Data-driven Parsing Evaluation for Child-Parent Interactions

Zoey Liu
Department of Linguistics
University of Florida, China
liu.ying@ufl.edu

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

Abstract

We present a syntactic dependency treebank for naturalistic child and child-directed spoken English. Our annotations largely follow the guidelines of the Universal Dependencies project (UD [Zeman et al., 2022]), with detailed extensions to lexical and syntactic structures unique to spontaneous spoken language, as opposed to written texts or prepared speech. Compared to existing UD-style spoken treebanks and other dependency corpora of child-parent interactions specifically, our dataset is much larger (44,744 utterances; 233,907 words) and contains data from 10 children covering a wide age range (18–66 months). We conduct thorough dependency parser evaluations using both graph-based and transition-based parsers, trained on three different types of out-of-domain written texts: news, tweets, and learner data. Out-of-domain parsers demonstrate reasonable performance for both child and parent data. In addition, parser performance for child data increases along children's developmental paths, especially between 18 and 48 months, and gradually approaches the performance for parent data. These results are further validated with in-domain training.

1 Introduction

Research on syntactic dependency parsing has experienced tremendous progress with the continuous development of the Universal Dependencies project (UD) (Zeman et al., 2022). That said, of the 228 treebanks in the latest version of UD (v2.10), only 12 consist of fully spoken data (see also Dobrovoljc, 2022), while the rest focus on different genres within the written domain. This means most (if not all) state-of-the-art off-the-shelf dependency parsers are oriented towards written texts rather than tailored specifically to spontaneous spoken language. Therefore a natural question arises: How well will parsing systems developed for written data perform on spontaneous spoken language?

Over the past decade, there have been efforts devoted to dependency parsing for the spoken domain, especially for (a subset of) the Switchboard corpus (Godfrey et al., 1992), which contains transcripts of telephone conversations in English. While some focused on parsing the full subset (Yoshikawa et al., 2016; Rasooli and Tetreault, 2013; Miller and Schuler, 2008), others attended to specific phenomena common in spoken data, such as speech repairment (Miller, 2009b,a). The Switchboard corpus was manually annotated only for constituency parses; the prior work relied on dependency parses automatically converted from constituency parses without manual verification.

In addition to English, dependency treebanks have been developed for the spoken domain of other languages, including French (Gerdes and Kahane, 2009; Bazillon et al., 2012), Czech (Mikulová et al., 2017), Russian (Kovriguina et al., 2018), Japanese (Ohno et al., 2005), and Mandarin Chinese (He et al., 2018). These treebanks, including Switchboard, however, were not always built using UD guidelines. In addition, the customized annotations and trained parsers are not always publicly available, making it less straightforward to carry out evaluation, especially since most dependency parsers are designed for UD-formatted treebanks.

This paper presents a wide-coverage dataset of spontaneous child-parent interactions (MacWhinney, 2000) annotated with syntactic dependencies largely following the UD standards. Our thorough annotation guidelines will be useful in the development of any spoken language dependency treebank.

Our work goes beyond prior work in several respects. First, compared to most of the other spoken dependency treebanks, which contain conversations between adults (Bechet et al., 2014; Dobrovoljc and Martinc, 2018; Dobrovoljc and Nivre, 2016a) or user-generated content (Davidson et al., 2019), our annotation guidelines attend to child and

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1734–1753, 2023. https://doi.org/10.1162/tacl_a_00624
Action Editor: Afra Alishahi. Submission batch: 5/2023; Revision batch: 9/2023; Published 12/2023.

child-directed spoken language. Second, in contrast to other spoken treebanks in the UD project (see Dobrovoljc (2022)), our dataset as a whole is of considerable size. In our corpus, there are 26,098 child utterances (116,428 words), and 18,646 parent utterances (117,479 words). Third, while there are some dependency corpora of child-parent interactions in English (Sagae et al., 2010), Japanese (Miyata et al., 2013), and Hebrew (Gretz et al., 2013), they include data from *only one or two children*, and detailed annotation guidelines pertaining to (child) spoken language are sometimes lacking. In this study, we provide annotations for utterances of 10 children across a much wider age range, therefore covering in more detail lexical and syntactic phenomena that are more common in child language (e.g., repetition, speech disfluency) and spoken language more broadly.

With this dataset, we ask two additional questions: (1) How do state-of-the-art dependency parsers trained on out-of-domain data perform on naturalistic spoken language of different interlocutors? To address this question, we evaluate parsers trained on three genres within the written domain: news texts, tweets, and learner data, all in English. (2) What is the relationship, if any, between parser performance and the developmental stage of the child? One might expect a positive correlation between the two, with the expectation that as the child continues to develop their language skills, they will utilize more and more cohesive syntactic structures, instead of, for example, producing grammatical errors, unintelligible speech, or word omissions. On the other hand, it is possible that parser performance might increase as the child reaches a certain developmental stage, then start decreasing, since the child might start producing sentences with more complex or expressive syntactic structures, which are potentially harder to analyze.

2 Why a *Dependency* Treebank?

We chose UD-style dependency annotations because of their general ease of adaptability to different domains or languages, as well as the continuous active development of different dependency parsing toolkits. Dependency parsing has recently attracted more attention than constituency parsing. Compared to constituency structures, dependency structures are ''simpler'', with easier adaptations to ''non-standard'' domains or typologically diverse languages. The popularity of dependency parsing can be largely attributed to the research community's considerable efforts into developing the UD project and the CoNLL shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007). Accordingly, tools for building dependency parsers (van der Goot et al., 2021; Qi et al., 2020; Varab and Schluter, 2019) are well-maintained; as new treebanks become available and are added to UD, new implementations will continue to become accessible. Developers can evaluate their parser models using our dataset to probe their parsers' capabilities.

Second, from the perspective of studying child language, previous literature has shown that linguistic features derived from dependency parses can be leveraged to replace traditional metrics such as the Index of Productive Syntax (Scarborough, 1990) and measure children's syntactic development in an automatic fashion (Lubetich and Sagae, 2014; Lu, 2009; Sagae et al., 2005, 2007). Others have utilized dependency trees to study how children produce syntactic alternations along their developmental trajectory (Liu and Wulff, 2023). These prior experiments, however, used parsers trained either on data from one or two children, or on fully written data; therefore the generalizability of the prior findings remains to be verified. Providing a dependency treebank (even with differences in annotation details) for data from more children and a wider age range will allow researchers in child language development to train parsers of wider coverage with good performance. The parsers can then be applied to unannotated child-parent interactions.

Third, while this work focuses on English, there are some corpora of child-parent interactions in other languages that also provide sufficient longitudinal data, though not on the scale of the English sections (e.g., the Hebrew Berman Longitudinal Corpus [Armon-Lotem and Berman, 2003]). We hope that our work can serve as a useful reference.

3 Related Work

Earlier work on dependency parsing of child-parent conversations (Sagae et al., 2001, 2004) focused on the Eve corpus (Brown, 1973) from CHILDES (MacWhinney, 2000), though the annotations did not follow UD guidelines. Their

annotations focused on structures that are also observed frequently in written data (e.g., subject, object, and sentential complements) rather than providing details about patterns specific to spoken language (e.g., disfluency, non-subject word omission). Subsequent research extended these annotation guidelines (Sagae et al., 2010) to child and child-directed spoken language in Japanese (Miyata et al., 2013) and Hebrew (Gretz et al., 2013).

