Language Resources and Evaluation
https://doi.org/10.1007/s10579-024-09734-y

ORIGINAL PAPER

Cross-linguistically consistent semantic and syntactic annotation of child-directed speech

Ida Szubert[1] · Omri Abend[3] · Nathan Schneider[4] · Samuel Gibbon[2] · Louis Mahon[1] · Sharon Goldwater[1] · Mark Steedman[1]

Accepted: 20 March 2024

Abstract

Corpora of child speech and child-directed speech (CDS) have enabled major contributions to the study of child language acquisition, yet semantic annotation for such corpora is still scarce and lacks a uniform standard. Semantic annotation of CDS is particularly important for understanding the nature of the input children receive and developing computational models of child language acquisition. For example, under the assumption that children are able to infer meaning representations for (at least some of) the utterances they hear, the acquisition task is to learn a grammar that can map novel adult utterances onto their corresponding meaning representations, in the face of noise and distraction by other contextually possible meanings. To study this problem and to develop computational models of it, we need corpora that provide both adult utterances and their meaning representations, ideally using annotation that is consistent across a range of languages in order to facilitate cross-linguistic comparative studies. This paper proposes a methodology for constructing such corpora of CDS paired with sentential logical forms, and uses this method to create two such corpora, in English and Hebrew. The approach enforces a cross-linguistically consistent representation, building on recent advances in dependency representation and semantic parsing. Specifically, the approach involves two steps. First, we annotate the corpora using the Universal Dependencies (UD) scheme for syntactic annotation, which has been developed to apply consistently to a wide variety of domains and typologically diverse languages. Next, we further annotate these data by applying an automatic method for transducing sentential logical forms (LFs) from UD structures. The UD and LF representations have complementary strengths: UD structures are language-neutral and support consistent and reliable annotation by multiple annotators, whereas LFs are neutral as to their syntactic derivation and transparently encode semantic relations. Using this approach, we provide syntactic and semantic annotation for two corpora from CHILDES: Brown's Adam corpus (English; we annotate $\approx 80\%$ of its child-directed utterances), all child-directed utterances from Berman's Hagar corpus (Hebrew). We verify the quality of the UD annotation using an inter-annotator agreement study, and manually evaluate the

Extended author information available on the last page of the article

Published online: 15 May 2024

transduced meaning representations. We then demonstrate the utility of the compiled corpora through (1) a longitudinal corpus study of the prevalence of different syntactic and semantic phenomena in the CDS, and (2) applying an existing computational model of language acquisition to the two corpora and briefly comparing the results across languages.

Keywords Child directed speech · Semantic annotation · Syntactic annotation · Cross-linguistic applicability

1 Introduction

As research in child language acquisition becomes increasingly data-driven, the availability of annotated corpora of child and child-directed speech (CDS) is increasingly important as a basis for understanding the process of child language acquisition from such input. The CHILDES project (MacWhinney, 2000) has been pivotal in the effort to streamline data collection and to standardize linguistic annotation in this domain. However, despite these achievements, CDS resources annotated with semantic annotation 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. For example, Sagae et al. (2010) developed a dependency annotation for CHILDES and applied it to English and Spanish, whereas Gretz et al. (2015) used a different dependency scheme when annotating Hebrew CDS. Neither of these schemes is standardly used in the field of natural language processing (NLP), limiting the application of NLP tools developed elsewhere. Meanwhile, the Dutch AnnCor CHILDES Treebank (Odijk et al., 2018) uses yet another dependency scheme, based on the Alpino parser (Bouma et al., 2001), and a sizeable portion of the English CHILDES Treebank has been annotated with constituency trees following the Penn Treebank annotation scheme (Pearl & Sprouse, 2013). Thus, when it comes to acquisition corpora, syntactic annotations are heterogeneous within and between languages, and do not necessarily reflect prevailing approaches for annotating other genres.

Despite a number of linguistic challenges in analyzing transcribed speech of adult-child interactions, we argue that datasets for studying syntactic acquisition need not be idiosyncratic. This work investigates, first, whether a syntactic framework that is now well-established in NLP—Universal Dependencies—can be applied to child-directed speech transcripts in multiple languages; and second, how language-agnostic rules can map such annotations into sentential logical forms suitable for studying the Semantic Bootstrapping Hypothesis, and the acquisition of grounded sentential semantics (e.g., Mao et al., 2021).

We motivate the components of this goal in turn.

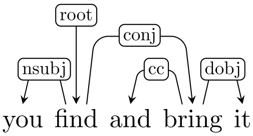
Child-directed speech In approaching syntax and semantics in acquisition data, we are mindful of the fact that empirical studies of language acquisition often focus on child-directed speech, i.e., utterances by adults who are interacting with the child learner. Despite the fact that the child's own utterances are in fact annotated in the original CHILDES corpora, we follow the above research in further annotating only the child-directed side of the data, leaving the child's own utterance unaffected, for two reasons. The first is that it is the child-directed component of the dialog that provides the language-specific input to the child's language-learning process, and the data for any model of how that process works. The second is that almost the only thing that we know about the structures or meaning representations that underlie early child utterances is that they are continuously changing—and thus, in our view, best modeled as latent structure.

Cross-linguistic applicability To the best of our knowledge, the present work is the first to apply a cross-linguistically consistent syntactic annotation scheme to CDS. This consistency is important to enable comparisons across typologically distinct languages: both corpus analyses investigating features of the adult input, and modelling studies testing theories of language acquisition. To illustrate its use, we annotate corpora in two languages: English and Hebrew. We also propose a methodology for producing cross-linguistically consistent *semantic* annotation of CDS.

Syntactic framework As a syntactic representation, from which we will generate the non-aligned logical forms that provide the input to the child or computational learning model, we use the Universal Dependencies (UD) standard (de Marneffe et al., 2021; Nivre et al., 2016), motivated by its demonstrated applicability to a wide variety of domains and languages, and its relative reliability for manual annotation of corpora (Berzak et al., 2016a). 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. The UD standard is briefly presented in Sect. 2. Our annotation reveals various distinctive characteristics of the CDS genre, for which we propose UD conventions (Sect. 2.2).

Logical forms and semantic bootstrapping 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 emphasize the role of "semantic bootstrapping", i.e., theories that construe grammar acquisition as the attachment of language-specific syntax to logical forms related to a universal conceptual structure (e.g., Abend et al., 2017; Bowerman, 1974; Briscoe, 2000; Buttery, 2006; Culicover & Wilkins, 1984; Pinker, 1979). Nevertheless, very few corpora of CDS are annotated with sentential meaning representations. Examples include verb- and preposition-sense annotation, as well as semantic role-labeling of data from English CHILDES by Moon et al. (2018), and sentential logical forms produced by Buttery (2006), Villavicencio (2002) and by Kwiatkowski et al. (2012). A related line of work automatically generated inputs for computational models of acquisition from a semantic lexicon (Alishahi & Stevenson, 2008). We are not aware of any semantically annotated CDS corpora for languages other than English. To address this gap, we further propose a method for automatically transducing LFs from UD structures, 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.

Semantics beyond syntax Although the LF level of representation is deterministically derived from the dependency level, this additional level of annotation is important since it is neutral with respect to surface word order and therefore comparatively language-independent—a key feature for developing and testing models of language acquisition. The transduction process we propose therefore abstracts away from syntactic detail, and transparently encodes information which is implicit in UD—in particular, long-range dependencies. As an example, consider the following, in which the subject "you" and the object "it" are shared between "find" and "bring"[1]



(1) LF: $\lambda e_1.\ and(find_{e_1}(you,\ it),\ bring_{e_1}(you,\ it))$

This information is only implicit in the UD structure, but is made explicit in the LF (though see Sect. 3.3). As is the case with any dependency annotation, some distinctions (such as coordination and scope) are underspecified in UD. We disambiguate some of these cases by refining the set of UD labels (see Sect. 2). Other cases cannot be handled effectively due to their underspecification in the UD formalism (as opposed to other grammar formalisms, such as, e.g., CCG (Steedman, 2000), or semantic schemes such as AMR (Banarescu et al., 2012). We discuss the relationship between our LF formalism and other semantic schemes in Sect. 3 and discuss its limitations in Sect. 3.3.

The conversion method is implemented by recursively building the LFs using unlexicalized rules that condition only on the UD dependency tree and Part of Speech (POS) tags.[2] As such, these rules can be applied to any UD-annotated sentence, regardless of its language. In this we follow the framework of Reddy et al. (2016), but cover a wider range of semantic phenomena, using a different representation language.[3]

Nature of the LFs Our LFs, detailed in Sect. 3, reflect what we take to be a fairly standard model-theoretic semantics. The focus is on compositional, as opposed to lexical, aspects of sentence meaning—i.e., aspects most crucial to modeling

[1] For simplicity, we notate most conceptual content in the LF as words (e.g., *you*: rather than *you′*), to be understood as logical constants.

[2] The only exceptions are the wh-pronouns, which are lexically conditioned.

[3] Reddy's representation is specialized for querying the Google/FreeBase Knowledge Graph.

the acquisition of syntax. Notably, in NLP there is a wider landscape of symbolic meaning representations applied to corpora, such as Universal Conceptual Cognitive Annotation (UCCA; Abend & Rappoport, 2013), Abstract Meaning Representations (AMR; Banarescu et al., 2013), and the Generative Lexicon (GL; Pustejovsky, 1998). Those representations, however, contain additional elements of meaning (like coreference and richer lexical semantics), and are therefore more challenging to annotate or parse.[4] Our LFs could, however, provide a starting point for inducing more elaborate semantic annotations in such frameworks.

New resource Using the proposed protocol of syntactic annotation, 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), as well as over 24K Hebrew utterances, constituting the entire Hagar CHILDES corpus (Berman, 1990). 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) (see below). In addition, the Adam Corpus was chosen because of the availability of the other labeled versions (Moon et al., 2018; Pearl & Sprouse, 2013), and because of the large amount of psycholinguistic study that has been applied to it (Brown, 1973; McNeill, 1966, *passim*).

To obtain gold-standard UD trees, we take advantage of the existing syntactic annotations in these corpora: we automatically convert them into approximate UD trees (Sect. 4.3), then hand-correct the converted outputs. We chose this procedure as we found it to be much faster than annotation from scratch, but note that it is not required: other corpora without preexisting dependency annotation could be annotated with UD parses directly. A schematic overview of the complete syntactic/semantic annotation methodology is given in Fig. 1.

We note that (Liu & Prud'hommeaux, 2021), in contemporaneous work, annotated the English Eve corpus with UD structures, using a semi-automatic approach akin to ours (but did not address other languages or the transduction of logical forms).

Evaluation We evaluate our method by first measuring inter-annotator agreement for UD parses in both corpora, showing that UD can be reliably applied to CDS in both languages (Sect. 5). Of all parsed sentences, our LF conversion tool is able to produce an output for 80.5% (English) and 72.7% (Hebrew). We then manually evaluate a small sample of these LFs and find that 82% of the LFs in both languages are fully correct. Most errors fall into a small number of categories, discussed in Sect. 3.3.

[4] AMR, moreover, was initially designed just for English, and without anchoring of concepts to words in the sentence, which makes it challenging to derive an AMR graph compositionally (attempts to do so include Blodgett & Schneider, 2019, 2021; Groschwitz et al., 2018; Szubert et al., 2018). A new framework, Uniform Meaning Representation (UMR; Van Gysel et al., 2021), aims to address some of these limitations, but is still under development.

Next, we provide some simple proof-of-concept analyses illustrating the benefit of these cross-linguistically consistent annotations (Sect. 6). We compare the usage frequency of different dependency types in our CDS corpora relative to written text corpora in the same languages, and between the English and Hebrew CDS corpora. Overall, we find that the CDS corpora are more similar to each other than to the text corpora in the same language. We also perform a longitudinal analysis, looking for systematic changes in the frequency of use of various syntactic constructions. We find that while in the English corpus only a small number of constructions increase in frequency (adjectival and relative clauses, noun compounding, and noun ellipsis), in the Hebrew one the changes are much more widespread. This can possibly be explained by the different ages of the children at the time of data collection. The finding for English could be relevant to the ongoing discussion as for whether the complexity of CDS changes or not over the longitudinal trajectory. Our findings can be interpreted as echoing the findings of Newport (1977), who also found that syntactic complexity in English CDS does not generally increase with time, except for the number of clauses, which shows a moderate increase.

Finally, as a proof of concept for demonstrating the utility of this work for the modeling of child language acquisition, we adapt the acquisition learning model by Abend et al. (2017) to learn from the transduced LFs (Sect. 7). Experiments are conducted for both English and Hebrew. Results show qualitatively similar trends to the ones reported by Abend et al. (2017).

To recap, we present the following contributions:

1. We show that the UD scheme can be applied to CDS with some additional guidelines, and conduct an inter-annotator agreement study to confirm this finding.
2. We compile two UD-annotated corpora of CDS, one in English and one in Hebrew.
3. We develop an automatic conversion method and codebase for converting UD-annotated CDS to logical forms.
4. We perform a longitudinal corpus study of the prevalence of different syntactic and semantic phenomena in CDS, across the two languages.
5. We show that a baseline grammar for both languages can be induced from the CDU-LF pairs in the corpora by the learner of Abend et al. (2017).

Our annotated data and transduction code are available at https://github.com/ida-szubert/CHILDES_UD2LF. The code for running the simulations is available at https://github.com/ida-szubert/ccg_acquisition_2.[5]

[5] The repositories are still updated from time to time, with small improvements/revisions (and the commit history is of course retained for reproducibility).

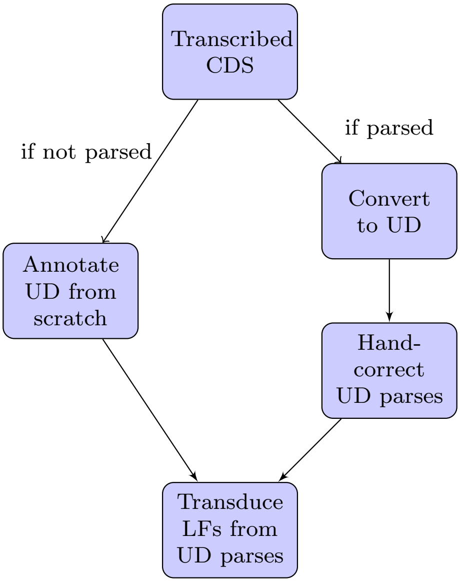


Fig. 1 Main stages of the proposed annotation methodology

2 The Universal Dependencies scheme

Universal Dependencies (de Marneffe et al., 2021; Nivre et al., 2016, UD) is a coarse-grained syntactic dependency scheme which has quickly become the de facto standard for annotating dependencies in many languages. It is designed to establish a unified standard for dependency annotation across languages and domains, to support rapid annotation, and to be suitable for parsing and helpful for downstream language understanding tasks. All these design principles fit naturally with the goals of this paper. Moreover, in order to attain cross-linguistic applicability, UD's design conventions are often similar to those made by semantic schemes (Hershcovich et al., 2019).

