Universal Dependencies and Semantics for English and Hebrew Child-directed Speech

Ida Szubert University of Edinburgh k.i.szubert@sms.ed.ac.uk
Omri Abend Hebrew University of Jerusalem omri.abend@mail.huji.ac.il
Nathan Schneider Georgetown University nathan.schneider@georgetown.edu

Samuel Gibbon **Sharon Goldwater** **Mark Steedman**
University of Edinburgh
{samuel.gibbon@, sgwater@inf., steedman@inf.}ed.ac.uk

1 Introduction

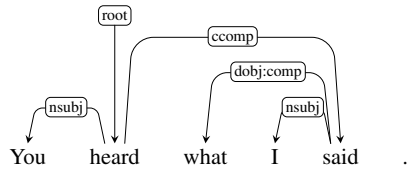
While corpora of child speech and child-directed speech (CDS) have enabled major contributions to the study of child language acquisition, semantic annotation for such corpora is still scarce and lacks a uniform standard. We compile two CDS corpora—in English and Hebrew—with syntactic and semantic annotations. We employ a methodology that enforces a cross-linguistically consistent representation, building on recent advances in dependency representation and semantic parsing. Our semi-automatic syntactic annotation follows the Universal Dependencies standard (UD; de Marneffe et al., 2021), adapted to suit the CDS genre. To induce semantic forms, we develop an automatic method for transducing UD structures into sentential logical forms (LFs), e.g. figure 1. The two representations have complementary strengths: UD structures are language-neutral and support direct annotation, whereas LFs are neutral as to the syntax-semantics interface, and transparently encode semantic distinctions.

What follows is a brief synopsis of the work, which is described in full in (Szubert et al., 2021).

2 Related Work

The CHILDES project (MacWhinney, 2000) has been pivotal in efforts to streamline linguistic data collection of child–caregiver interactions and to standardize linguistic annotation in this domain. CDS resources annotated with *semantic* annotation, however, are scarce, and lack a uniform standard. Indeed, even *syntactic* annotation is only available in CHILDES for a handful of languages, and these are not all annotated according to the same scheme.

Syntax. To the extent that CHILDES data has been syntactically annotated, various syntactic representations have been adopted (Sagae et al., 2010; Pearl and Sprouse, 2013; Odijk et al., 2018). Given the state of the art in multilingual parsing, we argue that UD is the best choice for cross-linguistically comparable syntactic annotation.[1]



LF: $\lambda e_1.\ heard_{e_1}(you,\ \text{WHAT}\ x[\lambda e_2.\ said_{e_2}(I,\ x)])$

Figure 1: Example syntactic and semantic annotation.

Semantics. Sentential logical forms (henceforth, LFs) are an essential building block in a complete linguistic analysis of CDS, and are needed for computational implementations of theories of acquisition that take a "semantic bootstrapping" approach, i.e., construe grammar acquisition as the attachment of language-specific syntax to logical forms related to a universal conceptual structure (e.g., Pinker, 1979; Briscoe, 2000; Abend et al., 2017b). Nevertheless, very few CDS corpora are annotated with sentential meaning representations. Examples include verb and preposition sense annotation, as well as semantic role labeling of data from English CHILDES (Moon et al., 2018), and sentential logical forms (Villavicencio, 2002; Buttery, 2006; Kwiatkowski et al., 2012). See (Alishahi and Stevenson, 2008) for a related line of work. We are not aware of any semantically annotated CDS corpus for languages other than English.

3 Semantic representation

Our goal is to demonstrate an approach to annotating CDS with *cross-linguistically consistent syntax and semantics*. For syntax, we use the Universal Dependencies (UD) standard, motivated by

[1] The English Eve corpus has been annotated with UD structures, using a semi-automatic approach akin to ours, in contemporaneous work (Liu and Prud'hommeaux, 2021).

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 235-240.
Held on-line February 7-9, 2022

its demonstrated applicability to a wide variety of domains and languages, and its relative ease and reliability for manual annotation of corpora. Moreover, as UD is the de facto standard for dependency annotation in NLP, it is supported by a large and expanding body of research work, and by a variety of parsers and other tools. For semantics, we automatically transduce these UD structures into logical forms—thereby obtaining cross-linguistic consistency for those annotations as well, while avoiding the difficult and error-prone procedure of annotating LFs over utterances from scratch. Figure 1 shows an example.

LF generation. The syntactic representation assumed as the input is UD with Universal POS tags. Our system is based on UDepLambda of Reddy et al. (2016), which we modified to accommodate a different target LF.

