The AnnCor CHILDES Treebank

Jan Odijk,Alexis Dimitriadis,Martijn van der Klis,Marjo van Koppen,Meie Otten,Remco van der Veen
Utrecht University
Trans 10 3512 JK Utrecht
{j.odijk, a.dimitriadis, m.h.vanderklis, j.m.vankoppen, m.otten1, r.p.vanderveen}@uu.nl

Abstract

This paper (1) presents the first partially manually verified treebank for Dutch CHILDES corpora, the AnnCor CHILDES Treebank; (2) argues explicitly that it is useful to assign adult grammar syntactic structures to utterances of children who are still in the process of acquiring the language; (3) argues that human annotation and automatic checks on this annotation must go hand in hand; (4) argues that explicit annotation guidelines and conventions must be developed and adhered to and emphasises consistency of the annotations as an important desirable property for annotations. It also describes the tools used for annotation and automated checks on edited syntactic structures, as well as extensions to an existing treebank query application (GrETEL) and the multiple formats in which the resources will be made available.

Keywords: treebank, Dutch, CHILDES, GrETEL, treebank querying

1. Introduction

We describe the approach to the development of the AnnCor CHILDES Treebank, a treebank for Dutch CHILDES corpora. The whole treebank has been automatically generated by the Alpino parser, but a part of this corpus has also been manually checked and, where needed, corrected. The AnnCor treebank is being created in the Utrecht University AnnCor project, which we describe in section 2..[1] We describe the syntactic annotation in this treebank in section 3.: we discuss some methodological issues with regard to the annotation (section 3.1.), the cleaning of CHILDES utterances (section 3.2.), the tool we use to inspect and edit syntactic structures (section 3.3.), annotation conventions we developed (section 3.4.), and checks on edited syntactic structures (section 3.5.). We have extended an existing treebank query application, and the extensions are described in section 4.. In section 5. we describe the formats in which the resources will be made available. In section 6. we discuss related work, and we end with conclusions and plans and suggestions for future work in section 7..

This paper (1) presents the AnnCor CHILDES Treebank, the first partially manually verified treebank for Dutch CHILDES corpora; (2) argues explicitly that it is useful to assign adult grammar syntactic structures to utterances of children who are still in the process of acquiring the language; (3) argues that human annotation and automatic checks on this annotation must go hand in hand to combine the human's intelligence and the software's rigor; (4) argues that explicit annotation guidelines and conventions must be developed and adhered to. Requiring consistency of the annotations is important and should be an essential ingredient of such annotation conventions.

2. The AnnCor Project

The AnnCor project[2] is an Utrecht University internal research infrastructure project that aims to create linguistically annotated corpora for the Dutch language and to enhance and extend an existing treebank query application in order to query the annotated corpora. Various types of corpora are being annotated, and various types of annotations are added. The corpora include learner corpora (texts produced by pupils at primary school), news corpora, narrative corpora, and language acquisition corpora (in particular, natural spoken interactions between parents and children). Annotations include annotations for learners' errors and their corrections, discourse annotations, and full syntactic structures. In this paper we focus on the creation of treebanks (i.e. text corpora in which each utterance is assigned a syntactic structure) for language acquisition data, in particular the Dutch CHILDES corpora (MacWhinney, 2000).[3] The CHILDES corpora contain annotated orthographic transcriptions of the interaction between multiple speakers, usually a target child and other participants (e.g the child's mother).
According to (Sagae et al., 2007), 'linguistic annotation of the corpora provides researchers with better means for exploring the development of grammatical constructions and their usage'. The research described in (Odijk, 2015; Odijk, 2016a) illustrates this for the study of the acquisition of particular syntactic modification and complementation phenomena using the Dutch CHILDES corpora. It is clear from these papers that such research cannot be done properly and efficiently without treebanks for these corpora. This can be illustrated with a simple example. If one wants to study the behavior of Dutch words such as *heel*, *erg* and *zeer* (all meaning 'very'), it is not sufficient to search for these strings: one will miss inflected variants such as *hele* and *erge* and find irrelevant occurrences of these words in different uses and meanings (such as *zeer* meaning 'pain'

[1] This paper contains many hyperlinks hidden under terms and acronyms. The presence of a hyperlink is visible in digital versions of the paper but may be badly visible or invisible in printed versions of the paper.

[2] `https://anncor.sites.uu.nl/`.