Formally, UD uses trees in which nodes are lexical items and directed edges represent dependencies labeled with types such as subject, modifier, etc. UD further includes conventions for annotating morphology, although only POS tags, morphological features and dependency structures are addressed in this work.[6] We use the UD guidelines version 1.0, as reference corpora for version 2.0 were not available at the time of annotation.[7]

We will now turn to UD's treatment of frequent constructions. A glossary of some common UD edge types used in this paper is given in Table 1.

[6] Our rules do not invoke specific morphological features, but we retain morphological annotations from CHILDES: see Sect. 3.1.

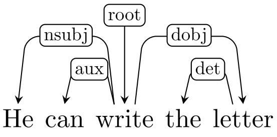
[7] The only exception to this rule is that we attach coordinating conjunctions (*cc*) to their following conjunct, as in version 2.0. Note that it is straightforward to convert this convention to the version 1.0 convention (attach each *cc* edge to the parent of its endpoint), where doing so inversely is non-trivial.

Throughout the rest of the paper we will use the CHILDES transliteration scheme for Hebrew, which directly reflects the writing system of Hebrew.

2.1 Major constructions in UD

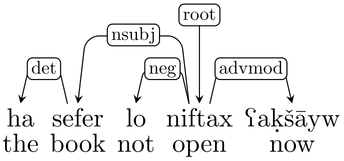
Auxiliaries and modals Auxiliary and modal verbs in UD are dependent on the matrix verb. For example, “can” in this example is dependent on “write”:

(2)



Adverbs and negation Adverbs and negation are treated similarly to auxiliaries and modals, and are also dependents of the matrix predicate.[8]

(3)



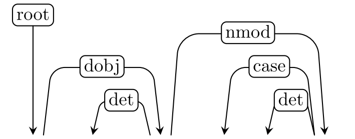
Noun phrases Noun phrases are headed by the lexical head in the case of common NPs, and by the first word in the case of proper nouns.

Adpositional phrases Adpositional phrases are represented as dependents of the head noun when found in a noun phrase. When found in a clause, adpositional phrases are represented as dependents of the matrix verb, and are invariably treated as modifiers so as to avoid drawing a hard distinction between core arguments and adjuncts (a difficult distinction to make in practice; see, e.g., Marcus et al., 1993).

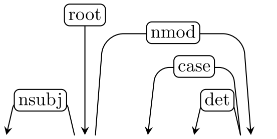
[8] See appendix for the transliteration scheme of the Hebrew letters. We adopt the one used in the Hagar corpus.

Table 1 Some common UD edge types that are used in this paper, and their definitions

Label	Short definition
Clause elements	
nsubj	Nominal subject
dobj	Direct object
ccomp	Clausal complement (finite or infinite), unless its subject is controlled
xcomp	Open clausal complement, i.e., predicative or clausal complement without its own subject
advmod	Modifying adverb
neg	Negation modifier (e.g., “not”, “no”)
aux	Auxiliary of a verbal predicate, including markers of tense, mood, modality, aspect, voice or evidentiality
nmod	Oblique: nominal functioning as an adjunct. (*nmod*s are also used for nominal modifiers in noun phrases, see below)
Inter-clause linkage	
conj	Relation between the conjuncts in a coordination to the first conjunct, which is considered the head
cc	Coordinating conjunction
advcl	Adverbial clause modifier, including temporal clause, consequence, conditional clause, and purpose clause
mark	Marker: the word introducing a clause subordinate to another clause, often a subordinating conjunction
parataxis	Several elements (often clauses or fragments) placed side by side without any explicit coordination, subordination, or argument relation
Nominal elements	
det	Determiner
case	Case marker, including adpositions
nmod	Nominal modifier of a noun or a noun phrase
nummod	Numeric modifier



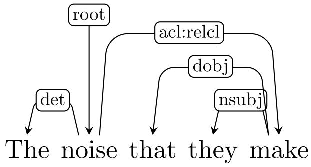
(4) Press the button on the chair



(5) They landed on this spot

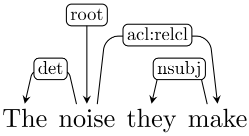
Relative clauses Relative clauses are internally analyzed just like matrix clauses, where the relative clause’s head is considered a dependent of the relativized element. The relative pronoun (where present) is marked with the role of the extracted element. For instance, in the case of object extraction, “that” will have a dependency label *dobj*:

(6)

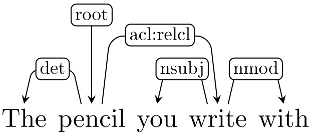


However, where no relative pronoun is present, the extracted slot is underspecified. For instance, “the noise they make” and “the pencil you write with” are analyzed similarly:

(7)



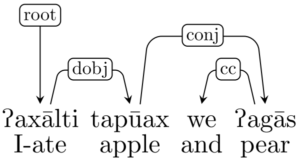
(8)



We therefore introduce two subtypes for the *acl:relcl* dependency label: *acl:relcl_subj* and *acl:relcl_obj* for subject and object relative clauses respectively. Where the extracted element is not the subject or the object, we keep the category *acl:relcl*, for instance in the case of adjuncts (e.g., “the pencil you write with”) or extraction from a complement clause (e.g., “the cat I was taught to like”). The subtyping could be further extended to specify the role of the head noun in those cases, but their frequency in our corpora did not merit further subtyping.

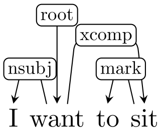
Coordination UD’s convention for coordination designates the headword of the first conjunct as the head (the other conjuncts are dependent on it with a *conj*-labeled edge), while the coordinating conjunctions are dependent on the conjunct following them with a *cc*-labeled edge.

(9)

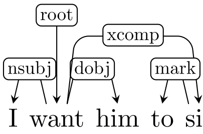


Open clausal complements An open clausal complement of a verb or an adjective (marked as *xcomp*) is defined in UD to be a predicative or clausal complement without its own subject. That is, the subject is inherited from some fixed argument position, often a subject or an object of a higher-level clause. Note that raising and control, which differ in the semantic valency of the matrix verb, are not distinguished in the UD parse.

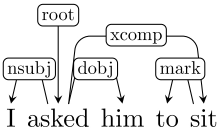
(10)



(11)



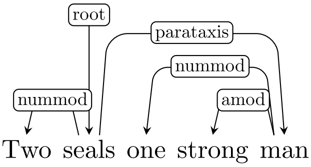
(12)



SUBJECT CONTROL RAISING TO OBJECT OBJECT CONTROL

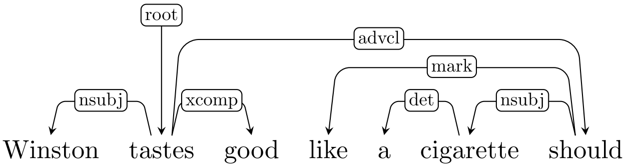
Parataxis Where an utterance consists of several clauses or fragments which are not linked through coordination or subordination, but are somewhat loosely related, UD marks the dependency between them as *parataxis*. For example:

(13)



Ellipsis and promotion Where the head word of a phrase is elided, UD's policy is to "promote" one of its children to be the headword. For example, in Example 14,[9] the auxiliary "should"—which would normally serve as a modifier of the matrix clause—instead serves as the head of the adverbial clause. UD does not distinguish between a promoted head and a regular head.

(14)



In order to make this distinction explicit, we subcategorize the dependency label of the promoted word's incoming edge to indicate that it was promoted to that position (in the above case, *advcl:promoted*). We only target VP ellipsis, due to its

[9] This is an authentic example from the Adam corpus, in which ellipsis is not common.

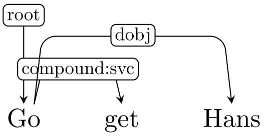
importance in theories of the syntax-semantics interface, but similar subcategorizations are in principle possible for other elliptical constructions.

2.2 Constructions idiosyncratic to CDS

New genres frequently impose new demands on UD annotation guidelines (as can be seen, for example, in the literature on UD for user-generated content; Sanguinetti et al., 2020). We turn to discussing a number of common phenomena from our corpus 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 (mostly parliamentary proceedings), and none for spoken Hebrew. Our corpora are thus different from most existing corpora in targeting spoken language, and in addressing the specific register of CDS.

Serial verb constructions Serial verb constructions (SVCs) are very restricted in English and Hebrew, but are fairly common in CDS. Examples in Adam only include the verb "go" in the first position (e.g. *Go get Hans.*).[10] Examples in Hagar are semantically similar, but include a somewhat broader class of verbs, such as "bōʔi tirʔī" (lit. *come see*). In the absence of clear UD guidelines as to how to treat this construction, we adopt the UDv2 sub-type for SVCs, *compound:svc*, and apply it to this case. For example:

(15)



Ambiguous fragments Many utterances do not constitute a complete clause, but only parts of it. In some cases, the syntax of such fragments may be underspecified. Examples include "frighten me for" from Adam, where it's unclear what the attachment of "for" is, and the following example from Hagar, where the role of "sgulīm" ("purple") is not clear:

Sgulīm, Hagāri, anī loʔ roʔā
Purple$_{pl}$, Hagari, I not see$_{1pl,fem,pres}$

In these cases, we instructed annotators to guess to the best of their ability what the sentence might mean and annotate it accordingly.

[10] This construction is sometimes referred to as quasi-SVC. See Pullum (1990) for discussion.

In-situ WH-pronouns While the grammar of both English and Hebrew requires that wh-pronouns ordinarily be fronted in questions, it is quite common to find in Adam cases where the pronouns stay in place. Examples: “A bird what?”, “Jiminy Cricket who?”, “do not what?”. The phenomenon occurs in Hagar as well, albeit less commonly. We annotate in-situ WH-pronouns the same as we annotate fronted wh-pronouns.

Word plays Some phrases and utterances appear to be playful manipulations of existing words, or belong to some private language between the adult and the child. It is not straightforward to determine what the propositional content of such cases is, if any. Examples include “romper bomper stomper boo” and “sorbalador” from Adam and “baladōn” and “bdibiyabi” from Hagar. Where the invented word is embedded within an otherwise intelligible utterance, annotators are instructed to infer its syntactic category from context. Where the syntax is unclear, we use the residual POS tag X and edge type *dep*. In such cases, our converter produces no LF for the utterance.

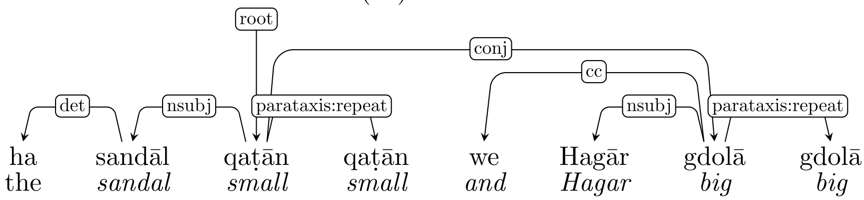
Non-standard vocabulary Other than word plays, examples of non-standard vocabulary include real words or phrases, used in a non-standard way. For example, “nūma nūma” means “sleep sleep” in Hebrew and is part of a nursery rhyme. In Hagar, it appears in “naʕaṣē le ha ʔefrōax nūma nūma”, which translates to “we will do to the chick nūma nūma”, probably meaning they will put it to bed. Other examples may be ungrammatical inflections of real words, e.g., “play games? boat somes”, where “boat somes” probably means “some boats”. We instruct annotators to assign edge labels to words according to their syntactic function, rather than according to their standard function in the target language. For example, “nūma nūma” will be considered a direct object in this case, despite being a verb morphologically.

Quotations We have 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. Examples: “Adam, can you say sits in the chair the boy?”, “It says gobble gobble”, “There’s a dot that says cross your printing set.”, “Did you say fright or did you say fight?”. We annotate quotations that do not contain a clause as direct objects, while quotations that do are annotated as complement clauses.

Repetitions Repetitions of a word or a phrase are common in CDS (Hoff-Ginsberg, 1985; Newport, 1977). The two major sub-classes are discursive repetitions (“no no don’t do that”; “bōʔi bōʔi” lit. “come come”) and onomatopoeias (“oink oink”; “tuk tuk” which is Hebrew for “knock knock”). Some repetitions elaborate on the first occurrence (“Adam’s Adam’s what?”;“ṭipā, ṭipā šel māyim” lit. “drop, drop of water”) or only partially repeat it (“ma ʕoṣīm po ma ʕoṣīm” gloss: “what do$_{pl,masc,pres}$ here what do$_{pl,masc,pres}$”, translation: “what does one do here, what does one do?”). The motivation for some repetitions is obscure, even in context (“guess he means ride buggy buggy”).

We introduce the subtype *parataxis:repeat* to indicate repetitions, except in cases where the repetition is constructional, as in “hold your hand way way up”, where the repetition is interpreted as an intensifier, and so both “way” instances are annotated as *advmod*.

(16)



Note that *parataxis:repeat* is different than the UD subtype *compound:redup*, common in some languages, which denotes the result of the morphosyntactic operation of reduplication.

3 Converting dependency structures to logical forms

The purpose of the system presented in this section is to generate semantic representations on the basis of syntactic ones in a way that is automatic and cross-linguistically applicable. The syntactic representation assumed as the input is in the form of UD, complete with Universal POS tags for each word.

The logical forms we use focus on *compositional sentential semantics*—in particular, argument structure phenomena. An example for the sentence “Do you think the baby whale might want some milk ?” is as follows:

$$\lambda_{e_1}.\ Q(do_{e_1}(think_{e_1}(you, \lambda_{e_2}.\ might_{e_2}(want_{e_2}$$
$$(\textsc{the}\ x[and\_comp(baby(x), whale(x))], \textsc{some}\ y[milk(y)])))))$$

Much of this notation should be familiar as a standard Neo-Davidsonian approach to logical semantics, expressed by lambda forms. Briefly, this LF uses two event variables e_1 and e_2, one for *think* and one for *might want*. These are introduced by λ terms and notated as subscripts on predicates associated with the event. The utterance is a polar question, denoted by Q. Two entity variables, x and y, are respectively introduced by the generalized quantifiers THE and SOME. Most content concepts are represented as semantic predicates with names derived from words in the sentence.

With a focus on predicate-argument structure, the LFs are similar in their core semantic content to other broad-coverage semantic schemes, such as AMR (Banarescu et al., 2013) and UCCA (Abend & Rappoport, 2013). For comparison, Fig. 2 presents the above sentence represented as a UCCA graph and an AMR graph.