We note two studies that carried out UD-style dependency annotations for child and/or child-directed spoken language. Liu and Prud'hommeaux (2021) took a semi-automatic approach to convert a subset of the existing dependency parses from the Eve corpus (Brown, 1973) to UD standards, with a focus on 18- to 27-month-old children. Concurrent work by Szubert et al. (2021) annotated dependency parses for two languages: English (from the Adam corpus [Brown, 1973]) and Hebrew (The Hagar corpus [Berman, 1990]), although they only looked at *child-directed utterances*. The annotations from Szubert et al. (2021) cover some phenomena prevalent in child spoken language, such as non-standard vocabulary and ambiguous fragments (with which our annotation standards align), but they do not include guidelines for annotating possible lexical omission or speech restart and repair, which we address.

When considering the annotations of the dozen spoken dependency treebanks in UD more broadly, they appear to mostly focus on fillers, discourse particles, and disfluency (Dobrovoljc, 2022; Kahane et al., 2021); in some cases, detailed information is lacking largely due to the fact that the treebanks are relatively small (Braggaar and van der Goot, 2021; Partanen et al., 2018). Given that our dataset is on a much larger scale, we are able to carefully note different speech-related phenomena along with providing clear annotation guidelines.

While fillers and discourse markers are usually treated consistently across existing UD treebanks, there remain significant inconsistencies in head-attachment and the (over-)application of treebank-specific dependency relations. As pointed out in Dobrovoljc (2022), while several treebanks rely on the *reparandum* dependency relation defined in UD to annotate disfluency, they do not all follow the UD guidelines which suggest applying *reparandum* right-to-left (i.e., treating the repairment as the syntactic head of the prior disfluent words) (Kahane et al., 2021). Other work introduces new treebank-specific dependency relation subtypes that do not abide by UD standards for clause discourse markers (e.g., using *parataxis:discourse* instead of *discourse* when clause-level discourse markers are automatically identifiable) or speech restart (e.g., using *parataxis:restart* instead of an augmentation to *reparandum*) (Dobrovoljc and Nivre, 2016b).

Here, our dataset adheres to the UD guidelines as much as possible. When introducing new dependency subtypes for speech repairment, restart, and repetition, we do so by extending *reparandum*. These subtypes also make it straightforward for automatic extraction of different structures as well as converting our annotations when necessary to make them better aligned with the customized annotations of other (spoken) treebanks.

4 Meet the Data

The dataset consists of transcripts of English naturalistic parent-child interactions from the CHILDES (MacWhinney, 2000) database, accessed through the `childes-db` interface (Sanchez et al., 2019). As we are interested in how parser accuracy changes at different developmental stages, we used *age* as a proxy for developmental stage and set 6-month intervals as bins. For each individual child, we calculated the total number of words produced by the child and by the parent(s) within each age bin of the child. Then, from each age bin, for both child and parent data, we randomly sampled a number of utterances that amounted to approximately 2,000 words; the criteria were relaxed in order to include data across a wide range of age bins. This resulted in spoken data of ten children from 6 corpora (Table 1).

5 Annotation

Our annotation guidelines largely followed those of UD (Zeman et al., 2022). Annotator A, with advanced training in dependency syntax, initially annotated data of age 18–24 months and 24–30 months from Abe and Sarah, as a way to take note of any domain-specific or challenging phenomena. These guidelines were discussed with annotator B and modified as needed. Then, given each age bin of every child, the two annotators annotated 10% of the data from both child and parent utterances. We calculated agreement scores using Cohen's

Child	Corpus	18–24	24–30	30–36	36–42	42–48	48–54	54–60	60–66
Abe	Kuczaj		2,007	2,007	2,022	2,020	2,036	2,021	1,386
Parent	(Kuczaj II, 1977)		2,028	2,040	2,025	2,020	2,017	2,019	–
Adam	Brown		2,012	2,004	2,016	2,021	2,012	2,026	2,025
Parent	(Brown, 1973)		2,012	2,009	2,020	2,008	2,015	2,007	1,234
Sarah	Brown		2,014	2,008	2,011	2,007	2,020	2,015	2,028
Parent			2,019	2,043	2,037	2,036	2,026	2,032	2,049
Thomas	Thomas		1,969	1,965	1,972	1,984	2,009	1,992	
Parent	(Lieven et al., 2009)		2,013	2,035	2,050	2,016	2,025	2,010	
Emma	Weist			2,007	2,014	2,029	2,025		
Parent	(Weist and Zevenbergen, 2008)			2,020	2,038	2,015	1,455		
Roman	Weist		1,980	1,975	1,990	2,004	1,999	2,009	
Parent			1,712	2,017	2,012	2,020	2,013	2,027	
Laura	Braunwald	1,930	1,925	1,955	1,912	1,956	1,943	1,954	
Parent	(Braunwald, 1985)	2,022	2,019	2,013	1,991	2,023	1,984	2,015	
Violet	Providence	1,730	1,855	1,889	1,886	1,894			
Parent	(Demuth et al., 2006)	2,007	1,995	2,025	2,044	2,030			
Naima	Providence	1,748	1,752	1,814	1,902	1,943			
Parent		2,019	2,007	2,005	2,032	2,019			
Lily	Providence	2,004	1,788	1,908	1,981	1,957	1,161		
Parent		1,996	2,016	1,999	2,038	2,006	2,000		

Table 1: Number of words for child and parent data at different age ranges (in months) of the children.

Kappa (Artstein and Poesio, 2008). The overall agreement score taking into account all syntactic head and dependency relation annotations is 0.97; the average agreement score across each dependency parse is 0.96. (Agreement scores for each child were around 0.97.) Final annotations were performed and verified by annotator A.

Here we describe in detail our approach to transcription orthography, tokenization, and dependency annotations for syntactic constructions that are unique to or more common in child production and spoken data more broadly.

5.1 Orthography and Tokenization

Regarding orthography of the transcripts, we made four decisions, all of which are on the basis of a principle that we call ''annotate what is actually there''. First, we did not perform orthographic normalization of most intelligible words in the data (e.g., *she wanna eat*); in other words, these words stayed true to their original forms taken from CHILDES. That said, the tokenization of certain cases was updated following UD. These cases include: (1) possessives (e.g., *Daddy's* → *Daddy 's*); (2) contractions (e.g., *I'm eating* → *I 'm eating*; *don't* → *do n't*); (3) combined conjunctives (e.g., *in_spite_of* → *in spite of*); (4) combined adverbs (e.g., *as_well* → *as well*); (5) other informal contraction (e.g., *gonna* → *gon na*); (6) childish expressions (e.g., *poo_poo*, *choo_choo*).

Second, unintelligible speech tokens were removed as it is not possible to know the number of words in the unintelligible region, whether the words were intentional, or in which syntactic role the unintelligible content might serve. Third, we preserved initial capitalization in the transcripts since typically only proper names were capitalized. Lastly, we omitted all punctuation except for apostrophe (to abide by UD standards) since punctuation marks tend to not be explicitly articulated in spontaneous spoken language.