[3] Accessible via `http://childes.talkbank.org/access/Dutch/`.

or ‘painful’). Significantly better search results are obtained when one can search in a treebank containing these words. Ambiguity is not restricted to these words: very many words are highly ambiguous, in particular the words that linguists are most interested in. Many of these ambiguities are resolved in treebanks. The AnnCor project aims to create exactly such treebanks, which, together with query applications, will become an integrated part of the Dutch part of the CLARIN research infrastructure (Odijk, 2016b; Odijk and van Hessen, 2017). These treebanks and the associated search and analysis applications can then contribute to an acceleration of language acquisition research and to a larger empirical basis for theories or hypotheses, thus providing a basis for carrying out groundbreaking research in which old questions can be investigated in new ways and new questions can be raised and investigated for the first time.

3. Syntactic Annotation

The syntactic annotation is added using the Alpino parser (Bouma et al., 2001). Since Alpino has been developed for written adult language such as newspapers, it is not surprising that it creates many wrong parses. The problem is twofold: CHILDES contains transcriptions of spoken utterances from a dialogue, and many of them are uttered by children that are still in the process of acquiring the language. The fully automatically generated parses can be inspected and queried in the PaQu application (Odijk et al., 2017).[4] In the AnnCor project we create a manually verified subcorpus, with a targeted size of 20% of the Dutch CHILDES *Van Kampen* subcorpus, sampled in a representative manner by selecting a contiguous subpart of each session of the subcorpus.[5] In addition, we manually verified and, if needed, corrected the parse trees for which it was very likely that they contain errors, as determined on the basis of a variety of heuristics for identifying potential errors. For example, syntactic structures containing nodes labeled as *dp* (discourse part) are often the result of the Alpino robustness module which is used when Alpino is not able to make a single full parse using its normal rules. This may be caused by fragmentary or ungrammatical input, or by parser errors.[6] Another example concerns coordinate structures, which often contain errors. All together we target to have 35% of the corpus manually verified.

3.1. Methodological Considerations

The CHILDES corpora contain orthographic transcriptions of the interaction between a target child and other participants. The latter are usually a parent, caretaker, or investigator, but may include others, among which other children. Parsing the utterances of adult speakers with a parser for a grammar which is supposed to reflect the competence of adult native speakers is unproblematic from a methodological point of view. However, it is not obvious that it makes sense to parse the children’s utterances with a parser for a grammar which is supposed to reflect the competence of adult native speakers. Of course, we do not know what internal grammar the children have (and we have no model for it), and this internal grammar changes over time. In fact, one of the goals of creating the AnnCor Treebank is to enable researchers to gain knowledge about the internal grammar of children. The parses of children’s utterances must not be seen as claims about the syntactic structure assigned by the children’s grammar but as a classification of children’s utterances and their parts in terms of the (best implemented model that we have for the) grammar for adults. We assume that children converge on a grammar structured like the grammar for adults and actually almost identical to it. Therefore it makes sense to classify children’s utterances in terms of the adults’ grammar, so that we can compare children’s and adults’ utterances.

Children often make utterances that are not well-formed according to the grammar for adults. Often these utterances appear to reflect a different internal grammar or performance factors typical for children (such as a more limited memory, less developed pronunciation abilities, etc). Such utterances pose problems for the creation of the treebank. Automatically assigned syntactic structures are most likely incorrect, and require manual correction. It is not always a priori evident what these corrections should look like. Therefore explicit conventions and guidelines have to be developed on what syntactic structure should be assigned to such utterances. These conventions and guidelines must be set up in such a way that they are maximally useful for research into language acquisition by children. We discuss some of these annotation conventions and guidelines in section 3.4..

3.2. Cleaning

CHILDES utterances are enriched with all kinds of annotations. An extensive description of these annotations can be found in (MacWhinney, 2015a). Many of these annotations are in-line annotations. Some examples are given in (1):[7]

(1) Example in-line annotations in CHILDES CHAT files:

a. < ik wi > [//] ik wil xxx bekertje doen.
< I wan > [//] I want xxx cup-DIM do
‘I want to do the little cup’

b. < doe maar even > [/] doe maar even op tafel.
< put PRT PRT > [/] put PRT PRT on table
’Just put on the table’

c. knor knor [=! pig sound] , ik heb honger.
oink oink [=! pig sound] , I have hunger
‘Oink oink, I am hungry’

These examples illustrate annotations for retracing ([//])and repetition ([/]), both with scope over the preceding part be-

[4] http://portal.clarin.nl/node/4182.

[5] The Van Kampen subcorpus contains 109k utterances with 526k tokens distributed over two target children, with monthly recordings for these children from their first through their fifth year.

[6] Gertjan van Noord, personal communication.