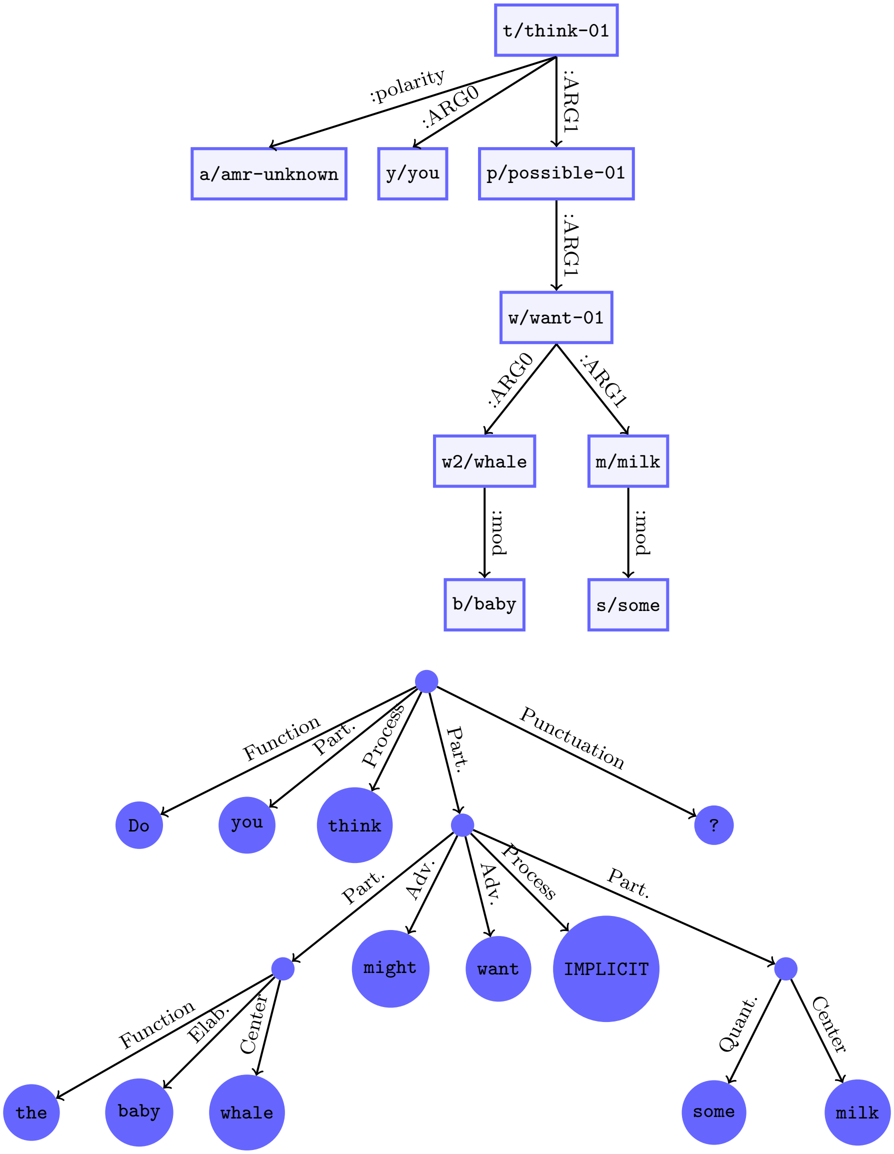


Fig. 2 Example AMR (top) and UCCA (bottom) graphs for the sentence "Do you think the baby whale might want some milk?" Abbreviations: Part. (Participant), Elab. (Elaborator), Quant. (Quantity) and Adv. (Adverbial)

All three representations capture the argument structure of the sentence, (semantic) head-dependent relations, and semantic types of the various constants and variables. AMR and UCCA go beyond the LFs in capturing elements of *lexical semantics* (e.g., word senses, semantic roles), as well as *discourse meaning* (e.g., coreference).

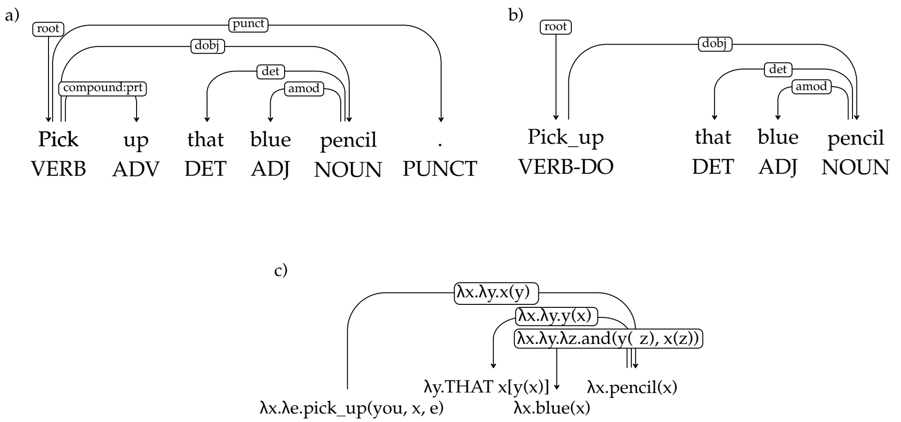


Fig. 3 **a** UD parse, **b** tree transformation to subcategorize verb POS, remove punctuation, and combine verb with its particle, **c** LF assignment to nodes and edges. (Color figure online)

Though it could be valuable to build upon our LFs to incorporate these other aspects of meaning, they are not a part of our investigation here.[11]

The LFs do offer some advantages over some of the aforementioned alternatives. First, they offer a straightforward decomposition into sub-parts that align with individual words. This is in contrast to schemes, like AMR, that do not offer such a decomposition (Szubert et al., 2018). This property of the LFs is useful for modeling or evaluating compositionality in the context of child language learning (see Sect. 7). Second, the LFs can be transduced using a flexible framework (detailed below), that can easily incorporate further (or fewer) distinctions, if provided with the relevant features in the input.

These considerations motivate our choosing LFs over other semantic schemes of semantic representation. We further note that the LFs reflect an underspecified approach to representation that is in line with (i.e., does not make any modeling decisions that contradict) more elaborate semantics that can be applied to these sentences, such as lexical semantics or quantifier semantics. However, we do not see the LFs as superior to other alternatives, and note that a similar resource and analysis could have been produced with other schemes as well.

The output Logical Forms (LF) are typed lambda calculus expressions, and the theoretical approach to semantic representation broadly follows the event semantics of Davidson (1967). Our system is based on UDepLambda of Reddy et al. (2016), which we modified to accommodate a different target LF. We stress that the LFs do not contain lexical semantic information about the words involved, and the

[11] That LFs are focused on compositional aspects of meaning is what allows us to induce them from UD trees without relying on additional resources such as lexicons. The LFs shallowly equate content words with concepts; they do not incorporate word sense disambiguation or semantic roles, but do use POS tags, morphological inflection features and multiword expressions to the extent encoded in UD (Sect. 3.2) for disambiguation purposes.

transcribed words themselves are generally used as their logical constants (e.g., "pencil" and "blue" are used in Fig. 3 to refer to the concept of a pencil and the color blue).

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.

In what follows we present the UD-to-LF conversion process and discuss our choice of LF.

3.1 Conversion process

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

- Tree transformation: as an initial step of conversion we modify the parse trees in order to facilitate the subsequent process of LF assignment. The transformations primarily include subcategorizing POS and dependency labels and removing semantically vacuous items. The rules used in this process (as well as LF assignment rules) consist of a tree regular expression (Tregex; Levy & Andrew, 2006) and an action to be taken when the pattern is matched. The example in Fig. 3b illustrates subcategorization of the POS tag of a verb whose only core argument is a direct object. A tregex is used which matches a verb with an outgoing *dobj, ccomp* or *xcomp* dependency but without a *nsubj* or *iobj*, and not in a subject control context (i.e. with an incoming *xcomp* edge); when a node is a match, we change its POS label to VERB-DO. Most transformation rules depend only on the syntactic context (POS tags and dependency labels), 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 in the sentence are assigned a logical form, based on their POS tag/edge label and their syntactic context, as in Fig. 3c. The LF assignment rules are not lexicalized. There are 230 assignment rules. For simplicity of presentation here, we write the logical constants in the LFs in the same way as their corresponding words. However, in the corpus the logical constants indicate the POS, lemma and inflection given in the CHILDES annotations. For example, the constant corresponding to the base form of the verb *think* would be *verb|think*, while *thinking* would be *participle|think-presentProgressive*. These symbols are treated atomically by the converter, so they serve as a way to minimally disambiguate different inflections with the same surface form, but otherwise the POS and morphology are not used by the converter.
- 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 manu-

ally created list of dependency priorities. With the order fixed, the sentence-level LF is obtained through beta-reduction, as shown in Fig. 4.

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.

3.2 Target logical forms

Our target is a Davidsonian-style event semantics, encoded in a typed lambda calculus.[12] In this section we describe the output we designed for the converter without claiming it to be "target semantics" understood as an ideal meaning representation.

An utterance is assumed to describe an event, and the LFs typically contain an event variable with scope over the whole expression. For example, the LF for the sentence *You found it* is

$$\lambda_{e_1}.\,found_{e_1}(you, it)$$

In the interest of legibility we show the event variable as a subscript of all predicates it has scope over instead of showing it as their argument. In the corpus all variables are typed, the event variable is always the last argument of the predicate.

We turn to discussing the resulting representations for a number of common phrase types and constructions.

3.2.1 Nominals

This category includes common and proper nouns, as well as pronouns.

Pronouns and proper nouns are treated as referring expressions and are represented as atomic terms. Common nouns are treated as non-referring and represented as functions of arity 1, requiring an argument to become referential. When determiners and quantifiers combine with common nouns, they provide such an argument by introducing a variable which they bind. Mass nouns and plural nouns, despite not requiring a determiner, are represented the same way as full determiner phrases, with a placeholder *BARE* determiner,[13]

[12] This paper does not focus on the type system, but roughly speaking, it works as follows. There are three base types in our calculus: t for truth type, v for variable (individual) type, and r for event type. Predicates with all argument slots saturated are functions from events to truth value, hence type $<r,t>$. In the target LFs only variables and constants are typed, but during derivation all expressions are typed.

[13] The actual format in the corpus is of the form *your(x, toy(x)) BARE(x, toy(x))*. In the paper, we adopt a format more familiar from literature on quantifier phrase semantics for the sake of readability.

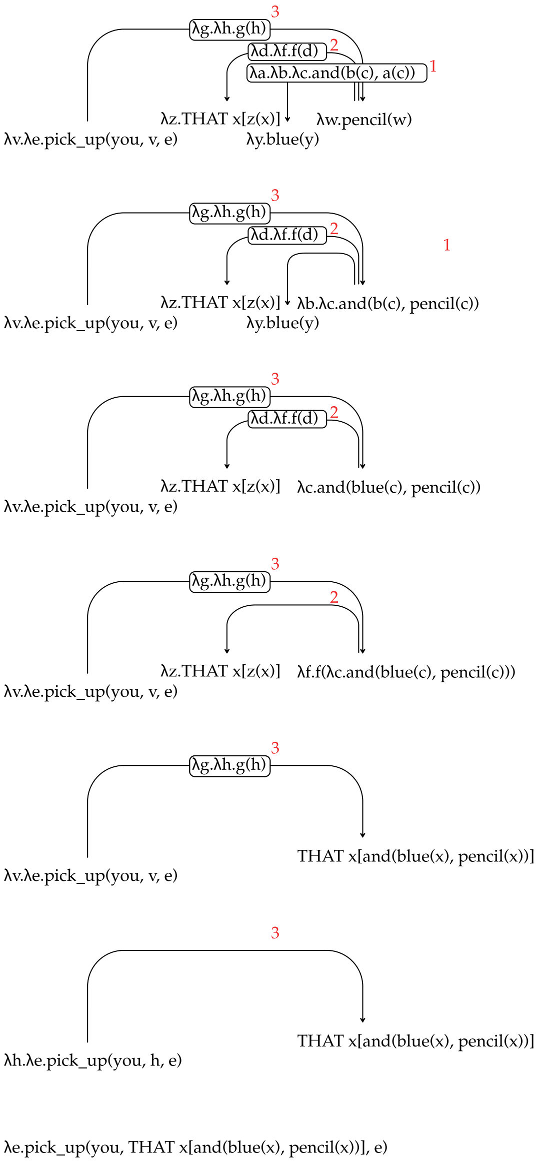


Fig. 4 Derivation of the LF for the sentence *Pick up that blue pencil*, starting after α-conversion of the LF expressions. Reduction proceeds by applying the LF of the dependency relation to the LF of the head, and applying the resulting LF to the LF of the dependent. The red numbers mark the order of composition determined in the tree binarization step. (Color figure online)

it: *it*
Adam: *Adam*
toy: *λx. toy(x)*
a toy: A *x*[*toy(x)*]
toys: BARE *x*[*toy(x)*]

Where proper nouns appear with a determiner, they are treated similarly to common nouns:

the Daddy: THE *x*[*Daddy(x)*]

Possessives are treated in the same way as determiners:

your toy: YOUR *x*[*toy(x)*]
Diandro's bottle: DIANDRO'S *x*[*bottle(x)*]

When used predicatively (notably, in copular constructions), nominals are treated as predicates with an arity of 2, taking as arguments a subject and an event variable. All nominals therefore have two possible types of LFs, non-predicative and predicative.

Where nominals are used predicatively, we do not interpret their determiner as having a determiner semantics in the LF, and instead simply interpret it as an application of the predicate defined by the nominal to the subject:

It is a raccoon: λ_{e_1}. *raccoon*$_{e_1}$ (*it*)
My pet is a raccoon: λ_{e_1}. *raccoon*$_{e_1}$(MY *x*[*pet*(*x*)])
This is the car: λ_{e_1}. *the_car*$_{e_1}$ *(this)*

Noun–noun compounds are represented by treating both nouns as arguments to a special *and_comp* predicate.

Show me a space boat: λ_{e_1}. *show*$_{e_1}$ *(you*, A *x*[*and_comp(space(x), boat(x))*], *me)*

3.2.2 Adjectives

Like common nouns, adjectives are represented as arity 1 predicates. We assume intersective semantics for adjectives, i.e., a nice carpenter is a thing which is nice and which is a carpenter.

nice: λx. *nice(x)*
nice carpenter: λx. *and(nice(x), carpenter(x))*

This is decidedly a simplification of the actual nuanced adjective semantics, e.g., a fake bear is not really a bear, and a good liar is not a person who is good and who is a liar.

Adjectives can head copular constructions, in which case they behave like arity 2 predicates, analogously to nominals in the same situation. It is possible for an adjective in those constructions to also have a clausal object, increasing arity to 3.

This carpenter was nice: λ_{e_1}. *nice*$_{e_1}$ (THIS *x*[*carpenter(x)*])
I am sorry to go: λ_{e_1}. *sorry*$_{e_1}$ *(I*, ${}_{e_2}$*go*$_{e_2}$ *(I))*

3.2.3 Verbs

Verbs are represented by predicates whose arity varies from 1 to 4, with possible arguments being subject, direct object, indirect object, clausal arguments (see below) and the Davidsonian event variable (represented in that order in the LF). The argument type is defined by its syntactic relation to the verb in the unmarked form. If a verb takes less than 4 arguments, we leave the other positions unfilled.