5.2 (Vague) Utterance Boundaries

Most conversational turns in CHILDES correspond to individual utterances; we therefore tried to annotate each one as a stand-alone sentence. That said, the initial utterance boundaries are not always adequate. This means that one conversational turn can, in some cases, be considered to have ''side-by-side'' sentences (Figure 1)[1]; we

[1]Examples presented in this paper are often modified from the original utterances for ease of presentation.

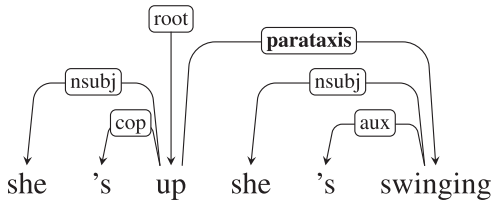


Figure 1: An example *(vague) utterance boundary*.

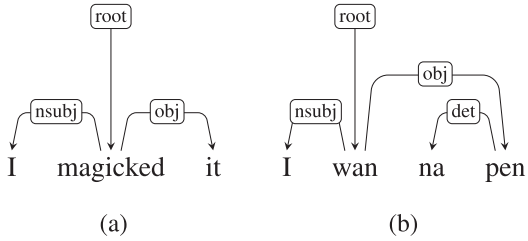


Figure 2: Examples of *creative lexical usage*.

followed the instructions of UD and annotated the first sentence to be the root; then later sentences in the instance were treated as *parataxis* of the root.

5.3 Creative Lexical Usage

Children commonly make lexical choices that do not necessarily follow the standards of parent production or (formal) written data (e.g., *mine pillow*). On the other hand, these cases may reflect the child's world since they may capture the child's own understanding of these words and their lexical (and syntactic) development. Therefore here we refer to such cases as *creative usage*; for each case, we analyzed their syntactic usage given the remaining structure of the sentence (Lee et al., 2017; Santorini, 1990), then assigned dependency parses accordingly. For example, in Figure 2a, the word *magicked* is creatively used as a verb that links the subject *I* and object *it*.

In some instances, it is relatively difficult to decide whether an utterance contains the child's creative usage of some lexical item or a potential transcription error inserted by the annotator. Given that transcribing spoken data manually requires large amounts of time and energy, it is not against expectations that the resulting transcriptions might have errors. For instance, with Figure 2b, it is not exactly clear whether the child really said *I wan na pen* (keeping in mind that *wanna* is normally meant to correspond to *want to*), or whether the transcription should have been *I want a pen*. In our approach, we compared how often each alternative occurs in our dataset, then made the final decision. Therefore between the two alternatives above, we chose to annotate *na* as the determiner of *pen*.

5.4 *Possible* Lexical Omission

As we are taking a data-driven approach, we tried to perform annotations based on how words or phrases are used within an utterance; this means that we avoided assuming potential word omissions as much as possible. We assigned dependency structures to an utterance if a reasonable syntactic tree (in manual annotations; not referring to the automatic parse done by parsers) could be derived without the assumption that certain words are missing given the context.

In other cases, we deemed the utterance as having a lexical omission if the omitted word could be *automatically retrievable*; this way other researchers will be able to formulate a different analysis for the utterance as they see fit. In particular, we considered two types of lexical omissions. The first type is copula omission, where the syntactic head of the copula is mostly a noun (e.g., *I **girl***), or an adjective (e.g., *he **heavy***); other times we assumed that a copula is omitted if the utterance can be interpreted as an expletive structure (e.g., *there book*, with *there* as the *expl* dependent of *book*). The second type is adposition omission, mostly the adposition that is the function head of an oblique phrase (Figure 3a) or the infinitival-*to* in a complement clause (Figure 3b).

5.5 Nominal Phrases

For certain nominal phrases that serve as adverbial modifiers in a given utterance (e.g., Figure 4a), and/or express time and dates, we tried to annotate them more carefully using subtypes of specific dependency relations (e.g., *nmod:tmod* or *obl:tmod*) (Schneider and Zeldes, 2021). For example, in Figure 4b, depending on their respective role, *morning* should be an oblique phrase of *go* whereas *tomorrow* modifies *morning*.

5.6 Ambiguity

The syntactic structure of a sentence can be ambiguous when the sentence is considered in isolation. Therefore we took into account the surrounding context of an utterance when performing annotations. In some cases, context can be helpful; for example, in (1), *like* can be treated as the verb of the sentence, rather than an adposition.

(1) *Parent: do you like this*; *Child: like this*

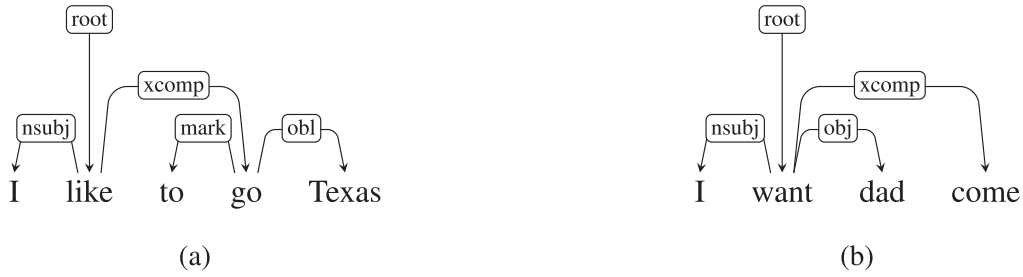


Figure 3: Examples of *possible lexical omission*: omission of adposition.

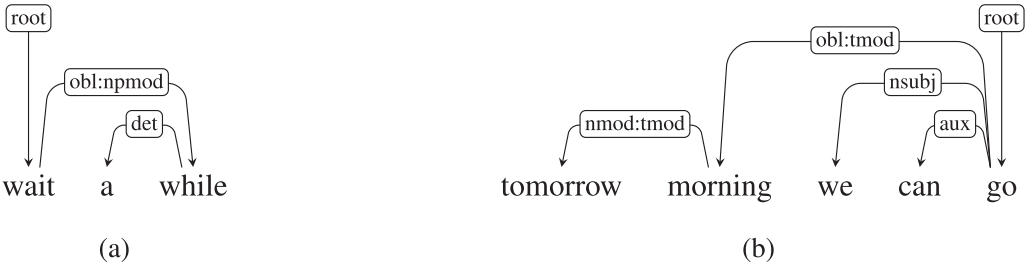


Figure 4: Examples of *nominal phrases*.

In other (rare) cases, context might not be useful; for instance, in (2), it is not clear whether *rain* should be a verb and the relation between the two words is *obl:tmod*, or a noun and the dependency relation is *nmod:tmod*. For these examples we opted for the simpler analysis given the characteristics of child utterances and treated *rain* as a noun.

(2) *Parent: eat your soup*; *Child: rain tonight*

Another source of ambiguity comes from whether to treat proper names or words like *mommy* and *daddy* as *vocative* or not, e.g., *Momma try it*. For these cases, we decided to consider them as *vocative* if this interpretation is reasonable, since subject omission is common in early child spoken language (Hughes and Allen, 2006).

