while this was not always the case with CHILDES. In particular, the English annotations in CHILDES choose function word as the syntactic head in structures that involve copula verb, prepositional phrase (PP), conjunction with coordinators, and negation.

Copula The treatment of copula structures in CHILDES is consistent across the utterances, where the copula verb is the root of the sentence. By contrast, UD has different annotations for the copula structures depending on whether the utterance is a statement (Figure 1), question (Figure 2), and whether there is an expletive in the sentence (Figure 3). For all utterances with a copula root from CHILDES, we determined whether they were questions/interrogatives or had expletives based on the lexical word of the annotated subjects, and the rest were treated as statements. With statements and questions in particular, the predicate of the copula verb was instead the new root of the sentence; any initial dependents of the copula that are not the subject or the predicate in the CHILDES annotations were converted to be dependents of the new root in the UD annotations.



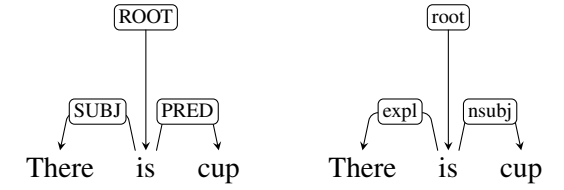
(a) CHILDES (b) UD

Figure 1: Annotations of copula structures in statements.



(a) CHILDES (b) UD

Figure 2: Annotations of copula structures in questions.

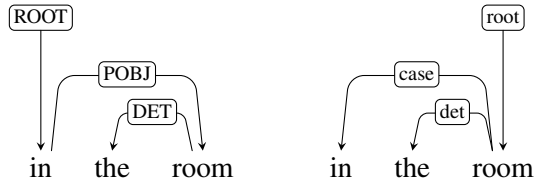


(a) CHILDES (b) UD

Figure 3: Annotations of copula structures in expletives.

Prepositional phrase In CHILDES, a PP has the preposition as its syntactic head, while UD chooses the lexical head as the head of the phrase in order for the annotations to be more comparable for languages with different case marking systems.

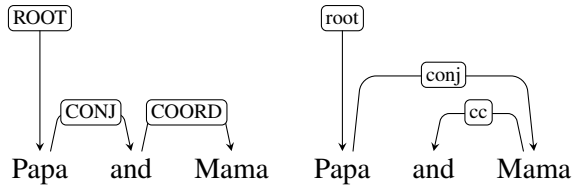
In our conversion, the initial prepositional object (POBJ) was the new head of the phrase and the preposition was treated as the *case* dependent (Figure 4). Any other initial dependents of the preposition were changed to be dependents of the lexical head in UD.



(a) CHILDES (b) UD

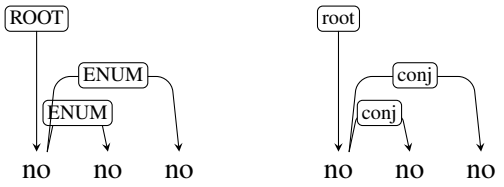
Figure 4: Annotations of prepositional phrases.

Conjunction The major difference in annotations of conjunctions with coordinators between CHILDES and UD, as illustrated in Figure 5, is that in CHILDES, for a series of coordinated items, the items other than the first one are dependents of their respective preceding coordinator, whereas they are all dependents of the first coordinated word in UD and also heads of their coordinators. This is distinct from structures with conjunctions that do not have coordinators (Figure 6), which have the same dependency parse in CHILDES and UD, except for the dependency relation (enumeration in CHILDES and conjunction in UD).



(a) CHILDES (b) UD

Figure 5: Annotations of conjunctions with coordinators.



(a) CHILDES (b) UD

Figure 6: Annotations of conjunctions without coordinators.

Negation In utterances with auxiliary or copula verbs modified by negation markers such as *not* or *n't*, the negation markers are annotated as the dependents of the auxiliary/copula in CHILDES. When converting to UD, if the auxiliary/copula is not the root of the sentence, we changed the dependencies of the negation markers so they have the same syntactic heads as the auxiliary/copula.