You gave Ursula the box: $\lambda_{e_1}.\ gave_{e_1}(you, \textsc{the}\ x[box(x)], Ursula)$
Mommy heard it: $\lambda_{e_1}.\ heard_{e_1}(Mommy, it)$

When a verb lacks an argument whose position precedes the positions of present arguments (with the exception of the event argument), we fill the slot of the missing argument with a "blank" symbol (_). Constructions necessitating this solution include passive voice[14] and some infinitival clausal arguments.

Daddy said to return the pen: $\lambda_{e_1}.\ saide_1$ (Daddy, $_{e_2}.\ return_{e_2}(\_, \textsc{the}\ x[pen(x)]))$
The tree is shaped (like that): $\lambda_{e_1}.\ shaped_{e_1}$ (_, $\textsc{the}\ x[tree(x)]$)

Subject-less clauses For every verb without a subject[15] we assume the clause is in imperative mood and supply *you* in the subject position in the LF.

Drink the juice: $\lambda_{e_1}.\ drink_{e_1}(you, \textsc{the}\ x[juice(x)])$

Auxiliary and modal verbs are predicates with an arity of 1, taking as their argument a proposition.

- He can write: $\lambda_{e_1}.\ can_{e_1}(write_{e_1}(he))$
- He could be writing: $\lambda_{e_1}.\ could_{e_1}(be_{e_1}(writing_{e_1}(he)))$

Particle verbs, including phrasal verbs, are merged into one lexical item of the form *verb_particle* whenever there are no other words intervening between the verb and its particle. Otherwise the particle is treated as a sentential modifier. The difference in treatment is motivated purely by the technical limitations of the converter, not theoretical considerations.

The paint came off: $\lambda_{e_1}.\ came\_off_{e_1}(\textsc{the}\ x[paint(x)])$
It picks the dirt up: $\lambda_{e_1}.\ and(picks_{e_1}(it, \textsc{the}\ x[dirt(x)]), up_{e_1})$

[14] As argument roles are defined semantically, in a passive clause the syntactic subject becomes one of the objects in the LF, and the subject spot is left empty.

[15] This does not include verbs which have a subject that is not directly connected to the verb in the UD parse—verbs controlled by a higher verb, verbs in infinitival complement clauses, verbs in relative clauses, verbs sharing a subject with a conjoined verb.

Serial verb constructions of the form “come get” or “go ask” and their Hebrew counterparts (e.g., “bōʔi tešvī”, lit. “come sit”) are treated in a special way, because semantically the first verb carries little propositional meaning and is purely discoursive in nature. Our converter reduces these expressions to the second verb only.

Go get two pennies: $\lambda_{e_1}.\ \mathit{get}_{e_1}(\mathit{you}, \textsc{two}\ x[\mathit{pennies}(x)])$

3.2.4 Adverbs

Verb-modifying adverbs are represented as predicates which take the event variable as their argument, and are conjoined with the matrix predicate using a general purpose *and*. We do not distinguish between VP-scoped and sentential adverbs because this distinction is not supported by the UD annotation.

She tried again: $\lambda_{e_1}.\ \mathit{and}(\mathit{tried}_{e_1}(\mathit{she}), \mathit{again}_{e_1})$
She certainly tried: $\lambda_{e_1}.\ \mathit{and}(\mathit{tried}_{e_1}(\mathit{she}), \mathit{certainly}_{e_1})$

Adjuncts (which are annotated as adverbs) modifying adjectives are arity 1 predicates whose argument is the LF representation of the modified adjective phrase.

a very kind boy: $\textsc{a}\ x[\mathit{and}(\mathit{very}(\mathit{kind}(x)), \mathit{boy}(x))]$
You are a very kind boy: $\lambda_{e_1}\ A\ \mathit{you}[\mathit{and}(\mathit{very}(\mathit{kind}_{e_1}(\mathit{you})), \mathit{boy}_{e_1}(\mathit{you}))]$

3.2.5 Prepositional phrases

Due to the difficulty in making the complement-adjunct distinction in UD, prepositional phrases (PP) within clauses are invariably considered as sentential modifiers (rather than arguments). A preposition is an arity 2 predicate, whose first argument is the prepositional object, and the second is the event variable, and the LF of the prepositional phrase is conjoined with the LF of the matrix predicate.

He played with Paul: $\lambda_{e_1}.\ \mathit{and}(\mathit{played}_{e_1}(\mathit{he}), \mathit{with}_{e_1}(\mathit{Paul}))$

A PP modification of a nominal is represented using the *att* relation, expressing the fact that the PP is in some sense an attribute of the nominal.

the juice on your shirt: $\textsc{the}\ x\ [\mathit{att}(\mathit{juice}(x), \mathit{on}(\textsc{your}\ y[\mathit{shirt}(y)]))]$

When a PP is used in a copular construction, the preposition is represented by an arity 3 predicate, taking as arguments the nominal inside the PP, the subject, and the event variable.

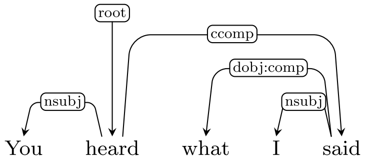
That is from Pinocchio: $\lambda_{e_1}.\mathit{from}_{e_1}(\mathit{Pinocchio}, \mathit{that})$

3.2.6 Relative clauses

Relative clauses provide additional information about a nominal which they modify. We represent the relation between a nominal and a relative clause as conjunction. There is no difference between LFs of normal and reduced relative clauses, or between restrictive and non-restrictive ones.

> We saw those mirrors that you liked: $\lambda_{e_1}.\ saw_{e_1}\ [we,\ \textsc{those}\ x\ (and(mirrors(x),\ \lambda_{e_2}\ liked_{e_2}.\ (you,\ x))])$
> The drum you were playing: $\textsc{the}\ x\ [and(drum(x),\ \lambda_{e_1}.\ were_{e_1}\ (playing_{e_1}\ (you,\ x)))]$

Free relative clauses, as in the example below, pose problems to the UD scheme. In absence of clear annotation guidelines, we decided to attach the relative clause to the matrix clause with the *ccomp* or *csubj* relation and annotate the wh-word in a way that reflects its role within the relative clause. We use whatever relation is appropriate, and subcategorize it with a complementizer subtag, *:comp*.



Using this annotation convention we can produce correct LFs for fused relative clauses:

> You heard what I said: $\lambda_{e_1}.\ heard_{e_1}\ (you,\ \textsc{what}\ x[\lambda_{e_2}.\ said_{e_2}\ (I,\ x)])$

3.2.7 Clausal arguments and modifiers

Clauses can function as arguments of verbs and, less often, other predicates. In LF clausal arguments are treated no different from nominal ones.

> I think that he can talk: $\lambda_{e_1}.\ think_{e_1}\ (I,\ \lambda_{e_2}.\ can_{e_2}\ (talk_{e_2}(he)))$
> He wants you to take a nap: $\lambda_{e_1}.\ wants_{e_1}\ (he,\ \lambda_{e_2}.\ take_{e_2}\ (you,\ \textsc{a}\ x[nap(x)]))$

Generating LFs for clausal complements is complicated by the ambiguity of the UD scheme which does not distinguish between raising to object and object control constructions. The actual semantics differs, but our converter heuristically treats all open clausal complements as if they were cases of object control and produces the LF accordingly. See discussion in Sect. 3.3.

Clausal modification of verbs is represented by treating the matrix clause and the subordinate clause as two arguments of the subordinating conjunction predicate. The predicates representing both clauses share the event variable.[16]

She sings when she is happy: $\lambda_{e_1}.\ \textit{when}(\textit{happy}_{e_1}(\textit{she}), \textit{sings}_{e_1}(\textit{she}))$

Clausal modifiers of nominals (other than relative clauses) come in two types. The first have relative clause semantics:

You saw a tree dancing: $\lambda_{e_1}.\ \textit{saw}_{e_1}(\textit{you}, \textsc{a}\ x[\textit{and}(\textit{tree}(x), \lambda_{e_2}.\ \textit{dancing}_{e_2}(x))])$

The second type of modification is more difficult to encode, as the specific semantic relation between the noun and the modifier is largely implicit. Examples include noun phrases such as *a battle to keep him out*, *one place to put things*, or *the way to play*. We resort to representing the relation with a generic *rel* predicate.

You showed me the way to play the game:
$\lambda_{e_1}.\ \textit{showed}_{e_1}(\textit{you}, \textit{me}, \textsc{the}\ x[\textit{rel}(\textit{way}(x), \lambda_{e_2}.\ \textit{play}_{e_2}(\_, \textsc{the}\ y[\textit{game}(y)]))])$

3.2.8 Negation

Negation of the main predicate of a clause is represented by arity 2 *not* predicate, taking as arguments the negated predicate applied to its arguments and the event variable. The LF of negated nominals follows UD in treating the negation as a determiner. We note that the auxiliary verb "do" is introduced into the logical form whenever it appears as a word in the sentence. Its role in the LF is to serve as a placeholder for tense.

I don't have any sugar: $\lambda_{e_1}.\ \textit{not}_{e_1}(\textit{do}_{e_1}(\textit{have}_{e_1}(I, \textsc{any}\ x[\textit{sugar}(x)])))$
I'm no clown: $\lambda_{e_1}\ \textsc{no}\ I[\textit{clown}_{e_1}(I)]$

3.2.9 Questions

Polar questions are represented by wrapping the LF of the corresponding indicative sentences in a *Q* predicate of arity 1.

Do you have a doll?: $\lambda_{e_1}.\ Q(\textit{do}_{e_1}(\textit{have}_{e_1}(\textit{you}, \textsc{a}\ x[\textit{doll}(x)])))$

Wh-questions are represented by binding in the outer scope the variable which stands for the thing being asked about. This variable is used in the LF in place of the wh-word.

[16] It may be argued that in some cases, the subordinate clause does not in fact correspond to the same event as the main clause. This is a fine distinction not made by UD, and consequently not made in the produced LFs either.

What did you take?: $\lambda x.\ \lambda_{e_1}.\ did_{e_1}\ (take_{e_1}\ (you,\ x))$

Since possessive modifiers are treated the same as quantifiers, we interpret *whose* questions as abstracting over a generalized quantifier, as in the following example (where we have replaced the variable *x* with WHOSE, in the interest of clarity):

Whose name are you writing?: λ WHOSE. $\lambda_{e_1}.\ are_{e_1}$ (writing$_{e_1}$ *(you*, WHOSE *y*[*name(y)*]*))*

3.2.10 Conjunctions

Conjunctions are represented by treating the conjuncts as arguments of the conjunction predicate.[17] In cases of clause conjunction there is only one event variable with scope over both clauses.

He had a fever or a cold:
$\lambda_{e_1}.\ had_{e_1}\ (he,\ or(\textsc{a}\ x[fever(x)],\ \textsc{a}\ y[cold(y)]))$
ʔaxālti tapūax(x) we ʔagās (lit. I-ate apple and pear):
λ_{e_1}. *ʔax ālti*$_{e_1}$ *(1sg, and(*λ_x. *tapūax(x)*, λ_y. *ʔagās(y)))*
Get a kleenex and wipe your mouth:
$\lambda_{e_1}.\ and(get_{e_1}\ (you,\ \textsc{a}\ x[kleenex(x)]\ )\ wipe_{e_1}\ (you,\ \textsc{your}\ y[mouth(y)]))$

Shared arguments of conjoined verbs[18] are explicitly repeated in the LF, as if they were overtly repeated in the sentence. We use a heuristic rule to decide whether an argument is shared or not, which we further discuss in Sect. 3.3.

You find and bring it: $\lambda_{e_1}.\ and(find_{e_1}\ (you,\ it),\ bring_{e_1}\ (you,\ it))$

3.2.11 Names and multiword expressions

We combine words annotated with the *mwe*, *name* and *goeswith* relations into a single lexical item and treat them as such in the LF. These three categories are used in UD to classify a restricted subset of multiword expressions: *name* connects the words of headless names; *mwe* connects fixed grammaticized expressions; *goeswith* connects two parts of the same word that are incorrectly rendered as separate tokens.[19] Many other multiword expressions are syntactically (semi-)regular but semantically idiomatic; our LFs do not capture these as single concepts, since doing so would, at present, require additional layers of annotation (or language-specific

[17] Two expressions can be conjoined only if they are of the same semantic type.

[18] There are only a few examples of conjoined predicates with shared arguments in the Adam corpus, but the converter can deal with cases more complicated than the one shown, such as conjoined verbs in relative clauses with the head noun being an argument of both (e.g. *I'll show you the book I wrote and he edited*).

[19] https://universaldependencies.org/docsv1/u/dep/index.html.

lexical resources and disambiguation). However, other efforts are underway to accommodate a broader range of multiword expressions within the UD framework (Savary et al., 2023), which will in turn enable their treatment by the converter.

3.2.12 Parataxis and discourse markers

The loose semantic connection associated with the UD relations of *discourse* and *parataxis* is represented by conjoining the LFs of both parts with a general purpose *and* predicate.

Wait, we forgot your snack: $\lambda_{e_1}.\ \mathit{and}(\mathit{wait}_{e_1}(\mathit{you}), \mathit{forgot}_{e_1}(\mathit{we}, \textsc{your}\ x[\mathit{snack}(x)]))$
I like it, thank you: $\lambda_{e_1}\ \mathit{and}(\mathit{like}_{e_1}(I, it), \mathit{thank\_you}_{e_1})$

3.2.13 Repetitions

When repetition annotated with *parataxis:repeat* occurs, the LF ignores the repeated element and represents it only once.

ṭipā, ṭipā šel māyim (lit. drop, drop of water): BARE *x*[*att(ṭipā(x), šel (*BARE *y*[*māyim(y)*]*))*]

3.3 Limitations

As observed throughout this section, our LFs encode compositional sentential semantics. The representation does not aim to capture aspects of meaning in the realm of lexical semantics or discourse.

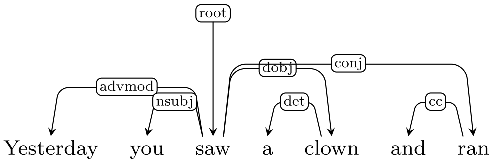
Even in the realm of compositional semantics, there are cases when the information available from the parse tree and POS tags is not sufficient to recover the correct LF. Most of the limitations of the converter have to do with shortcomings of UD as a syntactic annotation schema, discussed in Sect. 2. In cases of unresolvable structural ambiguity we generally choose to use the LF that represents the meaning which is more common in our English corpus. In a number of cases (listed below), where there is no such obvious more common option, we design the converter to simply fail.

Universal Dependencies builds its structures directly on the words of the sentences, and generally does not encode implicit elements or long-range dependencies. This often entails difficulties for our conversion method, in cases such as imperatives (e.g., "come over!"), where the person addressed is not explicit. Enhanced Universal Dependencies (Schuster & Manning, 2016) extend UD and construct graphs over the input tokens (rather than trees), that cover phenomena such as predicate ellipsis (e.g., gapping), and shared arguments due to coordination, control, raising and relative clauses. However, they do not address a variety of implicit arguments, even constructional ones, such as the person addressed in imperative forms. We do not use Enhanced UD in this work due to its language-specificity and use of non-tree structures, which considerably complicate the conversion method.