5.7 Speech Repairment

For speech repair, which captures one type of disfluency (Ferreira and Bailey, 2004), we used *reparandum* as suggested by the UD guidelines, where the speech repair is the syntactic head of the subtree that constitutes the disfluent speech (e.g., *six* in Figure 5a). If the disfluent speech contains discourse fillers or editing terms (e.g., *um*), these elements are annotated as the syntactic dependents of the repair with the relation *discourse*, which also avoids unnecessary crossing dependencies. In some cases, the disfluency subtrees are word fragments that do not form a complete phrase (e.g., *grab the* in Figure 5b); for these cases, we used the principle of promotion to analyze elements within the subtree structure of the disfluency if needed. For example, with Figure 5b, the word *grab* is most likely to be the head within the disfluency subtree; therefore we promoted the following *the* to be the *object* of *grab*, then analyzed the dependency relations of the residual structures in the instance.

To separate repairment from speech restart or abandonment (Section 5.8), we categorized an utterance as having speech repairment only if the repairment is sentence-medial.

5.8 Speech Restart

Another type of disfluency is speech restart. We generally considered an instance as having speech restart or abandonment if the abandoned elements occur at the beginning of the instance and do not form a coherent phrase together; in addition, the abandoned elements need to be different from the speech restart. For these cases, given that speech restart falls broadly under the umbrella of disfluency, and in order to distinguish restart from repairment above, we extended *reparandum* with a new dependency relation subtype: *reparandum:restart* to connect the abandoned elements as the dependents of the speech restart (Figure 6). This way the dependency relation will also go ‘‘right-to-left’’ (Dobrovoljc, 2022), following the usage of *reparandum*.

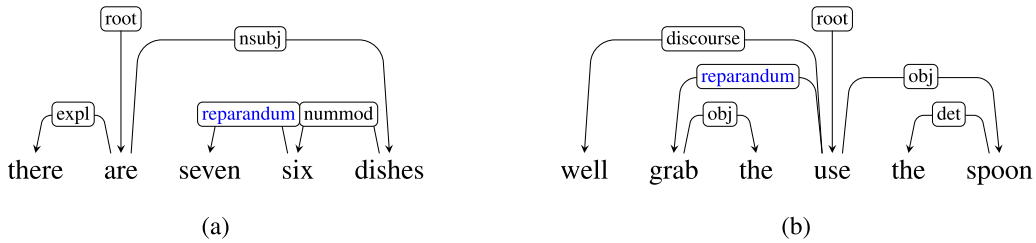


Figure 5: Examples of *repairment*; the dependency relation for speech repair is in blue.

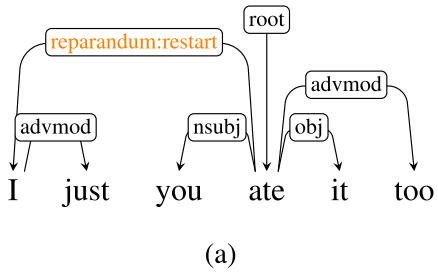


Figure 6: Examples of *restart*; the dependency relation between repeated elements is in orange.

5.9 Repetition

Overall we identify three major kinds of repetitions. For the first type, an utterance consists of repetitions of the same dependency subtree and the repeated subtree is a coherent phrase by itself. Examples include cases such as discursive repetition (e.g., *no no mommy*), onomatopoeia (e.g., *honk honk*), or repetition of other kinds of word or phrase (e.g., *this is my truck my truck*). For these cases, we treated the first appearance of the repeated subtree as the syntactic head with the following repetitions as the dependents connected with the relation *conj*. A special case is when the instance repeats a full sentence (or just containing a verbal phrase), e.g., *I did it I did it* (Figure 7a); for these examples we used *parataxis* to adhere to the annotations of side-by-side sentences noted by UD.

For the second type, repetition is used to emphasize the characteristics of certain objects or conditions (or serves as an intensifier [Szubert et al., 2021]); in these cases the repeated element is usually a single word with a part-of-speech (POS) tag of adjective or adverb. These cases were annotated similarly as those from the first type above (Figure 7b).

The third type of repetition pertains to disfluency. Whether to interpret an instance as a disfluent repetition is challenging when the instance does not have a corresponding audio to provide prosodic clues. Therefore to distinguish the two types, we considered an instance to have disfluent repetitions if the repetition appears at the beginning or in the middle of a single sentence; in addition, the repeated element must be (1) a series of word fragments that do not form a whole coherent phrase in that sentential context (Figure 7c); (2) or a single word whose POS tag is neither an adjective nor an adverb (Figure 7d). For these cases, to align with the fact that *conj* is usually applied left-to-right (i.e., syntactic head precedes its dependents), we again extended the usage of *reparandum* and applied a new dependency relation subtype, *reparandum:repetition*, to describe repetition in disfluency; speech repairment, restart, and disfluent repetition will hence be automatically distinguishable.

5.10 Other Structures

The last two types of structures are serial verb constructions (SVC), and tag questions. For SVC, we followed Szubert et al. (2021); for an utterance such as *he came see me play*, *see* was treated as the dependent of *came* and the relation is *compound:svc*. Tag questions are mostly used in parent utterances (e.g., *you like that book do n't you*); for these examples we abided by the UD annotations and used *parataxis* to connect the tag question to the main clause of the utterance.

6 Experimental Design

We now turn to evaluating dependency parsing with our dataset. We aim to address: (1) How *well* do competitive parsing architectures trained on out-of-domain data perform for spontaneous child-parent interactions? (2) Can in-domain data help with parsing performance for child speech, and if so, to what extent? (3) Are there any generalizable relationships between parser performance and child production at different developmental stages? To that end, we explored different parser

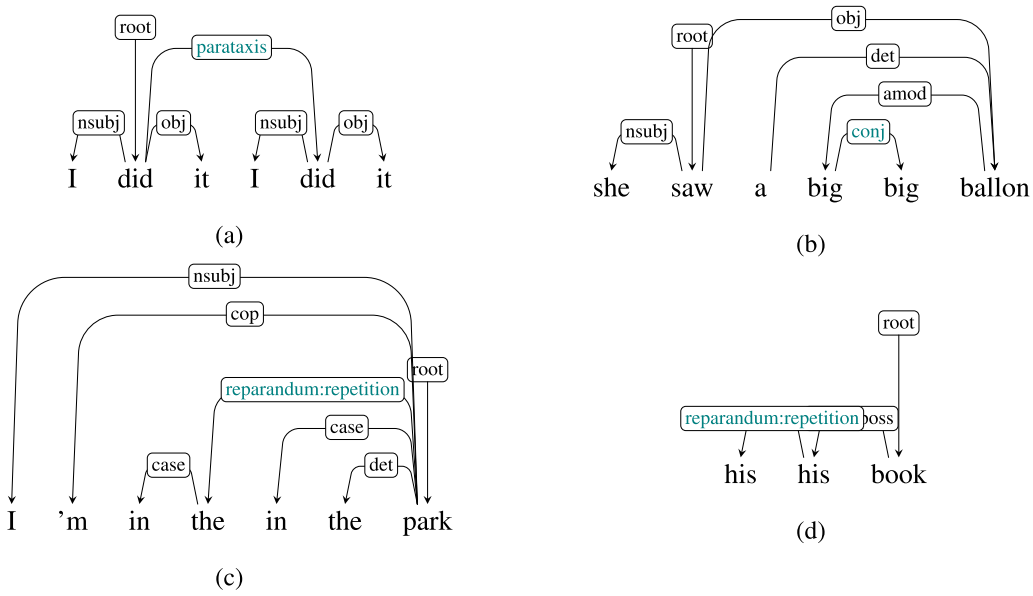


Figure 7: Examples of *repetition*; the dependency relation between repeated elements is in teal.