UDepLambda is a conversion system based on the assumption that Universal Dependencies can serve as a scaffolding for a compositional semantic structure—individual words and dependency relations are assigned their semantic representations, and those are then iteratively combined to yield the representation of the whole sentence. Our modification to UDepLambda consists of providing a new set of rules, which defines a semantics different from the default one used by UDepLambda. Our target is a Davidsonian-style event semantics Davidson (1967), encoded in a typed lambda calculus. An utterance is assumed to describe an event, and the LFs typically contain an event variable with scope over the whole expression. A comprehensive description of the target LF can be found in (Szubert et al., 2021).

Converting a UD parse to an LF is a three-stage process:

- Tree transformation: facilitates LF assignment. The transformations primarily include subcategorizing POS and dependency labels and removing semantically vacuous items. The rules consist of a tree regular expression (Tregex; Levy and Andrew, 2006) and an action to be taken when the pattern is matched. The example in figure 2(b) illustrates subcategorization of the POS tag of a verb whose only core argument is a direct object. Most rules depend only on the syntactic context, with the only exception being the lexicalized rules for recognizing question words. There are 120 rules in total.
- LF assignment: each dependency and each lexical item are assigned a logical form, based on their POS tag / edge label and their syntactic context, as in figure 2(c). The LF assignment rules are not lexicalized. There are 230 assignment rules.
- Tree binarization and LF reduction: The parse tree is binarized to fix the order of composition of word- and dependency-level LFs. Binarization follows a manually created list of dependency priorities. With the order fixed, the sentence-level LF is obtained through beta-reduction, as shown in figure 3.

All rules used in the conversion process are manually created and assigned priorities. UD trees are processed top-to-bottom and the first transformation and LF assignment rule which matches a given node or edge is applied.

Introducing subcategorizations at the tree transformation step is largely a matter of convenience. The same distinctions could in principle be encoded in LF assignment rules. However, introducing more fine-grained labels makes LF assignment rules easier to write and maintain.

4 Corpora

We annotate a large contiguous portion of Brown's **Adam** corpus from CHILDES (the first ≈80% of its child-directed utterances, comprising over 17K English utterances/108K tokens), as well as the entire **Hagar** CHILDES corpus (24K Hebrew utterances/154K tokens) (Berman, 1990).

Adam annotations cover 18,113 child-directed utterances (107,895 tokens) spanning from age 2 years 3 months to 3 years 11 months. Hagar annotations cover 24,172 utterances (154,312 tokens) spanning from age 1 year and 7 months to 3 years and 3 months.

The corpora were selected for their sizes, which are large for CDS corpora, and because they have an initial (non-UD) dependency annotation, part manual and part automatic, which makes our UD annotation process easier (Sagae et al., 2010). We automatically convert these existing parses into approximate UD trees, then hand-correct the converted outputs. Then we apply the UD-to-LF transduction procedure.

4.1 UD annotation

For the most part our UD annotations follow the standard guidelines. However, because of our corpora covering spoken language and specifically

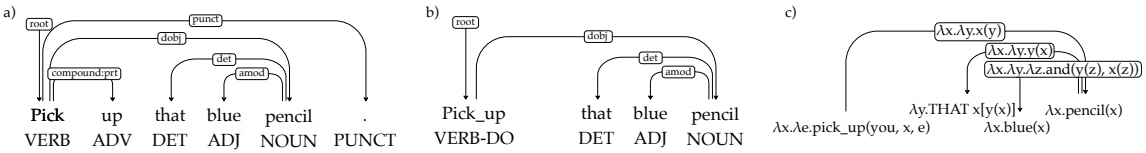


Figure 2: (a) UD parse; (b) tree transformation to subcategorize verb POS, remove punctuation, and combine verb with its particle; (c) LF assignment to nodes and edges.