[7] The sources are indicated by the session name (e.g. Sarah35) followed by the utterance number (e.g. 224), starting counting at 1. Utterances Sarah35.015, Sarah35.023 and Sarah35.224 from the *Van Kampen* corpus.

tween angled brackets, for unintelligible material (xxx) and for paralinguistic material ([=! ...]).[8]

The Alpino parser cannot deal with these annotations. A cleaning programme has been developed to remove the annotations and send a cleaned utterance to the Alpino parser.[9]

The cleaned variants of the utterances in (1) are:

(2) Example cleaned utterances:

a. ik wil xxx bekertje doen.

b. doe maar even op tafel.

c. knor knor , ik heb honger.

It is not always obvious how this cleaning should be done, and we have experimented with several variants, for example in an earlier variant we removed the xxx markings. However, this often led to clearly undesirable parses. Alpino analyses xxx as an unknown word, and assigns it a part of speech depending on the context (most often: noun), which is often correct or at least plausible. Even when it is wrong, the overall parse is generally easier to correct when xxx is present than when it is absent.

The cleaning program is available on GitHub[10] and has been integrated in the GrETEL upgrade described in section 4..

3.3. Editing syntactic structures

We use TrEd 2.0[11] as the editor for inspecting and correcting the syntactic structures (Pajas and Štěpánek, 2004). There already existed a TrEd extension[12] for syntactic structures in the format generated by Alpino, but it was only compatible with an older version of TrEd. We updated it to work with the current TrEd version, and created an 'extension repository'[13] to allow it to be installed from the TrEd 2.0 plug-in API. The editor provides an intuitively appealing graphical interface for inspecting and manipulating the syntactic structures, and is a desktop application, which works on multiple platforms (Linux, Windows, MacOS and several UNIX-based systems). Our experience is that most web-based interfaces are inferior to desktop interfaces, because of the requirement to be on-line, often unpredictable scrolling and cursor behavior and more limited options for keyboard short cuts. Therefore we did not employ a web-based tree editor. In addition, earlier projects that created treebanks for the Dutch language (in particular, the LASSY project (van Noord et al., 2013)) used TrEd so we could benefit from the experience with working with this tool gained there. Furthermore, the main platform our annotators work on is Windows, which restricts the options.

At the time we had to select a treebank editor (early 2015), there were not many alternatives to TrEd. See the Exmeralda Linguistic Annotation Wiki for an overview of annotation tools and related matters.

MPI's Synpathy could not be installed, and there is no support or further development. The EXMARaLDA Sextant `http://exmaralda.org/en/sextant-en/` tool (Wörner, 2009) was not available in a stable version.[14] *WebAnno* Version 2 (Yimam et al., 2014) is an annotation tool that is web-based and this version appears not particularly suited for annotating syntactic structures, though its successor version 3 (Eckart de Castilho et al., 2016) might be worth investigating further. *Atomic* is a new annotation editor (a desktop application) but claimed by the developers not to be stable yet (Druskat et al., 2014). The @nnotate tool appears to offer the required functionality but does not work on the Windows platform. Arborator deals with dependency relations in CONNL format only and just visualises edits in a textual CONLL file. The FLAT tool (van Gompel et al., 2017) was not ready for annotation of syntactic structures in 2015 and is a web application.

3.4. Annotation Conventions

Utterances from spoken language can contain many performance phenomena and errors for which it is not obvious how they should be analysed syntactically. The utterances used by the children contain many phenomena that are not part of the adult language. In addition, as in any annotated corpus, many phenomena can be analysed in multiple ways, none of which can be considered better than any other on purely linguistic grounds. It is important to analyse each construction in a consistent and uniform manner, so that it can be easily automatically identified and distinguished from other constructions in a treebank query application when the data are used in research. For this reason, it is important to develop and adhere to annotation conventions and guidelines.

For utterances made by adults we adhere to the annotation guidelines developed in the LASSY (van Noord et al., 2011) and Spoken Dutch Corpus Projects (Hoekstra et al., 2003), wherever applicable. For phenomena not covered there, we developed new annotation conventions. We will illustrate these with some examples.

As is well known, spoken language often contains incoherently structured utterances, with rephrasings, unfinished sentences, or just mispronunciations, all of which prove difficult for the Alpino parser to handle. Note that these problems are not limited to child speech, but also frequently occur in adult spoken language, which is often produced on the fly. In example (3), such phenomena are illustrated:[15]

[8] The glosses and translations given for these examples are not included in the CHILDES databases.

[9] Of course, removing these annotations makes it impossible to do research on these phenomena in combination with the syntactic structures. Therefore, we will translate these annotations into metadata in a new version of the cleaning program. See section 7..

[10] `https://github.com/JanOdijk/chamd`.

[11] TrEd is an abbreviation for Tree Editor. See `https://ufal.mff.cuni.cz/tred/`.

[12] `https://bitbucket.org/alpino/alpino`.