architectures, several dependency treebanks from a range of domains distinct from one another and from child-parent interactions, and different training schemes for in-domain experiments.[2]

Parsers We used two graph-based parsers, `Diaparser` (Attardi et al., 2021), `MaChamp` (van der Goot et al., 2021), and one transition-based parser, `UUParser` (de Lhoneux et al., 2017; Smith et al., 2018). We picked these parsers based on the level of detail in their documentation and implementation, as well as their availability. The goal of exploring different architectures here is to see whether qualitative observations (see Section 7) of parsing results hold regardless of the particular parser applied. For each of two graph-based parsers, we created different variants via combining with different language models (LMs) as encoders (`bert-base-cased` [`bert`] [Devlin et al., 2018], `roberta-base` [`roberta`] [Liu et al., 2019], and `twitter-roberta-base` [`twitter`] [Barbieri et al., 2020]). Here we mostly focus on results derived from `MaChamp`, which was able to achieve the best results on the out-of-domain data (Table 2); in addition, there were no noticeable differences in the overall patterns discussed in Section 7 across different parsers. We note that the goal of this work is not to determine which parser is the best for our dataset. Rather, we are exploring performance of a reasonably good parser across a variety of different training data configurations.

[2]Data and scripts are avaiable at: `https://github.com/ufcompling/spoken_parsing`.

Data	Parser	LM	LAS
EWT	MaChamp	`bert`	90.49
		`roberta`	91.27
		`twitter`	90.84
	Diaparser	`bert`	89.89
		`roberta`	88.40
		`twitter`	88.26
	UUParser	–	82.68
Tweebank	MaChamp	`bert`	79.57
		`roberta`	80.49
		`twitter`	80.26
	Diaparser	`bert`	80.01
		`roberta`	79.48
		`twitter`	79.65
	UUParser	–	69.54
ESL	MaChamp	`bert`	91.70
		`roberta`	92.27
		`twitter`	91.94
	Diaparser	`bert`	91.48
		`roberta`	90.74
		`twitter`	90.60
	UUParser	–	86.57

Table 2: Evaluation for out-of-domain datasets.

Out-of-domain Data We used three out-of-domain datasets of written English: (1) UD_English-EWT

(EWT) (Silveira et al., 2014), which contains texts from web media; (2) UD_Tweebank (Tweebank) (Liu et al., 2018), which consists of English language tweets; and (3) UD_English-ESL (ESL) (Berzak et al., 2016), which contains L2 English learner writing samples (Yannakoudakis et al., 2011). For each of the aforementioned datasets, we trained the parsers described in Section 6 with their default parameters. We calculated micro unlabeled attachment scores and labeled attachment scores (LAS) to evaluate parser performance. Throughout this paper we focus on reporting LAS. Parser evaluation results across three random seeds for all out-of-domain datasets are reported in Table 2.

7 Out-of-domain Experiments

We first performed automatic part-of-speech tagging for all child and parent data using the open source NLP library Stanza (Qi et al., 2020). We then applied each of the out-of-domain parsers to child and parent utterances within each 6-month age range of the child; parser performance was again indexed by LAS across 3 random seeds. We foresee two possible directions regarding the parsing results. On one hand, EWT has significantly more data than Tweebank and ESL, which might lead to overall better performance. On the other hand, among the three out-of-domain datasets, the domains of Tweebank and ESL are possibly more relevant or more similar to child-parent interactions, in the sense that they are less ''formal''. Additionally, the written texts of ESL may contain errors, which may bear similarity to child data, as children are also language learners. This potentially means that parsers trained from Tweebank or ESL might outperform those based on EWT. We note that the evaluation of parsers trained on out-of-domain data did not consider new dependency relation subtypes that were introduced in our annotations, which are *reparandum:restart* and *reparandum:repetition*.

7.1 Parent Data

On average, for all parents across different age ranges of their children, parsers trained on the out-of-domain ESL dataset with the `twitter` LM achieved the best result (83.88). By contrast, the best parsers from EWT, which were also trained with the `twitter` LM, performed noticeably worse; the average difference in LAS ranges from 2.11 in the age range of 48–54 months, to 3.95 for 18–24 months. On the other hand, the best parsers from Tweebank, trained with `bert`, achieved comparable performance (83.48) to the best parsers trained on the ESL dataset. This is noteworthy since Tweebank contains around 1/3 of the quantity of data in ESL (and less than 1/8 of the amount in EWT). In addition, the variability in performance across the different parsers trained on Tweebank is smaller than those trained on ESL.

So what might be the source of variability in performance for parsers trained from different out-of-domain treebanks, especially during early ages of the children? Comparing the best performing parsers trained on EWT and those from Tweebank and ESL, we see that for some utterances where the copula takes the form of ''*s*', parsers trained from EWT erroneously annotated the copula as the subject, assigning two subjects to the same syntactic head; this accounts for around 17.07% of all errors. This raises the worrisome question of why a structure where the syntactic head has two subjects would arise. The most plausible answer is that such structures exist in the training data. We found five such sentences in the EWT dataset. In these cases, the first subject was incorrectly annotated as headed by the verb of the subordinate clause at a lower level (e.g., in ***it*** *is not about how much* ***you*** *earn* (adapted), *it* was annotated as the subject of *earn*). Similarly, in cases where ''s' has a dependency relation of *aux*, the EWT-trained parsers also tended to parse it as *nsubj*; this accounted for 7.69% of the errors.

Let us now turn to the question of where the best-performing parsers trained from out-of-domain treebanks fall short. We note four cases here. For parent utterances in the early child age ranges (18–30 months), the dependency relation that results in the biggest discrepancy between parser and manual annotations is *nmod:poss* (13.05% on average; e.g., *my book*), where the parsers annotated the relation as *nmod* (97.35%), which is less preferred in the latest UD guidelines. The second case that caused confusion for the parsers is when sentences contain elements that should be annotated as *discourse* and/or *vocative* (7.69%; e.g., *hahaha I see*, *Roman oh that is beautiful*) but the parsers more consistently annotated the first word in these instance as the root of the sentence (55.08%).

One other noteworthy example is when the parsers think of a *vocative* as the subject (23.22%;

e.g., *Adam eat your soup*) or sometimes the object (11.78%; e.g., *sit Sarah*), since there is no punctuation in the annotations. The last one is *conj*, which we used for repetition or when the speaker appears to be listing individual nouns (e.g., *orange grape apple*); however, the parsers annotated the relation between pairs of nouns to be *compound* (9.05%), which was common in the out-of-domain treebanks. As the children age, the patterns described above still turned out to be the main explanations for parser errors, though to lesser extent.