3.3.1 Scope ambiguity

The major source of ambiguity when deriving an LF from a UD parse is UD's inability to represent scope phenomena. UD trees are not binary and contain no indication about order of composition of the children with the parent, which gives rise to various cases of unclear scope. This is an inevitable consequence of using dependency grammar as annotation rather than (for example) CCG (Steedman, 2000).

- **Argument sharing and modifier scope in verb coordination** Coordination structures are inherently ambiguous in UD, as the headword of the first conjunct serves also as the head of the entire coordination structure (for attempts to enhance UD with more informative annotation of coordination structures, see, e.g., Grünewald et al., 2021; Przepiórkowski & Patejuk, 2019). Arguments of the first conjunct and of the whole coordination structure are rendered indistinguishable. The same holds for modifiers of the first conjunct. For example, in



  it is unclear whether "clown" is an object of "saw" and "ran", or just of "saw", and whether both actions or just one happened yesterday.

  The heuristic we select is: if the head verb has an argument which the other verb lacks, assume that the argument is shared. This leads us to correctly represent sentences such as:

  You find and bring it: $\lambda_{e_1}.\ \mathit{and(find_{e_1}\ (you,\ it)\ bring_{e_1}(you,\ it))}$

  but also to produce some erroneous LFs:

  You saw a clown and ran:
  is $\lambda_{e_1}.\ \mathit{and(saw_{e_1}\ (you,}\ \textsc{a}\ \mathit{x[clown(x)]),\ ran_{e_1}\ (you,}\ \textsc{a}\ \mathit{x[clown(x)]))}$
  should be $\lambda_{e_1}.\ \mathit{and(saw_{e_1}\ (you,}\ \textsc{a}\ \mathit{x[clown(x)]),\ ran_{e_1}\ (you))}$

  With respect to modification, we assume that all modifiers attached to the first verb modify the whole conjunction:

  She ate and drank again: $\lambda_{e_1}.\ \mathit{and(and(ate_{e_1}(she),\ drank_{e_1}\ (she)),\ again_{e_1})}$

  In principle the order of words in the sentence could be used for disambiguation: in English shared objects would occur after the second conjuncts, while objects belonging only to the first conjunct would follow it directly. This, however, would require us to provide the converter with linear order information in addition to the UD parse, and make the converter language-specific.

- **Modal verb scope** UD treats auxiliaries and modals as modifiers of the matrix verb, giving rise to ambiguity in coordinate structures,[20] and ambiguity over the order of combination of modals and adverbs. The source of the difficulty is the lack of distinction between sentence and VP adverbials in UD. We heuristically decide to represent modals as always outscoping adverbs, correctly representing examples such as this:

  Somebody will stop suddenly: $\lambda_{e_1}.\ \mathit{will}_{e_1}\,(\mathit{and}(\mathit{stop}_{e_1}\,(\mathit{somebody}),\ \mathit{suddenly}_{e_1}))$

  but not others:

  Maybe somebody will stop:
  is $\lambda_{e_1}.$ will $_{e_1}\,(\mathit{and}(\mathit{stop}_{e_1}\,(\mathit{somebody}),\ \mathit{maybe}_{e_1}))$
  should be $\lambda_{e_1}.\ \mathit{and}(\mathit{will}_{e_1}\,(\mathit{stop}_{e_1}\,(\mathit{somebody})),\ \mathit{maybe}_{e_1})$
  Similarly as discussed above, a solution could be proposed which would rely on distinguishing between sentential and VP adverbials on the basis of word order, but this information is not available to the converter.

- **Modifier scope in NP coordination** Analogously to the ambiguity arising in verb coordination structures, any modifiers attached to the head noun of a noun coordination structure cause ambiguity. We choose to treat all modifiers as applying to the head of the conjunction only. This results in correct LFs for sentences such as:

  You got sweet pears and lemons:
  $\lambda_{e_1}.\ \mathit{got}_{e_1}\,(\mathit{you},\ \mathit{and}(\textsc{bare}\ x[\mathit{and}(\mathit{sweet}(x),\ \mathit{pears}(x))],\ \textsc{bare}\ y[\mathit{lemon}(y)]))$

  but not for sentences in which the modifier has scope over the conjoined structure:

  You got chocolate eggs and bunnies:
  is $\lambda_{e_1}.\ \mathit{got}_{e_1}\,(\mathit{you},\ \mathit{and}(\textsc{bare}\ x[\mathit{and}(\mathit{chocolate}(x),\ \mathit{eggs}(x))],\ \textsc{bare}\ y[\mathit{bunnies}(y)]))$
  should be $\lambda_{e_1}.\ \mathit{got}_{e_1}\,(\mathit{you},\ \textsc{bare}\ x[\mathit{and}(\mathit{chocolate}(x),\ \mathit{and}(\mathit{eggs}(x),\ \mathit{bunnies}(x))])$

3.3.2 Open clausal complements

UD does not distinguish between object control and raising-to-object structures, and so “I asked you to sit” and “I want you to sit” receive the same UD annotation, despite the fact that “asked” semantically takes “you” as an argument and “want” does not (see Sect. 2).

The converter interprets all open clausal complements as raising-to-object.

He wants you to take a nap (RAISING-TO-OBJECT): $\lambda_{e_1}.\ \mathit{wants}_{e_1}\,(\mathit{he},\ \lambda_{e_2}.\ \mathit{take}_{e_2}\,(\mathit{you},\ \textsc{a}\ x[\mathit{nap}(x)]))$

[20] For example, the modal verb applies to the first conjunct in *I will eat the banana but I prefer apples* but to both conjuncts in *You will sing and play*. Nevertheless, UD attaches the modal verb in both cases to the first conjunct.

Mommy asked you to come (OBJECT CONTROL):
is λ_{e_1}. $asked_{e_1}$ *(Mommy,* λ_{e_2}. $come_{e_2}$ *(you))*
should be λ_{e_1}. $asked_{e_1}$ *(Mommy, you,* λ_{e_2}. $come_{e_2}$ *(you))*

3.3.3 Relative clauses

As discussed in Sect. 2, UD annotation does not specify the role which the relativized noun takes on in the relative clause. In our UD annotation we subcategorize for subject and object relative clauses, but we do not mark the role of the noun if it is not a core argument. The converter fails on those non-subcategorized relative clauses.

For example, in this case the role that the relativized noun takes in the relative clause is that of an object. The converter therefore produces the correct LF as in here:

all things that you find: ALL x[*and(things(x),* λ_{e_1}. $find_{e_1}$ *(you, x))*]

However, in this example, the relativized noun takes the role of a prepositional object in an adjunct landed. Since this is not specified in the UD, the converter will fail on this example, rather produce the correct LF.

the spot they landed on: **should be** THE x[*and(spot(x),* λ_{e_1}. *and(*$landed_{e_1}$ *(they),* on_{e_1}*(x)))*]

Another difficulty is connected with free relative clauses, in which the head nominal is missing and a relativizer pronoun takes its place, e.g. “You heard what I said” in the figure in Sect. 3.2. In the LF we treat the wh-word as a determiner, which introduces a variable standing in for the missing nominal.

3.3.4 Clauses without overt subject

All clauses without a subject are assumed to be imperative (this does not include cases of external clausal subject, as in relative clauses, clausal modifiers of nominals, or in raising and control). We are thus assuming an implicit *you* subject and make it explicit in the LF, which can sometimes lead to mistakes.

See you later:
is λ_{e_1}. *and(*see_{e_1} *(you, you),* $later_{e_1}$*)*
should be λ_{e_1}. *and(*see_{e_1} *(I, you),* $later_{e_1}$*)*

While it is difficult to precisely quantify the frequency of the constructs described above, we do report statistics on the ratio of the utterances for which the converter fails, as well as conduct manual analysis of a sample of produced LFs in order to assess their quality. See Sect. 5.

4 Annotating the CHILDES Adam and Hagar Corpora

4.1 The corpora

Adam We annotate a total of 17,233 child-directed utterances from Brown's Adam corpus, covering sessions 1 to 41 and spanning from age 2 years 3 months to 3 years 11 months. 115 utterances which were incomplete (marked by the final token +...) were discarded. The corpus contains 107,895 tokens.

Hagar We annotate all child-directed utterances in Berman's Hagar corpus, comprising 24,172 utterances in total. The annotated corpus covers 134 sessions (recorded on 115 days, with multiple sessions on some days) from the child's ages of 1 year and 7 months to 3 years and 3 months. 192 incomplete utterances were discarded. The corpus contains 154,312 tokens.

We remained faithful to the existing tokenization of the CHILDES corpus, and so any annotation incorporated to Adam or Hagar, was incorporated into the new scheme. There have been a number of exceptions to this rule:

1. Compounds (tokens that included an underscore _ in them) in the original CHILDES corpus, were split to two tokens, in accordance with the UD guidelines.
2. Correction of words that had errors: some words (around 100 unique words) had errors, such as ya#higīd instead of yagīd. These were corrected by replacing the problematic words with the correct ones.
3. Splitting possessive pronouns in Hebrew from the main stem: possessive pronouns in Hebrew are generally clitics. In accordance with UD guidelines and the existing annotation for Hebrew, we split the clitics from the main stem.

In addition, incomplete sentences were discarded as well. We have also discarded sentences that contain the tokens 'xxx', 'yyy' or 'www' (indicating unidentifiable material, such as unintelligible words).

4.2 Annotator training

The Hagar treebank was annotated by three native speakers of Hebrew with a BA in linguistics. The majority of the Adam treebank (13,709 utterances) was annotated by a single annotator—a native English speaker with a BA in linguistics. The rest of the corpus (4404 utterances) was annotated by two of the Hebrew annotators, both highly proficient in English. Before annotating the treebank, our annotators received extensive training, which consisted of (1) a tutorial from a senior member of the team, (2) reading through the Universal POS tags and English UD guidelines,[21] and (3) annotating a subset of about 100 sentences from the CHILDES corpus, and discussing issues that came up in the annotation. The Hebrew annotators annotated a training batch of sentences in both languages. While working through the training

[21] http://universaldependencies.org/.

sentences, our annotators met several times with members of the team to seek advice and compare annotations. Upon satisfactory completion of the training sentences, the treebank annotation began.

4.3 Annotation procedure

Annotation was carried out using the web-based annotation tool Arborator[22] (Gerdes, 2013). Arborator uses a simple mouse-based graphical interface with movable arrows and drop-down menus to create labeled dependency trees. In order to expedite the annotation process, we leveraged existing POS tags and dependency trees over the utterances, which were automatically parsed using the transition-based parser of Sagae et al. (2010), and converted to UD through a method based on simple tree regular expressions, using the DepEdit tool[23] (Peng & Zeldes, 2018). The annotator's task was then to hand-correct the dependency relations and POS tags as appropriate. The code for preprocessing the data and for converting CHILDES dependencies to approximate UDs is freely available online.

Figure 5 presents a pre-annotated sentence given to our annotators through the Arborator interface. Annotations that the annotators were unsure of were marked as problematic in the annotation tool. Hard cases were extracted and discussed among the members of the team.

5 Statistics and evaluation

In order to evaluate the self-consistency of the compiled corpora we first measure the agreement between the annotators. For this purpose, each annotator was assigned a longitudinally contiguous sample of 500 utterances in each of the languages they worked on. The starting point of the annotation was the initial converted parser output (see Sect. 4.3).

For both Adam and Hagar, we find fairly high agreement scores comparable with those reported in the literature for English dependency annotation (Berzak et al., 2016a), and somewhat higher than the ones reported for low resource languages (Dirix et al., 2017; Nguyen, 2018). 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. About 0.4% in the LAS agreement in English is lost due to passive constructions occasionally not marked as such by the Hebrew annotators, possibly due to Hebrew UD not using the passive subject (*nsubjpass*) relation. 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.

[22] https://github.com/Arborator.

[23] https://gucorpling.org/depedit/.

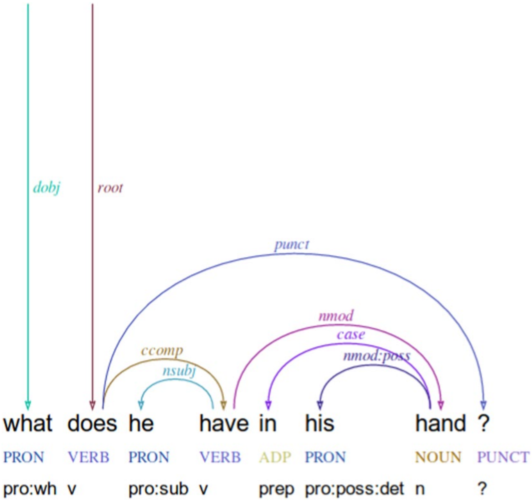

dobj
root
punct
nmod
case
nmod:poss
ccomp
nsubj
what does he have in his hand ?
PRON VERB PRON VERB ADP PRON NOUN PUNCT
pro:wh v pro:sub v prep pro:poss:det n ?

Fig. 5 A screenshot from the Arborator annotation interface, displaying an automatically converted UD parse, which was later hand-corrected

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% on the Hebrew corpus. We further evaluate the quality by manually evaluating the LFs of a sample of 100 utterances in English and 100 in Hebrew. We find that 82% of the English LFs in both English and Hebrew are correct. The LFs we judge to be incorrect generally exhibit at least one of the problems discussed in Sect. 3.3.

Table 2 presents statistics of the corpora, including the frequency per token of dependency labels in the full UD annotated corpus as well as in the portion of the corpus which was successfully converted to LF. It should be noted that an occurrence of a dependency type is counted as not converted if the sentence which contains it is not converted. It does not necessarily mean that this particular dependency was the source of the problem. Therefore the conversion rate of a dependency is only a noisy measure of how difficult a given construction is for the converter.

6 Corpus analyses

This section provides some initial analyses of both the syntactic and semantic aspects of our corpora. While simple, we hope these analyses, together with the modelling study in Sect. 7, will provide inspiration to other researchers regarding some of the questions that can be examined using these resources. Results here are mostly intended to demonstrate the potential utility of the proposed dataset, and should be interpreted with caution, taking into account the inter-annotator agreement (see Sect. 5). The analysis is based on the dependency structures, rather than LFs. The reason for doing so is that UD is annotated over other, non-CDS corpora in both English and Hebrew, which allows comparing the statistics of the compiled corpora to those of existing ones.