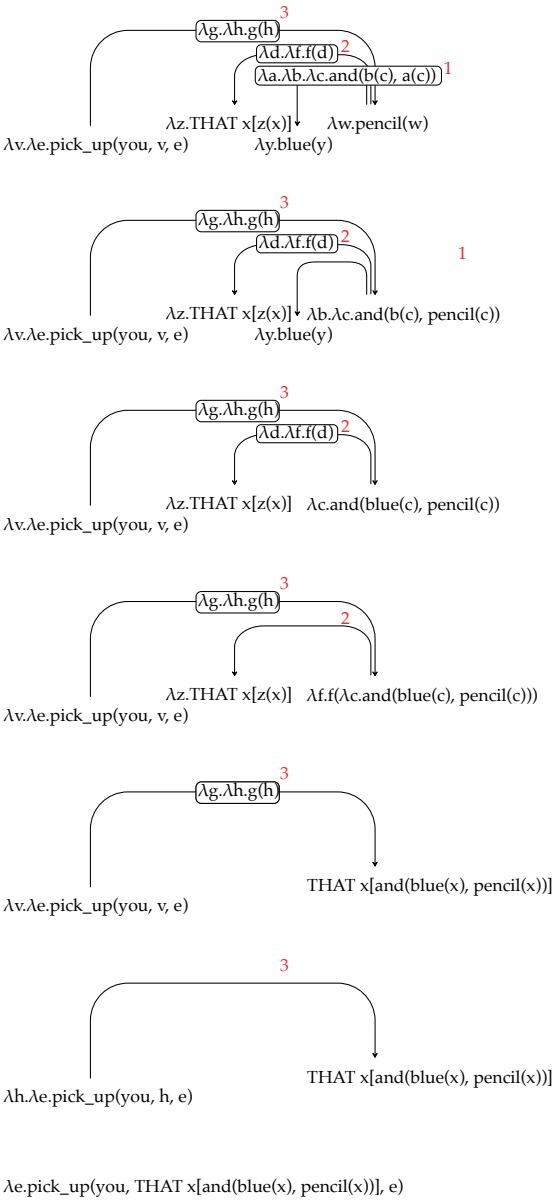


Figure 3: Derivation of the LF for the sentence *Pick up that blue pencil*. Reduction involves applying the LF of the relation to the LF of the head, and applying the output to the LF of the dependent. The red numbers mark the order of composition determined in the tree binarization step.

CDS, we have observed a number of common phenomena that are not often found in other UD corpora for English and Hebrew, which mostly target news and web texts. Indeed, there is little UD-annotated data of spoken English, and none for spoken Hebrew. The unusual constructions we have identified include:

- in-situ WH-pronouns: in English and, with lower frequency, Hebrew
- serial verb contructions: a construction fairly common in our corpora (e.g. *"go get Hans" "bō?i tir?ī"*, lit. *come see*) despite being in general very restricted in English and Hebrew.
- repetitions: both corpora display repetitions, which are a common feature of CDS. They primarily include discursve repetitions (e.g. *"no no don't do that"*) and onomatopeias (e.g. *"oink oink"*).

There is a number of other properties of the CDS genre which make UD annotation challenging and are worth mentioning. Many utterances do not constitute a complete clause and the syntax of such fragments may be underspecified (e.g. *"frighten me for"*). In these cases, we instructed annotators to guess to the best of their ability what the sentence might mean and annotate it accordingly. We have also observed many examples of utterances including quoted fragments, for instance the adult repeating what the child had said, or quoting rhymes, songs, and onomatopoeia. Sentences including quotes are not straightforward to analyze syntactically, and even more difficult to provide semantic representation for. We annotate quotations that do not contain a clause as direct objects, while quotations that do are annotated as complement clauses. Finally, some utterances appear to be playful manipulations of words with the propositional content being unclear or perhaps non-existent (e.g., *"romper bomper stomper boo"*). Where the invented word is embedded within an otherwise intelligible utterance, annotators are instructed to infer its syntactic category from context. Otherwise we use the residual POS tag X and edge type *dep*, in which case the converter produces no LF for the utterance.

4.2 LF annotation

In this synopsis of our work we will not discuss the details of our approach to semantic represenatation, but we will point out some challenges inherent to deriving semantics from a UD annotation. There are cases of underspecification when the information available from the parse tree and POS tags is not sufficient to recover the correct LF. Challenging constructions include examples of scope ambiguity (involving modal verbs or coordination structures), open clausal complements, relative clauses, and clauses without overt subject. We resolve the underspecification problem by heuristically chosing to encode in the UD-to-LF tranduction rules the most common semantic interpretation for a given construction. As an example, all clauses without over subject (excluding cases in which the subject exists, just outside of the clause) are assumed to be imperative and contain an implicit *"you"* subject, even though that is not always true - e.g. in *"See you soon"*.

4.3 Evaluation

For both Adam and Hagar, we find fairly high UD agreement scores comparable with those reported in the literature for English dependency annotation. We obtain a pairwise labeled attachment score (LAS) of 89.9% on Adam and an unlabeled score (UAS) of 95.0%, averaging over the three annotators. Average pairwise agreement on Hebrew is 86.7% LAS and 92.2% UAS. While using them facilitates the annotation process, we find that the converted parser outputs are of fairly low quality: about 40% of the edges are altered relative to the converted parser output in English, and about 30% of the edges in Hebrew.

Next, we evaluate the UD-to-LF conversion procedure. In terms of coverage, it achieves an 80% conversion rate on the English corpus and 72.7% for Hebrew. The converter fails due to ungrammaticality of the utterances, UD and POS annotation errors, and lack of coverage of uncommon syntactic constructions. By manually annotating samples of 100 utterances and comparing the automatically generated LFs, we find that 82% LFs in both English and Hebrew are correct. The transduction errors are primarily caused by the underspecification issues discussed above. More detailed statistics about which constructions appear to be the most challenging to generate LFs for can be found in (Szubert et al., 2021).