[13] `http://languagelink.let.uu.nl/~alexis/tred/extensions/alpino/`.

[14] We quote from the website: 'The development of the tool and the respective components is a "work in progres". This means: it is not guaranteed that the functions displayed in the software, or described elsewhere, will work. Some parts of this software and the schemas and models that underlie them, may change anytime. Therefore, no guarantee regarding the integrity of the data that will be edited with this tool, can be given.'

[15] LAURA28.264

(3) eh , wangetjes eh eve een keer , eve
um , cheek.DIM.PL um briefly a time , briefly
olie op wangetjes.
oil on cheek.DIM.PL
'Um, cheeks, um, probably once, briefly oil on cheeks.'

Since the structure in this sentence is chaotic, featuring filled pauses (*eh* 'um') which interrupt the flow of words, Alpino is unable to properly analyse the utterance. The lack of a verb to head the sentence makes it impossible to discover any reasonable analysis. Alpino also cannot distinguish the false start from the core part. We analyse the first part (until *een keer*) as a false start, which is filled with a sequence of fragments. The second part holds all the meaning, even on its own. We analyse such utterances as consisting of two parts. The meaningful part becomes the NUCL, or the nucleus, whereas the false start becomes the SAT, or the satellite.

An example of utterances that are not part of the adult language involves determiners. In acquiring adult language it is imminent that children eventually learn that some nouns combine with the article *de* 'the-UTR'[16] and others with the article *het* 'the-NEUT'. However, in early stages of acquisition children often do not use the different determiners correctly. Example (4)[17] shows such an utterance:[18]

(4) in de bad , he
in the-UTR bath , PRT
'in bath, right?'

Though it is clear that in this example *de bad* should be analysed as a noun phrase despite the gender mismatch, Alpino cannot do this. Since many researchers will want to know about determiner use by children acquiring language, it is important to manually correct such cases.

The following examples appear to contain a finite verb form (*lees* and *kocht*, respectively) where a participle is expected:[19]

(5) a. ik heb niet lees
I have not read-PRES
'I have not read'

b. Ik heb bolletjes kocht
I have roll-DIM-PL buy-PAST
'I have bought little rolls'

It is not a priori clear how such examples should be analysed: the child might be producing forms that do not conform to the adult language due to syntactic reasons, morphological reasons or phonological reasons. Each of these causes would imply a different analysis, but only after an intensive investigation of each phenomenon can one decide among them . In constructing the treebank we do not take a stand as to how such examples should be analysed, but we do treat each of them in a uniform way, so that each can be easily and automatically identified by researchers using a treebank query application. The examples in (5) are analysed in the treebank as participial verbal complements (*vc/ppart*) that contain a finite verb.[20]

A lot of utterances consist of an infinitive or perfect participle and its complements. Such a construction is not part of the adult language as a main clause. As one might expect, Alpino analyses them incorrectly. It occurs both with an overt subject and without. Without an overt subject it can occur as an infinitival or participial complement to other verbs in the adult language.

(6) a. ikke pap eten (Laura09.527)
I-emph porridge eat-INF
'I eat porridge'

b. en die maken (Laura13.042)
and that make-INF
'and make that one'

c. die weggelopen (Sarah10.024)
that away-walk-VD
'that one (has) run away'

In the AnnCor corpus we analyse such examples uniformly as infinitival (or participial) main clauses.

A spoken language effect that is difficult to capture by Alpino involves contractions. An example of such a contraction can be found in example (7):[21]

(7) Nee, das een andere, van de uiln !
No, that's a different, from the owls !
'No, that is a different one, from the owls!'

This example actually illustrates multiple problems. In spoken language the two words *dat is* 'that is' are sometimes pronounced as a contracted form *das*. It is orthographically incorrect to write this contraction as *das*. Therefore, Alpino cannot deal with it. A second problem is that the word *das* is a correct word of Dutch, so Alpino tries to assign a structure to the sentence in which this word is analysed as a noun (meaning 'tie' or 'badger'). In addition, the transcription in the CHILDES corpus violates the CHILDES transcription rules. In accordance with these rules (MacWhinney, 2015a, 47), this example of contraction should have been transcribed as *da(t) (i)s*. With such a transcription and the cleaning program (see section 3.2.) there would have been no problem for Alpino. We correct this by splitting up the contraction, though for such cases it would have been preferable to adapt the transcription and have the corrected transcription reparsed by Alpino.

The AnnCor documentation (Otten et al., 2018) describes these and many other annotation conventions in detail.

[16] UTR = uter, i.e. non-neuter, and NEUT = neuter.

[17] LAURA28.201

[18] The fact that this utterance is not a full sentence is in itself not problematic for Alpino.