Table 2 Dependency label counts and proportion of dependencies which were successfully converted to LFs for the Adam corpus (left) and Hagar corpus (right)

Adam			Hagar		
Dep label	Count	% converted	Dep label	Count	% converted
list	56	96	csubjpass	1	100
xcomp:promoted	40	82	vocative	1863	71
quant	11	82	ccomp:promoted	14	71
nmod:poss	1567	78	compound:svc	128	70
vocative	805	77	root	18,783	68
aux	6048	76	acl:promoted	3	67
parataxis:repeat	48	75	cop	580	66
root	14,724	74	dislocated	165	65
det	6278	74	dobj	6171	64
punct	14,858	74	parataxis:repeat	960	64
nsubj	12,816	74	nsubj	14,401	62
dep	23	74	punct	29,104	61
discourse	2412	74	nmod:smixut	615	60
dobj	6774	74	nmod	8160	59
amod	1138	74	name	132	59
dislocated	11	73	det	9439	59
nummod	259	73	amod	2185	58
case	4434	72	case	12,990	57
cop	3122	72	compound:prt	21	57
neg	2306	72	ccomp	1309	57
nmod	3617	71	acl:relcl:subj	151	56
compound:svc	59	71	compound:smixut	32	56
name	183	70	acl:relcl:obj	228	54
acl:relcl:obj	153	70	nmod:poss	958	54
dobj:promoted	128	70	advmod	6762	54
advmod	5194	69	discourse	3892	52
compound:prt	288	68	nummod	199	50
compound	1003	68	fixed	8	50
iobj	262	67	advcl:promoted	8	50
acl:promoted	3	67	root:promoted	359	50
remnant	12	67	mark	2243	48
goeswith	42	67	cc	3470	46
nmod:promoted	39	64	parataxis	3876	46
acl:relcl:subj	114	63	xcomp	1848	45
ccomp	1140	63	list	43	44
csubj	8	63	goeswith	44	43
nsubj:promoted	18	61	nsubjpass	48	42
parataxis	459	59	reparandum	257	40
mark	1973	59	remnant	32	38
advcl	568	58	advcl	573	35

Table 2 (continued)

Adam			Hagar		
Dep label	Count	% converted	Dep label	Count	% converted
xcomp	1405	57	mwe	831	35
reparandum	42	52	nmod:promoted	127	35
nsubjpass	35	51	csubj	93	33
auxpass	32	50	det:predet	157	33
comp	2	50	conj	1911	30
nmod:npmod	26	46	xcomp:promoted	7	29
cc	881	44	compound	492	28
det:predet	64	42	expl	69	28
mwe	85	40	aux	170	19
nmod:tmod	128	39	nmod:tmod	158	18
expl	211	38	acl	42	17
ccomp:promoted	32	34	dep	516	12
acl	72	33	nsubj:promoted	120	10
conj	675	32	dobj:promoted	135	7
advcl:promoted	7	29	appos	267	4
root:promoted	107	21	case:gen	36	3
appos	26	8	acl:relcl	85	1
csubjpass	1	0	csubj:promoted	2	0
acl:relcl	100	0	acl:relcl:subj:promoted	1	0

Ordered by % of occurrences converted

6.1 Analyses of syntactic dependencies

This section highlights some of the benefits of using the Universal Dependencies scheme. In particular, since this scheme is also used for adult-directed language, we can quantify some of the differences between our child-directed corpora and existing text corpora (Sect. 6.1.1). Perhaps of more interest to language acquisition researchers, the cross-linguistic consistency of the UD scheme also permits direct comparisons between the input to the child in different languages, as we demonstrate in Sect. 6.1.2. Longitudinal analyses are also possible, as shown in Sect. 6.1.3.

While our analyses are very simple frequency comparisons, other researchers might be interested in more subtle analyses, for example using the UD annotations to search for particular constructions of interest in one or more languages, to analyze these in more detail.

6.1.1 Comparison to general corpora of English and Hebrew

The dependency statistics of our CHILDES corpora can be compared to those of general treebanks of written English and Hebrew, English Web Treebank (Silveira et al., 2014) and Hebrew Dependency Treebank (HDT; McDonald et al., 2013;

Tsarfaty, 2013) respectively. Statistics are based on the entire corpora, ignoring the split into training, development and test sets. We focus our study comparison on the dependency annotation (rather than the LFs), as dependency structures decompose straightforwardly to atomic elements that can be counted and compared, and thus lend themselves more easily to statistical analysis.

Adam. As can be seen in Fig. 6a, not many dependency types are more frequent in child-directed language than in general English. The Adam corpus exhibits a higher prevalence of discourse phenomena and direct address to the interlocutor (*vocative*), which is explained by virtue of it being a corpus of conversational spoken language.

The higher frequency of basic relation types (*root, punct, nsubj, dobj*, and *aux*) is a result of the sentences being shorter than in the EWT corpus (a mean of 5.9 tokens per sentence as compared to 15). We also note that negation is more frequent in our corpus than in general English, and so is adverbial modification. The latter is perhaps attributable to a large number of questions about "why" and "how" in our corpus. Structures markedly more common in general English include adjectival modification, conjunction, compounding, prepositional phrases, clausal modifiers and passive voice. A slight difference is observed in the frequency of determiner use, possibly reflecting the fact that in the child-directed corpus we find many examples of naming things or affirming the child's utterance in which bare nominals are used, e.g. "Yes, scout", "Ice for boys and girls".

Hagar Comparing the Hagar corpus to HDT (Fig. 6b) we again observe higher frequency of the core dependencies in child-directed language because of the difference in average utterance length (an average of 6.4 tokens per sentence in the Hagar corpus and 19 tokens in HDT). The more discursive nature of the Hagar corpus is reflected in the higher prevalence of the *parataxis* and *discourse* relations. As in the case of English, negation and adverbial modification are slightly less frequent in general Hebrew. In contrast to English, however, the *aux* relation is more common in HDT than in the Hagar corpus. In Hebrew UD auxiliaries often express modality or aspect, which might characterize news text (source of HDT data) more than child-directed language.[24]

Similarly to English, general Hebrew displays noticeably higher frequencies of adjectival modification, conjunction, compounding, prepositional phrases, and clausal modifiers, but also possessives and indirect objects. The difference in *iobj* frequency might be attributable to the HDT corpus assuming different annotation guidelines and using the *iobj* label where we use *nmod*.[25] The frequency of determiner use is much higher in HDT, which may be explained by the lower frequency of *amod* and *nmod* in Hagar. These two dependency relations are the most common edge labels of the determiner heads in HDT (over 60% of the total number of such edges).

[24] In the Adam corpus the most common auxiliaries are forms of *do be*, and *can*, reflecting the high frequencies of do-support, copular constructions, asking the child to do something or answering questions about things being possible or allowed.

[25] In English, *iobj* applies to the first object in the double object construction (e.g. *Give* ***me*** *the book*). In Hebrew, this construction is very uncommon, and ditransitives in English are often translated to a dative PP (like *Give the book* ***to me***).

Fig. 6 Comparison of dependency type prevalence in child-directed speech and standard UD corpora of ▶ the same language. The plots show only dependencies with a difference in count per token of > 0.005 between each CDS corpus and its paired general text corpus. In each plot, dependencies are sorted according to the size of this difference: starting from the left are the dependencies with greater prevalence in CDS (sorted from larger to smaller differences with general text), followed by those with greater prevalence in general text (again, sorted from larger to smaller differences with CDS)

6.1.2 Comparison of Adam and Hagar corpora

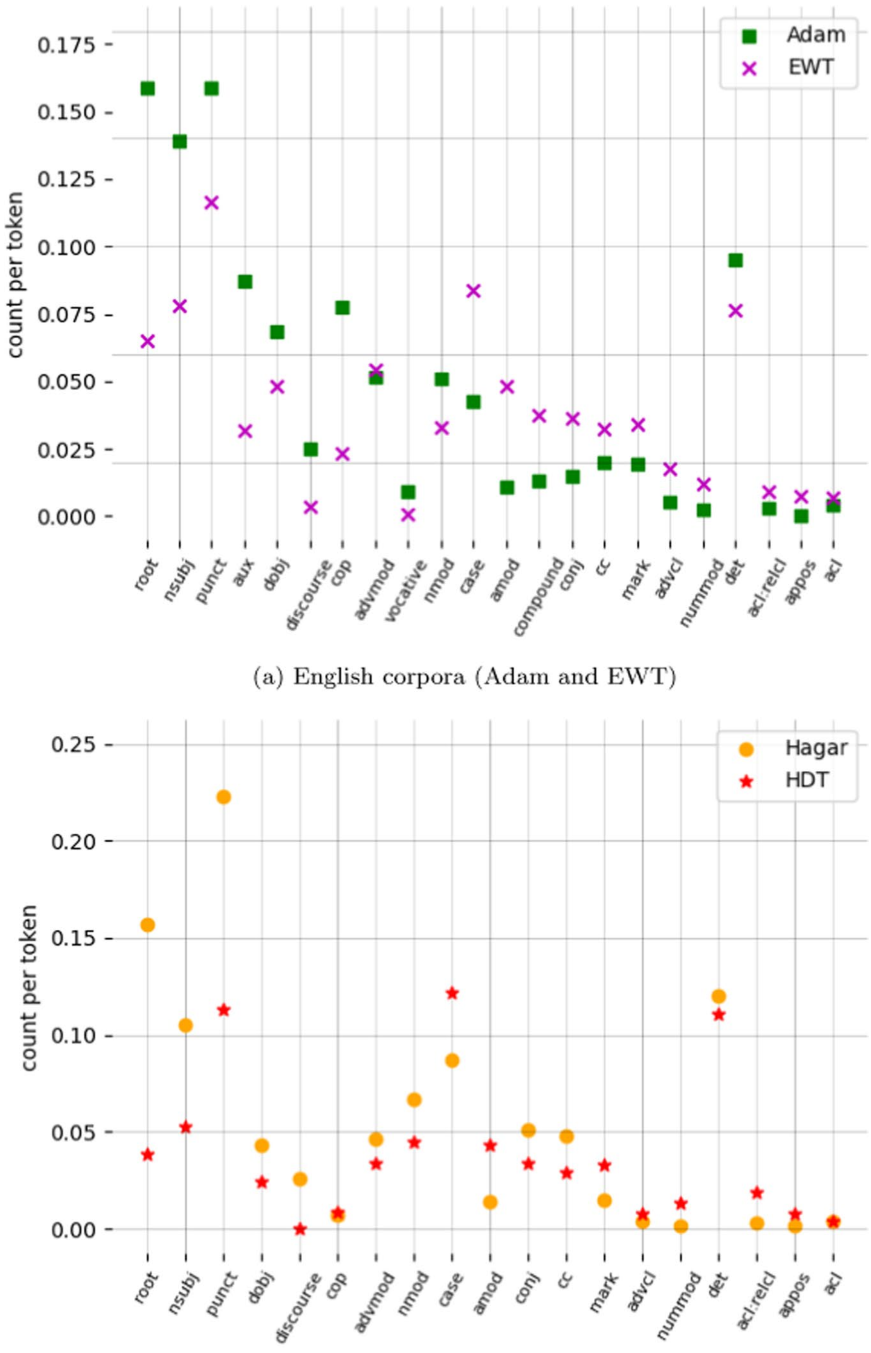
Figure 7 compares the two CHILDES corpora. There are not many notable differences between the frequency of occurrences of particular dependency relations between the English and Hebrew corpora. Sentences in the Adam corpus are on average shorter (5.9 tokens per utterance as compared to 6.4), which is reflected in the higher frequency of *root*, *nsubj*, and *dobj* relations in Fig. 7. The difference in *nsubj* is likely also related to Hebrew being a pro-drop language. Other differences also reflect diverging properties of the two languages: prevalence of *cop* in English is higher, because Hebrew lacks an overt copula; prevalence of *aux* is higher in English, since tense, which accounts for many of the *aux* instances in English, is encoded morphologically in Hebrew; prevalence of *case* and *nmod* in Hebrew is higher likely because of indirect objects being expressed using case markers.

Other observed differences, like more negation and possessives in English or more adjectives, conjunctions, and parataxis in Hebrew, might be idiosyncratic to the speakers. Other differences may be due to different transcription conventions. For instance, the Hebrew corpus contains markedly more commas.

6.1.3 Longitudinal analysis of syntactic dependencies

Taking advantage of our chronologically ordered data we inspect the changes in frequency of use of particular dependency labels over time. For each dependency and each session, we calculate the proportion of sentences which include that dependency. We check for the existence of longitudinal trends by examining whether the child's age is a significant predictor, in a linear regression model, of the frequency of each dependency. Below we discuss dependencies which exhibit a trend with $p < .01$.

In the Adam corpus, we find a significant increase in the use of the following constructions as the child gets older: adjectival clauses, object and “other” (i.e., not subject or object) relative clauses, and ellipsis affecting nouns in prepositional phrases (Fig. 8). In the Hagar corpus longitudinal changes are much more widespread (Fig. 9). The point of commonality is relative clauses—in the case of the Hagar corpus there are upward trends for subject and object relatives. The following constructions also significantly increase in use with time: adverbial clauses, adjectives, numerical and possessive modifiers, multiword expressions, disfluencies (reparandum), transitive verbs (direct object), conjunction, adverbs, subordinate clauses (mark), clausal complements, negation, and prepositional phrases, which in our annotation include all indirect objects.