7.2 Child Data

Overall parser performance for child data follows the patterns observed for parent data. Across all age ranges, the best parsers trained on ESL using `twitter` and those trained on Tweebank with `bert` yielded comparable performance (79.35 vs. 79.39). These parsers also outperformed the best parsers using the EWT treebank (76.48), despite the much larger size of EWT.

The types of dependency relations that resulted in discrepancies between parser output and manual annotations in child data are also similar to those noted for parent utterances, especially during children's early ages. For instance, when an utterance consists of two words, where the first was annotated as the subject of the second (*Adam home*), the parsers again preferred to label *Adam* as the *compound* of *home* (13.50%). The relation *vocative* was again sometimes annotated by the parsers as *nsubj* (12.32%) or *obj* (14.53%). Notice the small fluctuation in LAS in the 48–54 month age range. Here, we see parser errors that involve a head verb with an adverbial modifier, such as *bring them back*; rather than parsing *back* as *advmod* of *bring*, the parsers assigned *aux*.

In all, perhaps not surprisingly, parser performance is better overall for the parents. That said, when children reach later ages, the parsers' performance ($\sim$87.83) approaches that for parent utterances ($\sim$89.13), suggesting that children's acquisition of syntactic structures is becoming more parent-like. While parser scores for parent data slowly increase between the age range of 18–66 months, this progress is much more pronounced for child utterances; in other words, we do see an overall improvement of parser performance as children progress along the syntactic developmental trajectory, particularly within 18–48 months.

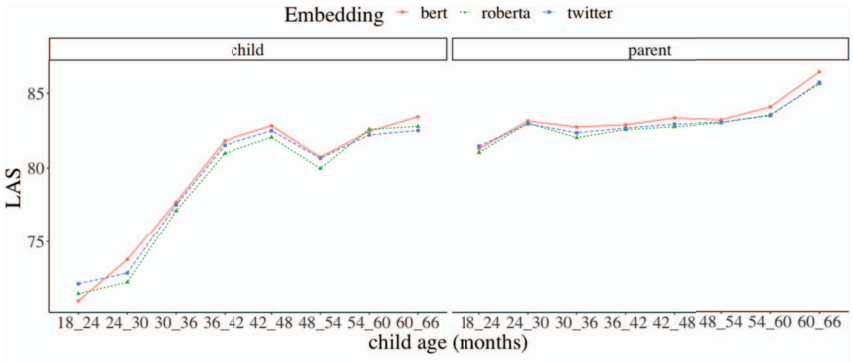
8 In-domain Experiments

In this section, we perform parser training with in-domain data of child-parent interactions. As seen in the results from Section 7, on average, parsers trained with MaChamp using the `twitter` LM seemed to achieve the most stable performance; we therefore adopted that same parser setup here.

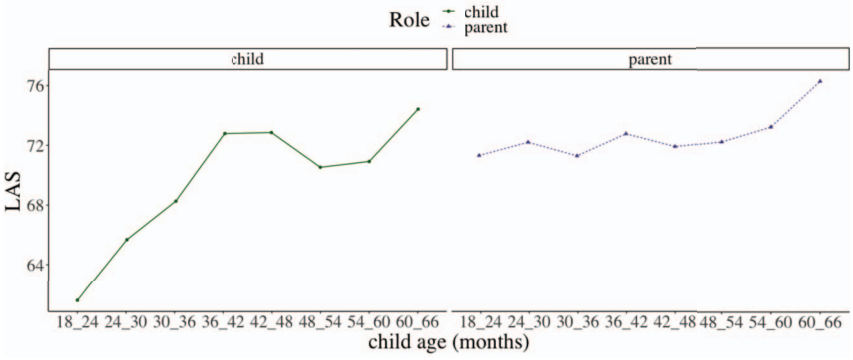
Our training scheme is as follows. Suppose that we wanted to evaluate parser performance for Adam (Brown Corpus). We first trained parsers using all the data from the other nine child-parent pairs; we then measured parser performance for both Adam's and his parents' data at each of Adam's available age ranges using LAS averaged across 3 random seeds.

Overall the trends of the LAS scores from in-domain training (Figure 9) are similar to those obtained from out-of-domain evaluation (Figure 8): As the children develop, the LAS scores increase. In contrast, the results for parent data appear to be much more stable across the age span of the children. (Note, however, that the actual numerical values of in-domain evaluation are not directly comparable to those of out-of-domain evaluation, given that the latter did not take into account new dependency subtypes that we introduced for speech disfluency.) For child data, performance ranges from 83.30 to 96.25, while for parent utterances, performance spans from 93.23 to 97.64. In early developmental stages (24–36 months), the parsers for child data seemed to be confused the most by utterances of 2–4 words, which potentially involve word omission (e.g., *are five*), discourse markers, or vocative elements. In these cases, the parsers tended to mis-analyze the root, preferring to treat shorter utterances as compound noun phrases and to assign the last words to be the root.

In addition, for child data specifically, we also explored another training scheme, where we trained parsers on the combined data produced by all 10 parents, then applied the parsers to the data of each individual child. Observations derived from this training scheme are qualitatively similar to those derived from the training scheme described above. To add statistical rigor when comparing results from the two different in-domain training settings, we combined data from all children, then measured the correlation



(a) MaChamp



(b) UUParser

Figure 8: Evaluation of out-of-domain Tweebank parsers (MaChamp and UUParser); at each age range, the parser score was averaged across all children (or parents) within that range.

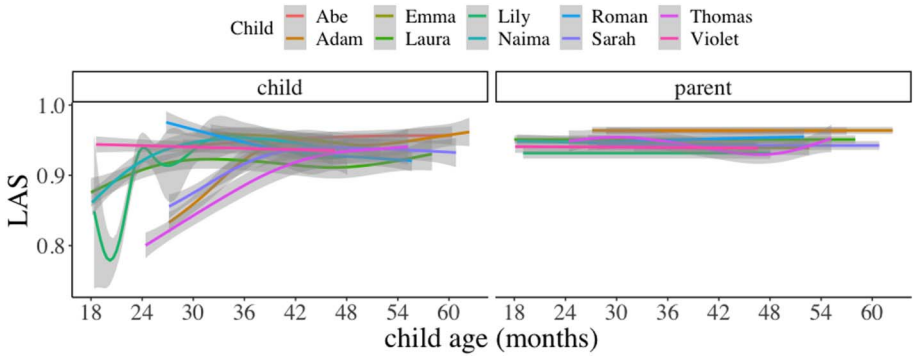


Figure 9: Evaluation of in-domain parsers trained with `twitter` LM, using child age as the index of developmental stage; to evaluate parser performance for a given child, the parsers were trained using the data from all the other nine child-parent interactions.

between the scores derived from the second training setting and those from the first. In particular, we used Spearman's ρ, which assesses if there is any monotonic (not necessarily linear) relationship between the two variables. Our results indicate that the observations from the two training setups are quantitatively comparable as well (Spearman's $\rho = 0.67$).