[19] Utterances Laura09.527 and Laura13.042 from the *Van Kampen* Corpus.

[20] Assigning a structure to such utterances differing from what Alpino assigns to them reduces the options for example-based search, which requires parsing by Alpino (see section 4.), so such examples will have to be searched by writing XPath queries or by adapting XPath queries generated by example-based querying.

[21] LAURA70.71

3.5. Checks on the Annotation

In general, about 10% of each batch of manually corrected sentences were checked by a second annotator. At the beginning of the project, and when a new annotator started, all output was double-checked.

We carried out an initial small experiment to determine interannotator agreement for an independently annotated sample, by comparing all combinations of the annotations of two annotators, resulting in an average F-score of 0.86 (6 annotator pairs, random sample of 100 utterances). We used F-score, since it is not clear how a metric such as Cohen's κ or related scores can be applied (since there is not really a fixed number of categories for classification). Though the F-score of .86 is reassuring, a test with a larger sample is desirable and planned.

The annotators have excellent knowledge of the language and its syntactic structures and this enables them to make such corrections. However, they are human and therefore will very likely make errors due to lack of attention, oversights, etc. This is especially so because the annotations can be very complex and a lot of small details have to be attended to. We try to avoid such errors as much as possible in a number of ways. First, the TrEd tool avoids potential errors by providing fixed dropdown lists for fixed ranges of values (part of speech codes, morpho-syntactic features, etc.). Second, we developed a new tool, called the AnnCor Check Engine (ACE), which checks properties of the syntactic structures. We provide some examples of such checks:

- certain grammatical relations can occur only once in a local tree (i.e., a parent node and its children nodes). For example, the grammatical relation *su* (for subject) can occur only once. If an annotator accidentally violated this constraint while editing a tree, ACE issues an error message.

- In earlier Dutch treebank creation projects (Spoken Dutch Corpus (Oostdijk et al., 2002) and LASSY (van Noord et al., 2013)), it was decided that syntactic structures should not contain unary branching nodes, and the Alpino output follows this policy. Thus in a sentence such as *he swims* the pronoun *he* is not dominated by an NP node (which would lead to unary branching) but is immediately dominated by the clausal node (and it bears the grammatical relation *su* for subject). ACE warns against violations against this ban on unary branching.

- The syntactic structures contain some redundancies. For example, past participle and infinitival complements have different labels: *ppart* and *inf* respectively. But a verb heading such a phrase has morphosyntactic features specifying whether it is a participle (*wvorm=vd*) or an infinitive (*wvorm=inf*). A phrase labeled *ppart* should contain a head verb that is a participle, and a phrase labeled *inf* should contain a head verb that is an infinitive. ACE warns against violations of this rule.

- We collected statistics on local trees from the Spoken Dutch Corpus and LASSY treebanks in a database. We assigned a score to each local tree, in principle equal to its frequency. If a syntactic structure contains a local tree configuration with a score equal to 0, ACE issues an error message. If the local tree configuration has a score below a certain threshold,[22] ACE issues a warning. Of course, results obtained in the past do not give guarantees for the future. For this reason, we set up a tuning phase in which we manually added legal configurations that happened not to occur in the earlier treebanks but did occur in the AnnCor CHILDES treebank. We also adapted the scores for well-formed configurations that occurred in the earlier treebanks with a frequency below the set threshold.

The tool currently checks for 34 different potential errors and its local tree configuration database contains several thousand legal local configurations. The tool may occasionally give incorrect error or warning messages. In such a case the annotators can mark this error or warning message as incorrect for this instance. This ensures that the message will not be issued for this instance again when the syntactic structure is checked in a later stage or by a different annotator. The functionality to mark certain phenomena as exceptions is surely needed when a syntactic structure has to be assigned to an utterance that deviates from what is allowed in adult Dutch, such as in the examples in (5), in which a finite verb heads an infinitival or participial complement.

4. Treebank Querying

In addition to creating the treebank, the Anncor project developed an application that can be used to explore the corpus.

We extended the existing treebank query application *GrETEL* (Version 3) developed in Leuven (Augustinus et al., 2012),[23] adding the possibility to upload one's own corpora and associated metadata, as well as functionality to analyse and filter on data and metadata in *GrETEL* Version 4.[24]

GrETEL is a web application that allows researchers to search in Dutch treebanks and to perform a limited analysis of the search results. It has a very user-friendly example-based interface, but also allows queries in the XML query language XPath.