4.4 Manual inspections

We reviewed all of the utterances thoroughly after automatic conversion. This was performed by an author of this paper, who has had extensive training in dependency linguistics. We tried to be as faithful as possible to the initial gold-standard annotations in the data and their interpretations accordingly.

Our inspections mainly concern syntactic dependencies that were not converted completely. These largely fall in three aspects. The first is temporal modifiers, which are treated as adverbial modifiers (JCT) in CHILDES. We changed them to either *obl:tmod* or *nmod:tmod* depending on whether they were modifying a verbal or nominal head. The second is clausal modifiers, which could be relative clauses or complements in CHILDES. We changed these to either *acl:relcl* or *ccomp* when converting to UD.

The third involves clausal conjunct (CJCT). These include coordinated clauses within one utterance (e.g. *I hear and watch Sarah*) that may or may not have the coordinators, or cases that seem less like one utterance but rather "side-by-side" sentences (e.g. *I say see Mommy*). The former was adapted to UD with *conj*; while for the latter we resorted to the *parataxis* relation (Figure 7). To be consistent with the UD style, we annotated the head of the first sentence to be the root of the whole utterance, and the heads of other sentences in parallel are dependents of the root.

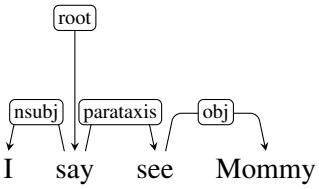


Figure 7: Annotations of clausal conjunctions.

In addition, we also carefully examined new dependencies that were added due to our modified tokenization scheme. We modified the initial tokenization of possessives (e.g. *Mommy's*), combined adverbs (e.g. *as_well*) and combined conjunctives (e.g. *in_case_that*). These are all treated as one

token in CHILDES, and in the UD annotations we split them into individual tokens based on the apostrophe or the underscore. While these tokenizations were first carried out during the automatic process, we checked to ensure that their dependency relations aligned with those in UD and that we were not including cases such as rhymes or onomatopoeia, which share similar tokenization standards with combined adverbs or conjunctives.

5 Experiments

5.1 Data

	18 mos	21 mos	24 mos	27 mos
Eve				
utterances	426	341	280	235
tokens	1,023	1,018	1,025	1,041
MLU	2.40	2.99	3.66	4.43
Parent				
utterances	203	210	190	180
tokens	1,034	1,039	1,041	1,046
MLU	5.09	4.95	5.48	5.8

Table 2: Descriptive statistics of the number of utterances, words and mean utterance length (MLU) for child and parent speech, at the corresponding age of the child in months (mos).

While the available corpora of English from CHILDES contain syntactic dependency parses, most of these annotations were derived automatically from the MEGRASP developed by Sagae et al. (2010), In the English sections, the Eve Corpus from the Brown Corpus (Brown, 1973) provides gold-standard manual annotations (Sagae et al., 2007) for its first fifteen transcripts out of the twenty files in total. Given the age range of the child Eve (18-27 months), we took around 1,000 tokens from child and parent utterances respectively, at the different ages of the child (Table 2) among the twenty files. Each utterance contained at least two tokens in order to not inflate the results of parsing and evaluations. We then converted the dependency annotations of these utterances to UD-style based on the procedures described in section 4.

5.2 Out-of-domain training

To evaluate parsing systems that were mainly developed for written corpora in English, we used Stanza (Qi et al., 2020), an open-source natural language processing toolkit that was developed with the multilingual corpora from UD. Using the default parameter setting for English in Stanza as well as Glove embeddings (Pennington et al., 2014), we first trained a POS tagger and a biaffine parser (Dozat et al., 2017; Qi et al., 2018) with the predefined train-development-test sets from the UD English Web Treebank (UD-EWT). The UD-EWT tagger and parser were then applied to our evaluation sets of child and parent speech.