On average across the data of all children, speech disfluency accounted for very small proportions of all dependency relations (repairment: 0.30%, restart: 0.48%, repetition: 0.08%), and the proportions are comparable when considering the data of all 10 children (repairment: 0.35%, restart: 0.78%, repetition: 0.10%). Between the three relation subtypes of disfluency, the parsers performed the best for repetition, though they tended to analyze the disfluent words or phrases as the syntactic heads of the speech repairment and restart.

9 Notes on Developmental Index

Our analysis thus far has largely relied on child age as the index of developmental stage. In this

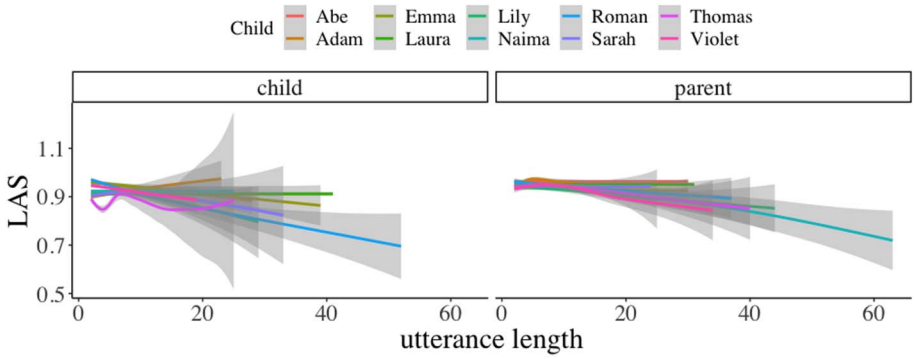


Figure 10: Evaluation of in-domain parsers trained with `twitter` LM, using utterance length as the index of developmental stage; in this case to evaluate parser performance for the data of a given child, the parsers were trained using the data from all the other nine child-parent interactions.

section, we looked into using utterance length as an alternative developmental index of child language production, with longer utterance length corresponding to later developmental stage. To that end, based on the results derived from the first in-domain training scheme from Section 8, we broke down parser performance given different utterance lengths for both child and parent speech. As illustrated in Figure 10, for child production, LAS scores decrease gradually as utterance lengths increase.

This observation, nevertheless, seems somewhat in opposition to the earlier patterns where parser performance appears to increase as children age, given that age and utterance length maintain a positive correlation with each other (Brown, 1973). This positive relationship is also evident in the children's data in our dataset; when fitting a linear regression model predicting utterance length as a function of age, the model yielded a significantly positive coefficient value for age ($\beta = 0.82, p < 0.001$).

To tease apart the respective effect of age and utterance length, we resorted to mixed-effect linear regression. For each utterance produced by children in our dataset, we calculated the following metrics: sentence length (McDonald and Nivre, 2011), dependency length (Ferreira and Bailey, 2004; Anderson et al., 2021), and LAS score. In the mixed-effect model (after step-wise backward regression), we included LAS score as the outcome variable, meanwhile adding age, sentence length, dependency length, and the training data size for the parsers; we included interactions between each pair of the aforementioned four fixed effects, as well as children as random effects.

Based on the mixed-effect regression results, age appears to have a significantly positive influence on LAS score ($\beta = 0.03$ (0.02, 0.04)), meaning that parser performance improves as age increases. On the other hand, utterance length seems to have a significantly negative effect on parser performance ($\beta = -0.007$ (-0.09, -0.005)), which aligns with observations from Figure 10. Perhaps a bit surprisingly, dependency length turns out to also play a positive role in parser performance, albeit only weakly ($\beta = 0.002$ (0.001, 0.002)). Comparing the coefficient values of the three factors, the role of age is much more pronounced than that of utterance length or dependency length. In particular, we attribute the relatively much weaker effect of utterance length to the fact that around 95.22% utterances of children in our data contain fewer than 10 tokens. (Note that this is a result from our data-driven approach to building the dataset originally rather than cherry-picking utterances from CHILDES.) We take these observations as indications that there does exist a relationship between children's developmental stage and parser performance, with the latter mostly affected by children's age while modulated by their utterance lengths.

10 Discussion and Conclusion

We present a wide-coverage dataset of child-parent interactions annotated with syntactic dependencies, along with detailed annotation guidelines extending the Universal Dependencies project. Evaluations from graph-based and transition-based dependency parsers with varying hyperparameters demonstrate that parsers trained using a relatively small amount of English tweets (Tweebank) are able to yield comparable or even

superior performance to parsers trained from much larger dependency treebanks. In addition, we observed the general trend that on average, parser performance increases as the children reach older ages, indicating that as children progress along their syntactic developmental trajectory, they start producing structures that are more cohesive but not too complex for the parsers to handle. The relationship between parser performance and children's age, as shown from our mixed-effect modeling, is modulated by utterance length, which has a significant yet quite weak and negative effect on LAS scores.

It is our hope that this dataset and the trained parsers will be useful for researchers studying both child language development and linguistic characteristics of spoken data more broadly. While in this work we adopted the UD annotation styles, prior work has also given guidelines for annotating spontaneous speech using constituency trees. The most notable example is the Switchboard corpus (Godfrey et al., 1992), which also provides annotations for speech restart, repair, word fragments, and so on. We do not wish to claim that our annotation guidelines are in any way *better* than those for Switchboard; rather, given that the CHILDES database contains child-parent interactions in different languages, we hope that our work here will motivate future relevant research that develops syntactic treebanks for child speech in languages other than English (see also Miyata et al., 2013; Gretz et al., 2013), potentially with wider coverage of children's age span. We started with English in this study because of data availability; with the ever growing size of CHILDES, others can utilize or adapt our annotation standards to different languages in order to facilitate research on crosslinguistic child language development.

Acknowledgments

We would like to thank our TACL action editor, Afra Alishahi, as well as the anonymous reviewers for their thoughtful feedback. We would also like to thank Parker Robbins for his efforts in the initial round of annotations.

References

Mark Anderson, Anders Søgaard, and Carlos Gómez-Rodríguez. 2021. Replicating and extending ''Because their treebanks leak'': Graph isomorphism, covariants, and parser performance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1090–1098, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.138

Sharon Armon-Lotem and Ruth A. Berman. 2003. The emergence of grammar: Early verbs and beyond. *Journal of Child Language*, 30(4):845–877. https://doi.org/10.1017/S0305000903005750, PubMed: 14686087

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596. https://doi.org/10.1162/coli.07-034-R2

Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine dependency and semantic graph parsing for enhanced universal dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 184–188, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.iwpt-1.19

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Thierry Bazillon, Melanie Deplano, Frederic Bechet, Alexis Nasr, and Benoit Favre. 2012. Syntactic annotation of spontaneous speech: Application to call-center conversation data. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1338–1342, Istanbul, Turkey. European Language Resources Association (ELRA).

Frederic Bechet, Alexis Nasr, and Benoit Favre. 2014. Adapting dependency parsing to spontaneous speech for open domain spoken language

understanding. In *Fifteenth Annual Conference of the International Speech Communication Association*. https://doi.org/10.21437/Interspeech.2014-39

Ruth A. Berman. 1990. On acquiring an (S)VO language: Subjectless sentences in children's Hebrew. *Linguistics*, 28(6):1135–1166. https://doi.org/10.1515/ling.1990.28.6.1135

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1070

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine. Association for Computational Linguistics.