The corpus upload functionality allows users to upload an archived collection of plain-text files. The software will tokenise and parse these files using the Alpino parser, and import them into the XML database BaseX (Grün, 2010) for querying with GrETEL. Users can specify their corpus as private (only searchable for them) or publicly available. Next to plain text input, input in the CHAT format is possible as well. In this case, the software uses, inter alia, the cleaning algorithm described in section 3.2.. One can also upload a treebank, i.e. a corpus in which each utterances has been assigned a syntactic structure that is compatible with the Alpino output format. Such syntactic structures can have been generated fully automatically, or be the result of manual annotation. Work is currently ongoing to

[22] set, after some experimentation, to 10.

[23] http://gretel.ccl.kuleuven.be/gretel3/index.php.

[24] http://gretel.hum.uu.nl/gretel4/.

provide a wider range of input formats (in particular, FoLiA (van Gompel and Reynaert, 2013) and TEI).[25]

For adding metadata to corpora, we use a format defined during the development of PaQu, which allows users to add metadata in the running text (see `http://zardoz.service.rug.nl:8067/info.html#cormeta` for details). The software reads in the metadata and will create faceted search in GrETEL to allow users to both analyse and filter their search results. Users can change the facets to their liking, e.g. to use a range filter instead of checkboxes for numeric metadata. After finding a result set of interest, this set can be further analysed in an analysis interface. This interface enables the creation of pivot tables and graphs, which allows rapid insight into the data. The result set can also be exported to a tab-separated value text format to allow further analysis in other tools.

We will illustrate GrETEL 4 by the example that we also used in (Odijk et al., 2018), but now applied to the CHILDES subcorpus *Van Kampen Sarah*. We are interested in utterances by young children containing three bare[26] verbs. GrETEL offers query by example (QBE) functionality, in which a sentence containing the desired construction is entered, parsed and used as the basis for a query. We use this function with the example from (8), in which the three bare verbs are in boldface:

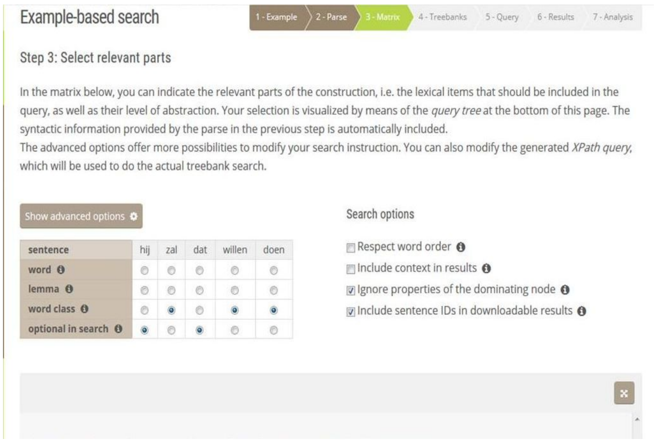
(8) hij **zal** dat **willen doen**
he will that want do
'he will want to do that'

This example sentence is now parsed by Alpino, and it parse result is shown. In order to have this example turned into a query we have to specify that the subject *hij* is not relevant for the construction we are interested in (the example contains it because Dutch sentences of this type must have a subject). Neither is the direct object *dat* (the example contains it because the verb *doen* is a transitive verb). The verbs must of course be included, but not these specific verbs: any verb that can occur in this construction will do. We therefore only require that they are verbs. The example is a main clause, but we also want to find examples in subordinate clauses. Therefore we specify that the properties of the top node of the parse tree containing the selected elements (*smain*, i.e. main clause) should be ignored. GrETEL offers a graphical interface to make such selections, which is illustrated in Figure 1.

This selection results in the XPath query (9):

(9)
```
//node[@cat and
       node[@rel="hd" and @pt="ww"] and
       node[@cat="inf" and @rel="vc" and
            node[@pt="ww" and @rel="hd"] and
            node[@rel="vc" and @cat="inf" and
                 node[@rel="hd" and @pt="ww"]]]]
```

which can be represented graphically as the query tree (10):



Example-based search

Step 3: Select relevant parts

In the matrix below, you can indicate the relevant parts of the construction, i.e. the lexical items that should be included in the query, as well as their level of abstraction. Your selection is visualized by means of the *query tree* at the bottom of this page. The syntactic information provided by the parse in the previous step is automatically included.
The advanced options offer more possibilities to modify your search instruction. You can also modify the generated *XPath query*, which will be used to do the actual treebank search.

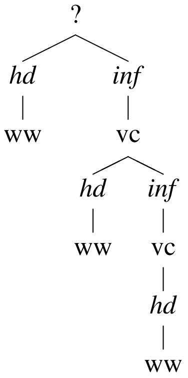
sentence	hij	zal	dat	willen	doen
word	○	○	○	○	○
lemma	○	○	○	○	○
word class	○	●	○	●	●
optional in search	●	○	●	○	○

Search options

☐ Respect word order
☐ Include context in results
☑ Ignore properties of the dominating node
☑ Include sentence IDs in downloadable results

Figure 1: Selection of the construction elements.