(a) English corpora (Adam and EWT)

(b) Hebrew corpora (Hagar and HDT)

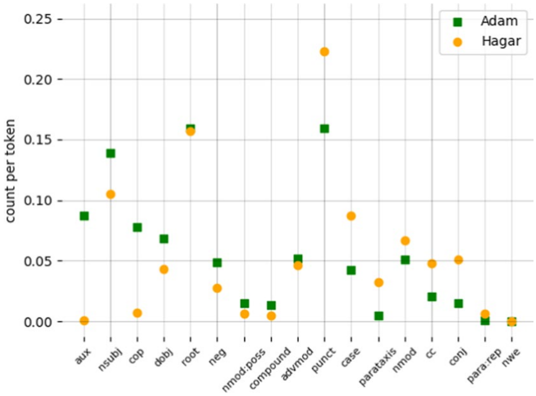
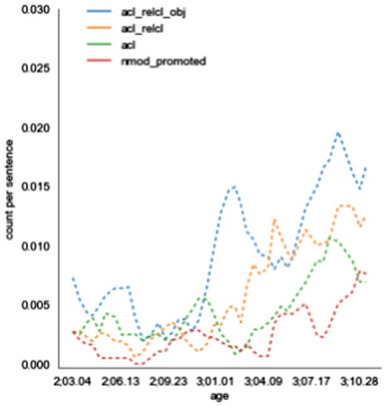


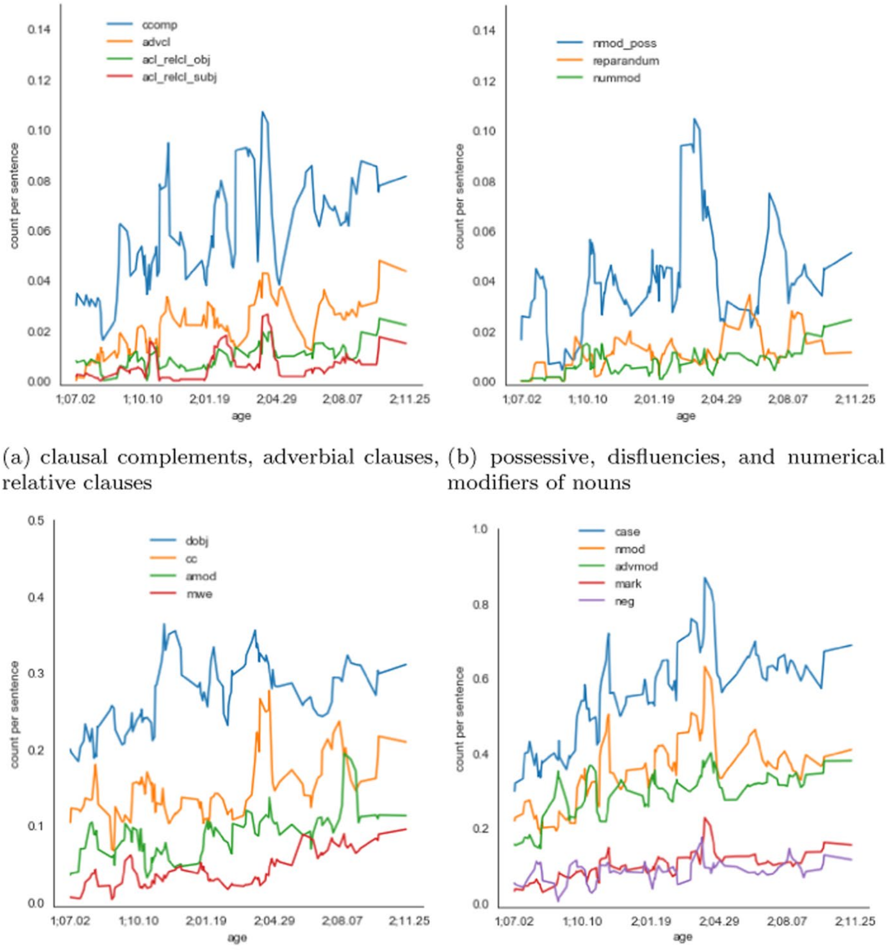
Fig. 7 Comparison of dependency type prevalence between the English and Hebrew CDS corpora. The plots show only dependencies with a difference in count per token of > 0.005, and are sorted according to the size of this difference: starting from the left are the dependencies with greater prevalence in the Adam corpus (sorted from larger to smaller differences with Hagar), followed by those with greater prevalence in the Hagar corpus (again, sorted from larger to smaller differences with Adam)

6.2 Longitudinal analysis of semantic complexity

As well as syntactic analyses, our corpora provide meaning representations, which allow additional types of research questions. Again, we provide just a simple proof of concept here, investigating whether the semantic complexity of the adults' utterances increases as the child gets older. Future work may wish to conduct other types of analyses that are not explicit in the UD syntax but are exposed by the LFs, such as statistics on the valency of different predicates, or the scope of quantifiers.

Fig. 8 Dependencies displaying an upward longitudinal trend in frequency in the Adam corpus. Frequencies are smoothed over 5 sessions





(a) clausal complements, adverbial clauses, relative clauses

(b) possessive, disfluencies, and numerical modifiers of nouns

(c) direct object, conjunction, adjectives, multiword expressions

(d) prepositional phrases, adverbs, subordinate clauses, negation

Fig. 9 Dependencies displaying an upward longitudinal trend in frequency in the Hagar corpus. Frequencies are smoothed over 5 sessions. (The grouping of dependencies is not meaningful but merely increases legibility)

6.2.1 Semantic complexity measures

In the context of this corpus analysis, we propose a very constrained definition of semantic complexity. We consider complexity in the sense of structural complexity of the predicate-argument relationships in the utterance—the depth of nesting and the number of predicates, arguments and modifiers.

The most pertinent question when it comes to the longitudinal analysis of the semantic complexity of CDS is whether the adult utterances express increasingly complex meanings as the child gets older. There are many axes on which complexity

The full LF for the sentence *"What happened to your finger?"* is shown below:

$\lambda a.\ \lambda e_1.\ and(happened_{e_1}(a),\ to_{e_1}(\text{YOUR}\ x[finger(x)]))$

From it, we extract the following five sub-expressions (corresponding to the dashed boxes in the tree):

- $finger(x)$
- $\text{YOUR}\ x[finger(x)]$
- $to_{e_1}(\text{YOUR}\ x[finger(x)])$
- $happened_{e_1}(a)$
- $\lambda e_1.\ and(happened_{e_1}(a),\ to_{e_1}(\text{YOUR}\ x[finger(x)]))$

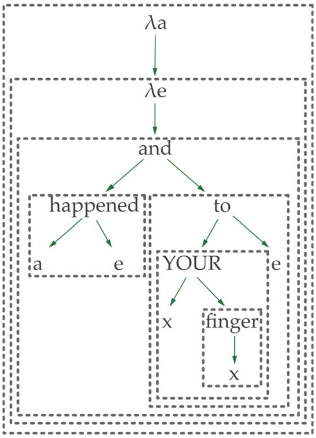


Fig. 10 Extracting sub-expressions from an example LF. If one views the LF as a tree in which variable bindings, predicates, and variables are nodes, then sub-expressions correspond to sub-trees of that tree (indicated by dashed boxes). The number of sub-expressions reflects both the branching factor and nesting level of the tree

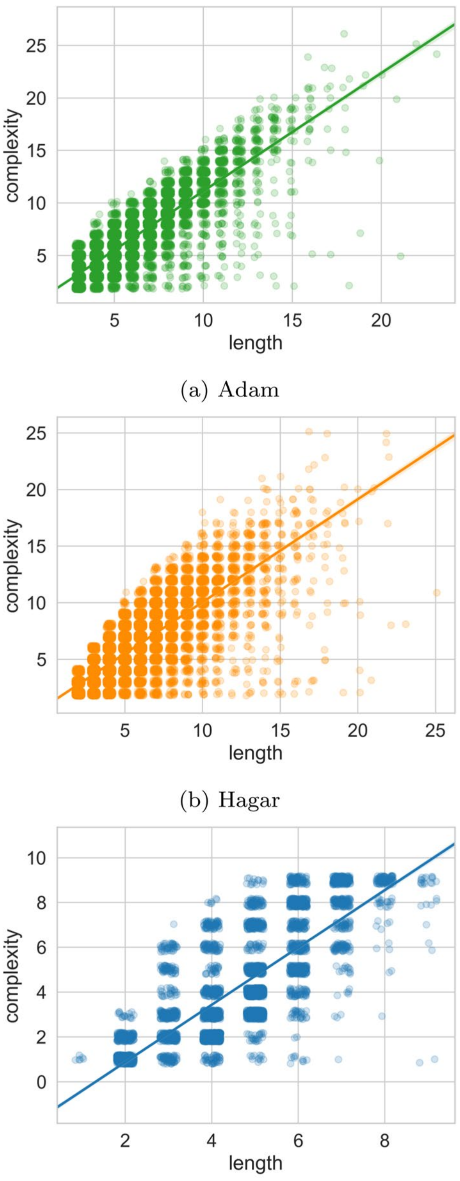
can increase—concepts being more abstract, referents of expressions being less contextually obvious, language being more metaphorical, etc. Our newly available data creates the opportunity to study this question in terms of sentential predicate-argument structures. That is, the question we answer in this analysis is: does the predicate-argument structure of the CDS grow more complex as the child grows older?

One way to approach the issue is to count the number of sub-expressions in the LF. For example, from the LF of the sentence "What happened to your finger?", we obtain five sub-expressions, as illustrated in Fig. 10.

It is expected that the number of sub-expressions will correlate strongly with the number of tokens in the utterance. For both corpora it is indeed the case, as can be seen in Fig. 11 with the Pearson's coefficient of $r = 0.74$ for Adam and $r = 0.76$ for Hagar. Even though the correlation is strong, we can also observe a relatively wide spread of complexities for any given value of length. The coefficient of determination in OLS regression shows that utterance length accounts for 54.1% of variation in complexity in Adam and 58.1% in Hagar. This indicates that our complexity measure captures information beyond just the number of tokens in an utterance.

To illustrate the improvement of our automated approach of LF generation over a more restricted dataset, and in particular to highlight the usefulness of the relatively high coverage of syntactic constructions in our transducer, we also analyse the semantic complexity of Brown's Eve dataset (Brown, 1973). To this end we use the transduced LFs by Abend et al. (2017), which were created semi-automatically from the morphosyntactic annotation of Sagae et al. (2010), and filtered to only include utterances of length up to 10, due to the limitations of their conversion method. Results (Fig. 11) present a seemingly even stronger correlation and less variation on the Eve dataset than for Adam and Hagar. In the case of the Eve corpus, utterance length accounts for 66.2% of the variation in complexity.

Fig. 11 Relationship between LF complexity (number of sub-expressions) and utterance length (number of tokens). Each point represents an individual utterance in the corpus. Solid lines illustrate the linear regression line and the shaded region around the lines the 95% confidence interval for that regression (very tight in all 3 graphs)



(a) Adam

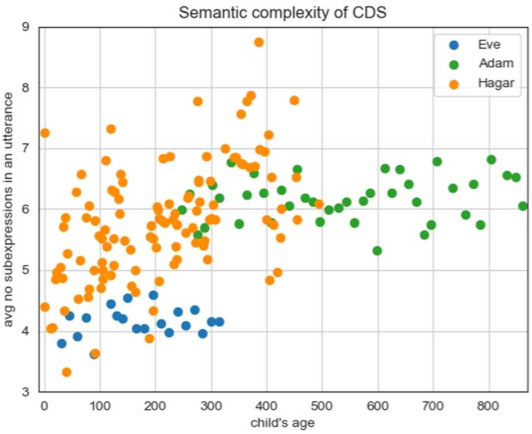
(b) Hagar

(c) Eve

6.2.2 Longitudinal analysis

Does CDS complexity change with the child's age? Figure 12 shows the distribution of semantic complexity (averaged over all utterances in a session) relative to the child's age. While the complexity in the Adam corpus remains relatively stable over time, there might be an upwards longitudinal trend in Hagar. Pearson's coefficient

Fig. 12 The average complexity of child-directed utterances in a session plotted against the child's age in days



confirms a weak correlation between average complexity and child's age in Hagar ($r = .35$, $p < 0.001$) and no correlation in Adam ($r = .11$, $p = .5$). Looking further into the Hagar corpus, the OLS regression's coefficient of determination indicates that the child's age accounts for 11% of the variation in complexity beyond that explained by utterance length alone. Interestingly, Fig. 12 also suggests that the LFs in the Eve corpus might not adequately reflect the longitudinal changes in CDS utterance complexity. The Hagar corpus presents an increase in semantic complexity at the age range covered by Eve, and we therefore would have expected to see a similar one in Eve. The fact that such a trend is not observed is probably due to some limitation of the Eve LFs, which were formed by a different extraction method to ours (see Sect. 6). We would expect an increase in meaning complexity in line with the child's cognitive development over this age range. The fact that our LFs show an increase in complexity may suggest that they capture the relevant semantic information in the text, and if so, this is evidence for the superiority of our method over the method used for compiling the Eve LFs dataset, which does not reflect such a trend. However, our results do not allow us to decide whether this is the case or not, and there may be individual or cultural differences across our data.

7 LF annotated corpora as data for acquisition simulations

This section presents a set of preliminary experiments that demonstrate how the presented corpora may be used for simulations of the learning dynamics that resemble child language acquisition processes in children. The cross-linguistic consistency of the scheme allows us to evaluate the cross-linguistic applicability of the model, which is essential for establishing the validity of an acquisition model.

7.1 The learning model: an outline

We adapt the language acquisition model of Abend et al. (2017) to the proposed LFs. The model is a computational implementation of the semantic bootstrapping hypothesis (Bowerman, 1973; Pinker, 1979), whose goal is to generalize from input pairs of observed utterances and inferred meanings in order to interpret new utterances whose meaning is unavailable contextually. Unlike "parameter-setting" approaches (e.g., Yang, 2002), we do not assume that the grammar of natural languages can be described by a finite number of finitely-valued parameters. Instead, the proposed model searches a structured space of all possible grammars as defined by an established formal theory of the syntax–semantics interface—CCG.

The proposed model employs Bayesian learning to jointly model (a) learning of the lexicon: the mapping between words (or generally: any portion of the input string) and portions of the sentential meaning, and (b) syntax learning: the rules governing the combination of the lexical elements into utterances. By jointly modeling lexical learning and syntactic acquisition, and assuming that the inferred meanings available to the child are at the level of utterances rather than individual words, the model provides a working account of how these two aspects of language can be learned simultaneously in a mutually reinforcing fashion.

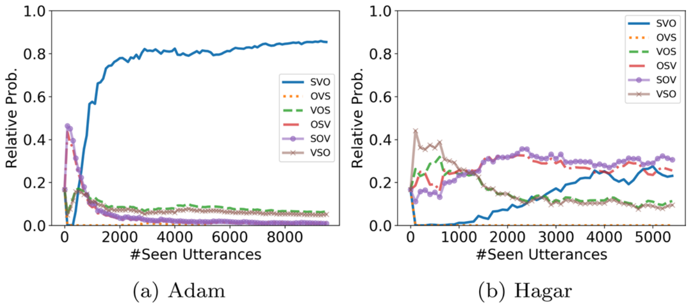
7.2 Learning of word order

Both English and Hebrew are regarded as languages with Subject-Verb-Object as their basic word order. We report here experiments that show that when experimenting with the learner on the proposed corpora, the probability of SVO indeed increases during learning.

We run the learner on the Adam and Hagar corpus, with their corresponding LFs, and compare our results to those reported by Abend et al. (2017) for the Eve corpus (using length-bounded sentences and using a different, semi-automatic approach for generating the LFs). Experiments are performed without introducing any intentional noise to the training data, and therefore correspond to their "No Distractors" setting.

Figure 13 presents our results. On Adam the model learns that English transitive sentences are SVO; learning curves are steep, despite the lack of an explicit signal. Comparing the trends to the ones reported on Eve by Abend et al. (2017), we find that while SVO emerges as the overwhelmingly most probable order in both cases, in the simulation based on the Adam corpus, other orders are considered more probable for around the first 1000 utterances, while with the Eve corpus, SVO overwhelms other hypotheses within the first 100 sentences. It appears that the Eve corpus is too limited in terms of sentence structure variety and complexity to allow for examining the period of acquisition before the basic word order is determined in the mind of the learner.

In Hebrew, learning is considerably more gradual. In fact, after training on 4000 utterance-LF pairs, the model has managed to demote the (incorrect) VSO, VOS and OSV orders, but remains indecisive as to whether SVO or the verb-final orders are correct. A steady increase, however, is presented in the probability of the correct



(a) Adam

(b) Hagar

Fig. 13 Learning that English and Hebrew are SVO languages. Plots show the relative posterior probability assigned by the model to the six possible categories of transitive verbs

SVO order. This more gradual trajectory may be due to the more flexible word order presented by Hebrew, as opposed to the relative rigidity of English.