Unlabeled attachment score (UAS) and labeled attachment score (LAS) were taken as indexes of parsing performance (Kübler et al., 2009). Since the utterances in each of our evaluation sets came from the same speaker, and overall our data was taken from one child and her parents, this indicates that the utterances under examinations were not exactly independent from one another. Due to this reason as well as the small size of our evaluation data, we used bootstrapping (Berg-Kirkpatrick et al., 2012; Dror et al., 2018; Efron and Tibshirani, 1994) for significance testing. Given an evaluation set with a total of N utterances which have been automatically parsed, we randomly selected N sentences from the set with replacement, then calculated the UAS and LAS of the sample. This process was repeated for 10,000 iterations, which yielded an empirical distribution of the LAS and UAS values respectively. We then computed the mean and the 95% confidence intervals of these empirical distributions.

Parsing	
UAS	89.33 (88.60, 90.03)
LAS	86.30 (85.52, 87.07)

Table 3: Parsing accuracy for UD-EWT test set evaluated with the UD-EWT parser. UAS refers to unlabeled attachment score, while LAS refers to labeled attachment score. Significance testing was performed with bootstrapping.

5.3 Discrepancy analysis

As certain utterances can have multiple interpretations according to whether contexts are taken into account (see section 3), especially cases in child speech, and since our UD dependency annotations rely heavily on the existing dependency relations from CHILDES, we are hesitant to label automatic parses that are different from the annotations in our UD conversions as errors. Instead we examined the parse results as a way to see what discrepancies there were between our annotations and the parser that is targeted towards written data.

Data	Parser	18 mos	21 mos	24 mos	27 mos
Parent	***UD-EWT***				
	UAS	**95.95**	**92.18**	92.51	91.30
	LAS	**91.10**	**88.72**	**88.36**	**87.09**
Child	***UD-EWT***				
	UAS	84.86	74.24	89.94	86.26
	LAS	68.13	59.36	80.58	79.89
	UD-EWT fine-tuned with parent speech				
	UAS	82.40	74.47	92.97	89.53
	LAS	72.12	66.14	**87.19**	83.85
	UD-EWT fine-tuned with child speech				
	UAS	88.59	**82.77**†	90.15	83.09
	LAS	**77.15**	**73.38**	83.31	75.59
	Parent speech				
	UAS	78.23	74.45	92.9	89.54
	LAS	66.01	66.13	**87.24**	83.88
	Parent speech fine-tuned with child speech				
	UAS	91.02	79.53	90.64	88.29
	LAS	**81.05**	**71.68**	83.52	80.12

Table 4: Parser accuracy for (1) parent speech with the UD-EWT parser; (2) child speech with the UD-EWT parser; and (3) child speech with various parser training configurations. At each of the four ages, the scores for (1) and (3) were compared with (2). Boldface indicates a significant difference, derived by comparing their respective 95% confidence intervals after bootstrapping.

5.3.1 Parent speech

As presented in Table 4, by contrast, the UD-EWT parser appears to perform relatively well when applied to parent speech. The results are comparable to those when evaluating the parser on just the test set of UD-EWT and mostly better than those when the same parser was evaluated with child speech; these patterns seem to be more or less consistent across the different ages of the child. This pattern is possibly due to the following factors: (1) parent utterances are on average longer than those produced by the child yet still generally shorter than written texts; (2) parent speech is less likely to have omitted words and repetitions, which helps the parser to analyze the syntactic structures of the utterances more deterministically.

Manual inspection across the evaluation sets reveals three major discrepancies between the automatic parses of parent speech and our annotations. First, in CHILDES, for noun phrases where the first noun serves more or less as a modifier of the second noun, during our UD conversion, the first noun was annotated as a nominal modifier (*nmod*) of the second noun, such as the word *grape* in the phrase *grape juice*. The automatic parser, however, consistently treated the first noun as a *compound* dependent of the second. This accounts for around 9.68% of all parsing discrepancies in parent speech.