(10)



We run the query on the *Van Kampen Sarah* subcorpus[27] (*vksarah* in Gretel) and get 147 hits. We can now analyse the search results in terms of data (the elements that match nodes in the query tree) and metadata such as speaker, age, role, etc.). We illustrate the analysis page in Figure 2, which specifies the frequency of the most superordinate verbs used by Sarah. The application allows many aspects of the results to be counted and tabulated. For example, 18 utterances have Sarah (code SAR) as speaker, 129 are by the mother Jacqueline (code JAC). The child uses the construction already at the age of 28 months, but it occurs only sporadically until month 46, after which its frequency increases. 14 of the utterances by the child contain triples of verbs that are not in the input provided by the mother (in this, admittedly small, sample), and of the ones that do occur in the mother's input only one belongs to the triples frequently used by the mother (*zullen gaan doen* 'will go do'). These findings are consistent with the findings for the CHILDES subcorpus *Van Kampen Laura* reported in (Odijk et al., 2018), though Sarah starts using the construction earlier than Laura.

We refer to (Odijk et al., 2018) for many more details on the analysis options offered by GrETEL 4.

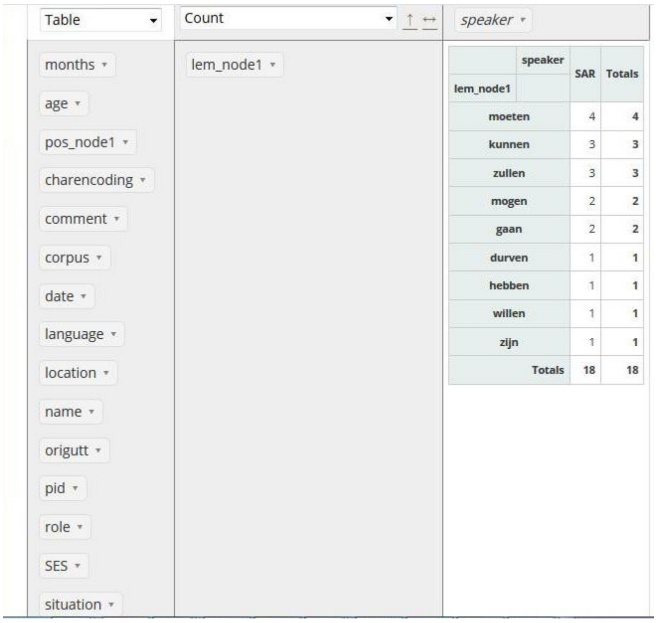
5. Updated CHAT files

We will make the parsed data available as downloadable files and as part of the GrETEL application. We will also provide the treebank in the CHAT format, in the MOR and

[25] `http://www.tei-c.org/`.

[26] i.e without *te*, cf. English *to*

[27] This corpus contains 44,869 utterances



speaker lem_node1	SAR	Totals
moeten	4	4
kunnen	3	3
zullen	3	3
mogen	2	2
gaan	2	2
durven	1	1
hebben	1	1
willen	1	1
zijn	1	1
Totals	18	18

Figure 2: Screenshot of the analysis page: frequency of the superordinate verb used by Sarah.

GRA tiers (MacWhinney, 2017). The MOR tier contains lemma, morphological and morpho-syntactic information for each word occurrence, and the GRA tier contains syntactic dependency relations between word occurrences. The representation in these tiers will be integrated in upgrades of the Dutch CHILDES corpora, so that these data can also be analysed with standard CHILDES tools such as *CLAN* (MacWhinney, 2015b) or CHILDES-recommended query and visualisation tools such as ANNIS (Krause and Zeldes, 2016).

To that end, we have created a tool to convert D-COI style tags as used in Alpino into MOR-style tags, a matter which is not completely trivial because the concepts behind the two different tagging methods differ radically: basically, D-COI tags represent a morpho-syntactic characterisation that abstracts from their concrete realisation with morphs, while MOR-style tags represent abstract characterisations of sequences of allomorphs. We plan to report on this in a separate paper (Odijk et al., in preparation).

6. Related Work

To our knowledge, (Sagae et al., 2001) was the first to parse utterances in CHILDES corpora, for English. They parsed the child-directed (adult) utterances only. These already form a challenge because they are transcriptions of ‘casual and conversational’ speech, ‘differing significantly from written natural language’.