4.4 Analysis of corpora

We focus our the analysis on the UD annotation as dependency structures decompose straightforwardly to atomic elements that can be counted and compared. By comparing our Adam and Hagar corpora to the English Web Treebank (Silveira et al., 2014) and Hebrew Dependency Treebank (HDT; Tsarfaty, 2013; McDonald et al., 2013) respectively, we predictably find a higher prevalence of discourse-related dependencies in CDS and a lower prevalence of structures such as adjectival modification, conjunction, compounding, prepositional phrases, clausal modifiers and passive voice. The differences in dependency type frequency between our English and Hebrew corpus are mostly straightforwardly related to typological differences - as a pro-drop language Hebrew has a lower prevalence of *nsubj*; *cop* is more frequent in English because Hebrew lacks an overt copula; *aux* is more frequent in English since tense, which accounts for many examples, is encoded morphologically in Hebrew; prevalence of *case* and *nmod* in Hebrew is higher likely because of indirect objects being expressed using case markers. In (Szubert et al., 2021) we present a longitudinal analysis of the changes in the syntactic composition of the CDS over time.

5 Conclusion

We present a scalable approach to generating meaning representations based on a widly used, cross-linguistically applicable syntactic annotation scheme. While the ability of computational models of acquisition to generalize to different languages is a basic requirement, it has seldom been evaluated empirically, much due to the unavailability of relevant resources. This work immediately enables such comparative investigation in Hebrew and English. Moreover, given the cross-linguistic applicability of UD and the generality of the conversion method, this work is likely to lead to the compilation of similar resources for many languages more, thus supporting broadly cross-linguistic corpus research on child directed speech. Previous work (Abend et al., 2017a) showed that a model of a child's acquisition of grammar can be induced from semantic annotation of the kind discussed here. Future work will apply this model to the compiled corpora, thereby allowing comparative computational research of grammar acquisition in the two languages.

Acknowledgments

The project SEMANTAX has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 742137). The research was funded in part by a James S. McDonnell Foundation Scholar Award (#220020374) to Sharon Goldwater, and a grant by the Israel Science Foundation (grant No. 929/17) to Omri Abend. We would like to acknowledge the contribution of the annotators employed in this project: Zohar Afek, Tamar Johnson, Yelena Lisuk and Adi Shamir.

References

Omri Abend, Tom Kwiatkowski, Nathaniel Smith, Sharon Goldwater, and Mark Steedman. 2017a. Bootstrapping language acquisition. *Cognition*, 164:116–143.

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater, and Mark Steedman. 2017b. Bootstrapping language acquisition. *Cognition*, 164:116–143.

Afra Alishahi and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive Science*, 32(5):789–834.

Ruth A. Berman. 1990. On acquiring an (S)VO language: subjectless sentences in children's Hebrew. *Linguistics*, 28(6):1135–1166.

Ted Briscoe. 2000. Grammatical acquisition: inductive bias and coevolution of language and the language acquisition device. *Language*, 76(2):245–296.

Paula J. Buttery. 2006. Computational models for first language acquisition. Ph.D. dissertation UCAM-CL-TR-675, University of Cambridge, Computer Laboratory, Cambridge, UK.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press, Pittsburgh, PA.

Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proc. of EACL*.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proc. of LREC*, pages 2231–2234, Genoa, Italy.

Zoey Liu and Emily Prud'hommeaux. 2021. Dependency parsing evaluation for low-resource spontaneous speech. In *Proc. of the Second Workshop on Domain Adaptation for NLP*.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Erlbaum, Mahwah, NJ.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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 *Proc. of ACL*, pages 92–97, Sofia, Bulgaria.

Lori Moon, Christos Christodoulopoulos, Fisher Cynthia, Sandra Franco, and Dan Roth. 2018. Gold standard annotations for preposition and verb sense with semantic role labels in adult-child interactions. In *Proc. of COLING*.

Jan Odijk, Alexis Dimitriadis, Martijn van der Klis, Marjo van Koppen, Meie Otten, and Remco van der Veen. 2018. The AnnCor CHILDES Treebank. In *Proc. of LREC*, Miyazaki, Japan.

Lisa Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68.

Steven Pinker. 1979. Formal models of language learning. *Cognition*, 7(3):217–283.

Siva Reddy, Oscar Täckström, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. 2016. Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics*, 4:127–140.

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.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proc. of LREC*, pages 2897–2904, Reykjavík, Iceland.

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:2109.10952 [cs]*.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. In *Proc. of ACL*, pages 578–584, Sofia, Bulgaria.

Aline Villavicencio. 2002. The acquisition of a unification-based generalised categorial grammar. Ph.D. dissertation UCAM-CL-TR-533, University of Cambridge, Computer Laboratory, Cambridge, UK.