Secondly, prepositions that lack their own lexical object and co-occur with verbs are labeled in CHILDES as either adverbial modifiers or verb particles. If the meaning of the preposition is essential in deriving the compositional meaning of the verb and the preposition (e.g *look it up*), then it is treated as a particle; otherwise the preposition is considered an adverbial modifier of the verb (e.g. *tear if off*). The latter in our UD annotations was directly converted to *advmod*. On the other hand, the parser treated the prepositions in cases like this almost entirely as particles with the *compound:prt* dependency. This accounts for about 9.89% of all parsing differences in parent speech.

The discrepancies described above are not particularly crucial, since the parser annotated related cases in a consistent fashion, which could be adapted easily if necessary. Additionally, there is not always a clear criterion determining whether two nouns should be considered as a compound or whether a preposition should be a particle of a verb. In comparison, the last main difference between the automatic parse and our annotations, which accounts for around 14.41% of all discrepancies in

parent speech, is related to correctly labeling words that are vocatives or discourse markers of the utterance. Such instances are less straightforward to resolve. For instance, the utterance *who are you calling Eve* is directed to the child. Thus the word *Eve* should be treated as a vocative of the root of the sentence, *calling*; yet the parser instead labeled Eve as a direct object of calling.

5.3.2 Child speech

Looking at child speech, the three main parsing discrepancies found in parent utterances are also observed (*nmod* / *compound*: 10.08%; *advmod* / *compound:prt*: 4.60%; *discourse* and *vocative*: 10.26%). That said, a large number (around 10.08%) of differences between the automatic parses of child speech and our annotations result from utterances lacking a clear syntactic structure, which potentially leads the utterance to having more than one interpretation, especially when context is disregarded. These utterances are mostly telegraphic speech consisting of two tokens.

For example, the utterance *Eve writing* could mean the writing belongs to Eve, or Eve is doing the action of writing. Based on initial annotations from CHILDES as well as the surrounding context of this sentence, *Eve* was converted to be the subject of *writing* in our annotations. Yet the parser annotated it as the *compound* or sometimes an adjective modifier (*amod*) of *writing*. This type of utterances is more common when the child is at the age of 18 or 21 months at least in our evaluation data, which drives the parsing differences for child speech at the different ages in Table 4.

Another reason why the parser does not seem to align well with our dependency annotations when the child is 18 months and especially 21 months old is due to repetition or conjunction without coordinators in speech. For example, the utterances *writing a b c* was annotated with a dependency structure as Figure 8. Nevertheless, the automatic parser's treatment of cases as such was more random; it would either label *a* as a determiner of *b*, a compound or a nominal modifier of *c*.

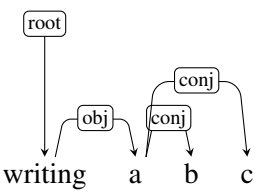


Figure 8: Annotations of *writing a b c*

5.4 In-domain training

To investigate whether the parsing performance for child speech could be improved with help from domain-specific spoken data, we experimented with four different training settings without changing our annotation decisions analyzed above.

The first and second approaches applied fine-tuning to the UD-EWT parser initially trained on just written texts with either parent or child speech. When fine-tuning with parent speech, we concatenated all parent utterances in our evaluation data. When fine-tuning with child speech, we concatenated all child utterances except for those to be evaluated (e.g. fine-tuning with data from the ages of 21, 24 and 27 months when evaluating child speech at the age of 18 months). The third approach trained an biaffine parser with the same parameter settings as those for the UD-EWT parser, except with all utterances of the parent speech in our evaluation sets. The last approach fine-tuned the parser trained with just parent speech with additional data from child speech; the fine-tuning was done in the same way as described in the second approach.