We report on these experiments in order to illustrate the potential usefulness of the corpora and LFs reported here. A deeper investigation of these and other trends in the acquisition of grammar is needed in order to draw cognitive conclusions from such experiments.

8 Discussion

Having demonstrated the utility of our resource, we consider limitations and opportunities for future extensions.

First, we note that the approach of annotating dependency syntax and automatically transducing it to logical forms is practical but not perfect. We have discussed limitations of the logical forms (Sect. 3.3) and estimated the error present in syntactic annotation and conversion to semantics (Sect. 4). We believe the accuracy is sufficient for examining broad trends (e.g., over the course of acquisition, as in the pilot study in Sect. 5). But further work on the representations may be required to support research that relies on grounding in a world model, for example.

Second, the style of semantic representation (as rather conventional logical forms in the formal semantics tradition) is suited to some modes of investigation but not all. Other semantic representations that exist in broad-coverage corpora might better capture elements such as discourse context (Kamp & Reyle, 1993), lexical semantics (Banarescu et al., 2012; Pustejovsky, 1998), or typologically motivated scene structures (Abend & Rappoport, 2013). Future studies might profit from enriching our data with such representations, building on the LFs that are there to expose syntactically nonlocal semantic dependencies.

Third, the corpus has been designed to facilitate research on the semantic bootstrapping hypothesis. As such, semantic representations are provided for child-directed speech, in order to simulate the meaning that is presumably available to the child in an interaction. Our focus on child-directed speech is typical of much of the acquisition literature using CHILDES data (*inter alia* Fazly et al., 2010; Huebner et al., 2021; Perfors et al., 2011; Yedetore et al., 2023). Nevertheless, some lines of research may benefit from syntactic and semantic annotation of child utterances as well. Annotating child language is difficult because it requires interpretation of utterances exhibiting non-mature syntax, and guidelines to support this (cf. UD annotation of adult learner syntax; Berzak et al., 2016b). We leave this to future work.

Finally, we have investigated two languages in this study as a case study of the considerations needed for cross-linguistic work with our approach. Two languages are, of course, not sufficient to demonstrate that a representation is "universal" or that annotating any new language will be trivial. But we argue that building upon a highly multilingual syntactic framework (UD) and adopting a fairly neutral representation of meaning (LFs) provides a solid foundation for developing syntactically and semantically rich resources for child-directed speech in new languages, and facilitates cross-linguistic comparison as well.

9 Conclusion

Cross-linguistically consistent linguistic annotation of child-directed speech is essential for corpus studies and computational modeling of child language acquisition. We have presented a methodology for syntactic annotation on CDS using Universal Dependencies and a conversion method for transducing logical forms from the resulting trees. We show that the methodology can be reliably applied to English and Hebrew, and propose a way to address common phenomena in CDS that are scarce in standard UD corpora. We then turn to a discussion of the limitations of the current method, suggesting paths for future improvement. Finally, we apply the proposed methodology to two corpora from CHILDES, the English Adam corpus and the Hebrew Hagar corpus, yielding sizable, cross-linguistically consistent annotated resources.

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., 2017) showed that a model of a child's acquisition of grammar can be induced from semantic annotation of the kind discussed here. We apply their model to the compiled corpora as a preliminary demonstration of the possibility of comparative computational research on grammar acquisition in the two languages.

Appendix: Hebrew transcription conventions

We adopt the transcription conventions used in the Hagar corpus. The consonants and their transliterations are:

א	ʔ
ב	b/v
ג	g
ד	d
ה	h
ו	w
ז	z
ח	ḳ
ט	ṭ
י	y
ך	k/ḳ
ל	l
ם	m
ן	n
ס	s
ע	ʕ
ף	p/f
ץ	c
ק	q
ר	r
שׁ	š
שׂ	ṣ
ת	t

The vowels in use are (stressed and unstressed):

ā ē ī ō ū
a e i o u

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10579-024-09734-y.

Acknowledgements 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 grants by the Israel Science Foundation (Grant Nos. 929/17 and 2424/21) 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, and the contribution of Yael Katz, who partnered in the preprocessing of the Hebrew annotation.

Author contributions I.S. has written the converter code, processed the data, modified the learner code (from the original Abend et al., 2017 code), ran the simulations and analysis and generated the figures and tables. I.S. and L.M. worked on postprocessing the annotated data. I.S. and M.S. have led the design of the LFs. Sa.Gi. has initially led the annotation of Adam, and O.A. coordinated the annotation effort on Hagar, but all authors but L.M. contributed to these efforts. N.S. has written code to process the CHILDES corpora and conducted UD annotator training. All authors partnered in the writing, reviewing and modifying of the paper.

Funding Open access funding provided by Hebrew University of Jerusalem.

Declarations

Conflict of interest The authors have no conflicts of interest that relate to this manuscript.

Graphics program All graphs were created using the Python seaborn package (https://seaborn.pydata.org/).

References

Abend, O., Kwiatkowski, T., Smith, N., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition, 164*, 116–143.

Abend, O., & Rappoport, A. (2013). Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of ACL* (pp. 228–238). http://aclweb.org/anthology/P13-1023

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science, 32*(5), 789–834.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2012). Abstract meaning representation (AMR) 1.0 specification.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, Sofia, Bulgaria* (pp. 178–186). http://www.aclweb.org/anthology/W13-2322

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

Berzak, Y., Huang, Y., Barbu, A., Korhonen, A., & Katz, B. (2016a). Anchoring and agreement in syntactic annotations. In *Proceedings of the 2016 conference on empirical methods in natural language processing, association for computational linguistics* (pp. 2215–2224). https://doi.org/10.18653/v1/D16-1239, http://aclweb.org/anthology/D16-1239

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., & Katz, B. (2016b). Universal Dependencies for learner English. In *Proceedings of ACL, Berlin, Germany* (pp. 737–746). http://www.aclweb.org/anthology/P16-1070

Blodgett, A., & Schneider, N. (2019). An improved approach for semantic graph composition with CCG. In *Proceedings of IWCS, Gothenburg, Sweden* (pp. 55–70). https://www.aclweb.org/anthology/W19-0405

Blodgett, A., & Schneider, N. (2021). Probabilistic, structure-aware algorithms for improved variety, accuracy, and coverage of AMR alignments. In *Proceedings of ACL-IJCNLP, Online* (pp. 3310–3321). https://aclanthology.org/2021.acl-long.257

Bouma, G., Noord, G. V., & Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers, 37*, 45–59.

Bowerman, M. (1973). Structural relationships in children's utterances: Syntactic or semantic? In *Cognitive development and the acquisition of language*. Academic Press.

Bowerman, M. (1974). Learning the structure of causative verbs: A study in the relationship of cognitive, semantic and syntactic development. *Papers and Reports on Child Language Development, 8*, 142–178.

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

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Buttery, P. (2006). Computational models for first language acquisition. PhD Thesis, University of Cambridge.

Culicover, P., & Wilkins, W. (1984). *Locality in linguistic theory*. Academic Press.

Davidson, D. (1967). The logical form of action sentences. In N. Rescher (Ed.), *The logic of decision and action* (pp. 81–95). University of Pittsburgh Press.

de Marneffe, M., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics, 47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Dirix, P., Augustinus, L., van Niekerk, D., & Van Eynde, F. (2017). Universal Dependencies for Afrikaans. In *Proceedings of the NoDaLiDa 2017 workshop on Universal Dependencies* (Vol. 135, pp. 38–47).

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science, 34*, 1017–1063.

Gerdes, K. (2013). Collaborative dependency annotation. In *Proceedings of the second international conference on dependency linguistics* (DepLing 2013) (pp. 88–97).

Gretz, S., Itai, A., MacWhinney, B., Nir, B., & Wintner, S. (2015). Parsing Hebrew CHILDES transcripts. *Language Resources and Evaluation, 49*(1), 107–145.

Groschwitz, J., Lindemann, M., Fowlie, M., Johnson, M., & Koller, A. (2018). AMR dependency parsing with a typed semantic algebra. In *Proceedings of ACL, Melbourne, Australia* (pp. 1831–1841). http://aclweb.org/anthology/P18-1170

Grünewald, S., Piccirilli, P., & Friedrich, A. (2021). Coordinate constructions in English enhanced Universal Dependencies: Analysis and computational modeling. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume, association for computational linguistics, online* (pp. 795–809). https://aclanthology.org/2021.eacl-main.67

Hershcovich, D., Abend, O., & Rappoport, A. (2019). Content differences in syntactic and semantic representation. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers), association for computational linguistics, Minneapolis, Minnesota* (pp. 478–488). https://doi.org/10.18653/v1/N19-1047, https://www.aclweb.org/anthology/N19-1047

Hoff-Ginsberg, E. (1985). Some contributions of mothers' speech to their children's syntactic growth. *Journal of Child Language, 12*(2), 367–385.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational* *natural language learning, association for computational linguistics, online* (pp. 624–646). https://doi.org/10.18653/v1/2021.conll-1.49, https://aclanthology.org/2021.conll-1.49

Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Kluwer.

Kwiatkowski, T., Goldwater, S., Zettlemoyer, L., & Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th conference of the European chapter of the ACL (EACL 2012), ACL, Avignon* (pp. 234–244).

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of LREC, Genoa, Italy* (pp. 2231–2234).

Liu, Z., & Prud'hommeaux, E. (2021). Dependency parsing evaluation for low-resource spontaneous speech. In *Proceedings of the second workshop on domain adaptation for NLP, Kyiv, Ukraine* (pp. 156–165). https://aclanthology.org/2021.adaptnlp-1.16

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Erlbaum.

Mao, J., Shi, F., Wu, J., Levy, R., & Tenenbaum, J. (2021). Grammar-based grounded lexicon learning. In *Proceedings of the thirty-fifth conference on neural information processing systems* (pp. 7865–7878).

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313–330.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., & Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of ACL, Sofia, Bulgaria* (pp. 92–97). http://www.aclweb.org/anthology/P13-2017

McNeill, D. (1966). Developmental psycholinguistics. In F. Smith & G. Miller (Eds.), *The Genesis of Language* (pp. 15–84). MIT Press.

Moon, L., Christodoulopoulos, C., Fisher, C., Franco, S., & Roth, D. (2018). Gold standard annotations for preposition and verb sense with semantic role labels in adult-child interactions. In *Proceedings of the 27th international conference on computational linguistics, association for computational linguistics, Santa Fe, New Mexico, USA* (pp. 3004–3014). https://www.aclweb.org/anthology/C18-1254

Newport, E. (1977). Mother, I'd rather do it by myself: Some effects and non-effects of maternal speech-style. In C. Snow & C. Fergusson (Eds.), *Talking to children: Language input and acquisition* (pp. 109–149). Cambridge University Press.

Nguyen, K. H. (2018). Bktreebank: Building a Vietnamese dependency treebank. In *LREC* (pp. 2164–2168).

Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC* (pp. 1659–1666).

Odijk, J., Dimitriadis, A., Van der Klis, M., Van Koppen, M., Otten, M., & Van der Veen, R. (2018). The AnnCor CHILDES treebank. In *Proceedings of LREC* (pp. 2275-2283).

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

Peng, S., & Zeldes, A. (2018). All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations. In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018), Santa Fe, NM* (pp. 167–177). https://www.aclweb.org/anthology/W18-4918

Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition, 118*, 306–338.

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

Przepiórkowski, A., & Patejuk, A. (2019). Nested coordination in Universal Dependencies. In *Proceedings of the third workshop on Universal Dependencies (UDW, SyntaxFest 2019), Association for Computational Linguistics, Paris, France* (pp. 58–69). https://doi.org/10.18653/v1/W19-8007, https://aclanthology.org/W19-8007

Pullum, G. K. (1990). Constraints on intransitive quasi-serial verb constructions in modem colloquial English. *Working Papers in Linguistics, 39*, 218–239.

Pustejovsky, J. (1998). *The generative lexicon*. MIT Press.

Reddy, S., Täckström, O., Collins, M., Kwiatkowski, T., Das, D., Steedman, M., & Lapata, M. (2016). Transforming dependency structures to logical forms for semantic parsing. *Transactions of the Association for Computational Linguistics, 4*, 127–140.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language, 37*, 705–729.

Sanguinetti, M., Cassidy, L., Bosco, C., Çetinoğlu, O., Cignarella, A. T., Lynn, T., Rehbein, I., Ruppenhofer, J., Seddah, D., & Zeldes, A. (2020). Treebanking user-generated content: A UD based overview of guidelines, corpora and unified recommendations. In Proceedings of LREC (pp. 5240–5250).

Savary, A., Stymne, S., Barbu Mititelu, V., Schneider, N., Ramisch, C., & Nivre, J. (2023). PARSEME meets Universal Dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*. https://doi.org/10.3384/nejlt.2000-1533.2023.4453

Schuster, S., & Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of LREC, ELRA*. https://nlp.stanford.edu/pubs/schuster2016enhanced.pdf

Silveira, N., Dozat, T., Marneffe, M. D., Bowman, S. R., Connor, M., Bauer, J., & Manning, C. D. (2014). A gold standard dependency corpus for English. In Calzolari, N., Choukri, K,, Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., & Piperidis, S. (Eds.), *Proceedings of LREC* (pp. 2897–2904). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1089_Paper.pdf

Steedman, M. (2000). *The syntactic process*. MIT Press.

Szubert, I., Lopez, A., & Schneider, N. (2018). A structured syntax-semantics interface for English-AMR alignment. In *Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long papers), Association for Computational Linguistics, New Orleans, Louisiana* (pp. 1169–1180). https://doi.org/10.18653/v1/N18-1106, https://aclanthology.org/N18-1106

Tsarfaty, R. (2013). A unified morpho-syntactic scheme of Stanford dependencies. In *Proceedings of ACL, Sofia, Bulgaria* (pp. 578–584). https://www.aclweb.org/anthology/P13-2103

Van Gysel, J. E. L., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C., Hajič, J., Martin, J. H., Oepen, S., Palmer, M., Pustejovsky, J., Vallejos, R., & Xue, N. (2021). Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz, 35*(3), 343–360. https://doi.org/10.1007/s13218-021-00722-w

Villavicencio, A. (2002). The acquisition of a unification-based generalised categorial grammar. PhD thesis, University of Cambridge.

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford University Press.

Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. Proc. of ACL (9370–9393)

Authors and Affiliations

Ida Szubert[1] · Omri Abend[3] · Nathan Schneider[4] · Samuel Gibbon[2] · Louis Mahon[1] · Sharon Goldwater[1] · Mark Steedman[1]

✉ Omri Abend
omri.abend@mail.huji.ac.il

[1] School of Informatics, University of Edinburgh, Edinburgh, UK

[2] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

[3] School of Computer Science and Engineering and the Department of Cognitive Science, Hebrew University of Jerusalem, Jerusalem, Israel

[4] Departments of Linguistics and Computer Science, Georgetown University, Washington, D.C., USA