Susan R. Braunwald. 1985. The development of connectives. *Journal of Pragmatics*, 9(4):513–525. https://doi.org/10.1016/0378-2166(85)90019-0

Roger Brown. 1973. *A First Language: The Early Stages*. Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City. Association for Computational Linguistics. https://doi.org/10.3115/1596276.1596305

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. https://doi.org/10.18653/v1/D19-1162

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, 49(2):137–173. https://doi.org/10.1177/00238309060490020201, PubMed: 17037120

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Kaja Dobrovoljc. 2022. Spoken language treebanks in Universal Dependencies: An overview. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Kaja Dobrovoljc and Matej Martinc. 2018. Er . . . well, it matters, right? On the role of data representations in spoken language dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 37–46, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6005

Kaja Dobrovoljc and Joakim Nivre. 2016a. 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).

Kaja Dobrovoljc and Joakim Nivre. 2016b. 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).

Fernanda Ferreira and Karl G. D. Bailey. 2004. Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, 8(5):231–237. https://doi.org/10.1016/j.tics.2004.03.011, PubMed: 15120682

Kim Gerdes and Sylvain Kahane. 2009. Speaking in piles: Paradigmatic annotation of French spoken corpus. In *Fifth Corpus Linguistics*

Conference, pages 1–15, Liverpool, United Kingdom.

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. https://doi.org/10.1109/ICASSP.1992.225858

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-demos.22

Shai Gretz, Alon Itai, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2013. Parsing Hebrew CHILDES transcripts. *Language Resources and Evaluation*, 49(1):107–145. https://doi.org/10.1007/s10579-013-9256-x

Ruifang He, Yaru Wang, Dawei Song, Peng Zhang, Yuan Jia, and Aijun Li. 2018. A dependency parser for spontaneous Chinese spoken language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17(4):1–13. https://doi.org/10.1145/3196278

Mary Hughes and Shanley Allen. 2006. A discourse-pragmatic analysis of subject omission in child English. In *Proceedings of the 30th Annual Boston University Conference on Language Development*, volume 1, pages 293–304. Cascadilla Press, Somerville, MA.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Liubov Kovriguina, Ivan Shilin, Alina Putintseva, and Alexander Shipilo. 2018. Multilevel annotation in the corpus for parsing Russian spontaneous speech. In *Speech and Computer*, pages 311–320, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-99579-3_33

Stan A. Kuczaj II. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600. https://doi.org/10.1016/S0022-5371(77)80021-2

John Lee, Herman Leung, and Keying Li. 2017. Towards Universal Dependencies for learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71, Gothenburg, Sweden. Association for Computational Linguistics.

Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2017. Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *Proceedings of the The 15th International Conference on Parsing Technologies (IWPT)*. Pisa, Italy.

Elena Lieven, Dorothé Salomo, and Michael Tomasello. 2009. Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–508. https://doi.org/10.1515/COGL.2009.022

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. https://doi.org/10.18653/v1/N18-1088

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zoey Liu and Emily Prud'hommeaux. 2021. Dependency parsing evaluation for low-resource spontaneous speech. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 156–165, Kyiv, Ukraine. Association for Computational Linguistics.

Zoey Liu and Stefanie Wulff. 2023. The development of dependency length minimization in early child language: A case study of the dative alternation. In *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)*, pages 1–8, Washington, D.C. Association for Computational Linguistics.

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1):3–28. https://doi.org/10.1075/ijcl.14.1.02lu

Shannon Lubetich and Kenji Sagae. 2014. Data-driven measurement of child language development with simple syntactic templates. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2151–2160, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk. Transcription Format and Programs*, volume 1, Psychology Press.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating Dependency Parsers. *Computational Linguistics*, 37(1):197–230. https://doi.org/10.1162/coli_a_00039

Marie Mikulová, Jiří Mírovský, Anja Nedoluzhko, Petr Pajas, Jan Štěpánek, and Jan Hajič. 2017. Pdtsc 2.0 - Spoken corpus with rich multi-layer structural annotation. In *Text, Speech, and Dialogue*, pages 129–137, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-319-64206-2_15

Tim Miller. 2009a. Improved syntactic models for parsing speech with repairs. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 656–664, Boulder, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/1620754.1620850

Tim Miller. 2009b. Word buffering models for improved speech repair parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 737–745, Singapore. Association for Computational Linguistics. https://doi.org/10.3115/1699571.1699609

Tim Miller and William Schuler. 2008. A unified syntactic model for parsing fluent and disfluent speech. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 105–108, Columbus, Ohio. Association for Computational Linguistics. https://doi.org/10.3115/1557690.1557718

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.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932, Prague, Czech Republic. Association for Computational Linguistics.

Tomohiro Ohno, Shigeki Matsubara, Nobuo Kawaguchi, and Yasuyoshi Inagaki. 2005. Robust dependency parsing of spontaneous Japanese spoken language. *IEICE Transactions on Information and Systems*, E88-D(3): 545–552. https://doi.org/10.1093/ietisy/e88-d.3.545

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6015

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.

Mohammad Sadegh Rasooli and Joel Tetreault. 2013. Joint parsing and disfluency detection

in linear time. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 124–129, Seattle, Washington, USA. 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. https://doi.org/10.3115/1629795.1629799

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. https://doi.org/10.1017/S0305000909990407, PubMed: 20334720

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.

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. https://doi.org/10.3115/1219840.1219865

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).

Alessandro Sanchez, Stephan C. Meylan, Mika Braginsky, Kyle E. MacDonald, Daniel Yurovsky, and Michael C. Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, 51(4):1928–1941. https://doi.org/10.3758/s13428-018-1176-7, PubMed: 30623390

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. University of Pennsylvania, School of Engineering and Applied Science.

Hollis S. Scarborough. 1990. Index of productive syntax. *Applied psycholinguistics*, 11(1):1–22. https://doi.org/10.1017/S0142716400008262

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 160–172, Sofia, Bulgaria. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-2011

Ida Szubert, Omri Abend, Nathan Schneider, Samuel Gibbon, Sharon Goldwater, and Mark Steedman. 2021. Cross-linguistically consistent semantic and syntactic annotation of child-directed speech. *arXiv preprint arXiv:2109.10952*.

Daniel Varab and Natalie Schluter. 2019. UniParse: A universal graph-based parsing toolkit. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 406–410, Turku, Finland. Linköping University Electronic Press.

Richard M. Weist and Andrea A. Zevenbergen. 2008. Autobiographical memory and past time reference. *Language Learning and Development*, 4(4):291–308. https://doi.org/10.1080/15475440802293490

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics.

Masashi Yoshikawa, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Joint transition-based dependency parsing and disfluency detection for automatic speech recognition texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1041, Austin, Texas. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1109

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Erik Andersen, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Yifat Ben Moshe, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ishola Ọlájídé, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung

Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín-Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Öztürk Başaran Balkız, Teresa Paccosi, Alessio Palmero Aprosio, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A. Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum

Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2022. Universal dependencies 2.10. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.