The training data of the UD-EWT parser is about 48 times larger than that of all parent utterances; yet based on results from Table 4, the parser trained with just parent speech performs on par with the UD-EWT parser. The overall best parsing performance was achieved when the parsers were fine-tuned with child speech. This is the most obvious with child speech at 18 or 21 months; whereas in cases with already decent parsing results such as child speech at 27 months, there does not appear to be a significant gain with fine-tuning.

6 Conclusion

In our low-resource evaluation settings, we found that an out-of-domain parser trained only with written texts performed well with parent speech, but not necessarily so with child speech. On the other hand, similar or better performance was achieved when the parser was fine-tuned with a very limited amount of in-domain data.

Given the relatively small size of our evaluation data and the narrow age range of the child, it is certain that there are other common phenomenon unique to child-speech that have not been covered, such as disfluency. For future work, we would like to conduct a more thorough and larger-scale parsing evaluation with speech across wider age ranges of different children. In addition, we plan

to extend our study with further experimentation of different domain adaptation methods to other spoken domains such as medical conversations as well as languages. Besides English, CHILDES also provides dependency annotations in child-parent interactions in other languages, such as Spanish (Sagae et al., 2010), Japanese (Miyata et al., 2013), Dutch (Odijk et al., 2018) and Hebrew (Gretz et al., 2015). Since annotation standards in in these corpora largely abide by those of English (Sagae et al., 2007), our semi-automatic UD conversion scheme described here could be applied to these languages, along with modifications of language-specific morphosyntactic structures.

References

Allison Adams. 2017. Dependency parsing and dialogue systems: an investigation of dependency parsing for commercial application.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Gosse Bouma, Yuji Matsumoto, Stephan Oepen, Kenji Sagae, Djamé Seddah, Weiwei Sun, Anders Søgaard, Reut Tsarfaty, and Dan Zeman, editors. 2020. *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*. Association for Computational Linguistics, Online.

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

Sam Davidson, Dian Yu, and Zhou Yu. 2019. Dependency parsing for spoken dialog systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1513–1519, Hong Kong, China. Association for Computational Linguistics.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573, Portorož, Slovenia. European Language Resources Association (ELRA).

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30, Vancouver, Canada. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Fernanda Ferreira and Michael Anes. 1994. Why study spoken language? In Morton Ann Gernsbacher, editor, *Handbook of psycholinguistics*, page 33–56. Academic Press.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Shai Gretz, Alon Itai, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2015. Parsing hebrew childes transcripts. *Language Resources and Evaluation*, 49(1):107–145.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar. Association for Computational Linguistics.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. *Synthesis lectures on human language technologies*, 1(1):1–127.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Susanne Miyata, Kenji Sagae, and Brian MacWhinney. 2013. The syntax parser grasp for childes (in japanese). *Journal of Health and Medical Science*, (3):45–62.

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*, pages 2275–2283. European Language Resources Association (ELRA).

Roy C O'Donnell. 1974. Syntactic differences between speech and writing. *American Speech*, 49(1/2):102–110.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

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. 2007. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague, Czech Republic. 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. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 197–204, Ann Arbor, Michigan. Association for Computational Linguistics.

Kenji Sagae, Brian MacWhinney, and Alon Lavie. 2004. Adding syntactic annotations to transcripts of parent-child dialogs. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Justin Spence, Zoey Liu, Palakurthy Kayla, and Lee-Wynant Tyler. 2018. Syntactic annotation of a hupa text corpus. In *Working Papers in Athabaskan Languages: Alaska Native Language Center Working Papers*, pages 37–53. Fairbanks, AK: ANLC Publications.

Weiwei Sun, Yufei Chen, Xiaojun Wan, and Meichun Liu. 2019. Parsing Chinese sentences with grammatical relations. *Computational Linguistics*, 45(1):95–136.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1105–1116, Hong Kong, China. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, et al. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.