(Sagae et al., 2007) proposed an annotation scheme for representing syntactic information as grammatical relations in CHILDES data largely based on (Sagae et al., 2004), a manually curated gold-standard corpus of 65,000 words annotated according to this scheme, and a parser (called *MEGRASP*) that was trained on the annotated corpus and produces highly accurate grammatical relations for both child and adult utterances, for English . We have not developed a parser specific to the CHILDES corpus but started from an existing parser developed for adult language, but the treebank resulting from our project can of course be used to train a child language parser.

(Pearl and Sprouse, 2013a) and (Pearl and Sprouse, 2013b) describe the creation of the CHILDES Treebank for the child-directed speech in various English CHILDES subcorpora in order to investigate the types of learning biases that are necessary to learn these constraints from the input, with the goal of determining whether any innate domain-specific biases are necessary.

(Laakso, 2005) reports on attempts to parse English CHILDES corpora automatically with a variety of rule-based and statistical parsers, showing that each of them has poor performance though the statistical parsers were slightly more successful.

(Gretz et al., 2015) describes a novel annotation scheme of dependency relations reflecting constructions of child and child-directed Hebrew utterances. A subset of the corpus was annotated with dependency relations according to this scheme, and was used to train two parsers (MaltParser and MEGRASP) with which the rest of the data were parsed.

(Dredze et al., 2007) describe results of their research on adaptation in the 2007 CoNLL Shared Task on Domain Adaptation, which involved, inter alia, CHILDES data. Their error analysis for this task suggests that a primary source of error is differences in annotation guidelines between treebanks, which clearly indicates that consistency of annotation is crucial for the usefulness of treebanks, both for humans as research material and for machine learning based software.

7. Conclusions and Future Work

We have described the approach to the development of the AnnCor CHILDES Treebank for Dutch. The treebank is still under development (we aim to make it available through the CLARIN research infrastructure in October 2018), but some results are already available, e.g. the extensions in Version 4 of the GrETEL query application, and the automatically generated parses have been made available in the PaQu application by our Groningen colleagues.

There are a number of aspects that we would like to work on in the future: (1) create options to adapt the transcription and process the adapted transcription (see section 3.4.); (2) changing a wrong syntactic parse into a correct one can be quite difficult if the wrong parse differs significantly from the correct one. The Alpino parses allows directives in the input string to guide the parsing process (so-called ‘bracketed input’).[28] Though we occasionally already use this feature, we would like to make it an integrated feature of the Alpino extension to the TrEd editor; (3) The CHILDES annotations and information in tiers related to an utterance such as in example (1) are currently ignored by the cleaning program. However, one might convert them into a format that can be used by the GrETEL query engine to extend the query options for searching for metadata and information on these other tiers. This requires at least an extension of the metadata notation used by GrETEL (e.g., to specify the

[28] See `https://www.let.rug.nl/vannoord/alp/Alpino/AlpinoUserGuide.html`.

span for which the annotation holds, as in example (1)), and probably also extensions in the query and analysis components.

8. Acknowledgements

This work was financed by the Utrecht University internal AnnCor research infrastructure project and by the CLARIAH-CORE project funded by the Dutch National Science Foundation (NWO).

9. Bibliographical References

Augustinus, L., Vandeghinste, V., and Eynde, F. V. (2012). Example-based treebank querying. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Bouma, G., van Noord, G., and Malouf, R. (2001). Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.

Dredze, M., Blitzer, J., Talukdar, P. P., Ganchev, K., Graca, J., and Pereira, F. C. (2007). Frustratingly hard domain adaptation for dependency parsing. In *EMNLP-CoNLL*, pages 1051–1055.

Druskat, S., Bierkandt, L., Gast, V., Rzymski, C., and Zipser, F. (2014). Atomic: an open-source software platform for multi-level corpus annotation. In *Proceedings of the 12th edition of the KONVENS conference*, volume 1, Hildesheim, Germany.

Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the LT4DH workshop at COLING*, pages 76–84, Osaka, Japan, dec.

Gretz, S., Itai, A., Macwhinney, B., Nir, B., and Wintner, S. (2015). Parsing Hebrew CHILDES transcripts. *Lang. Resour. Eval.*, 49(1):107–145, March.

Grün, C. (2010). *Storing and querying large XML instances*. Ph.D. thesis, University of Konstanz, Konstanz, Germany.

Hoekstra, H., Moortgat, M., Renmans, B., Schouppe, M., Schuurman, I., and van der Wouden, T. (2003). CGN syntactische annotatie. CGN report, Utrecht University, Utrecht, the Netherlands.

Krause, T. and Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31. `http://dsh.oxfordjournals.org/content/31/1/118`.

Laakso, A. (2005). On parsing CHILDES. Submitted to Midwest Computational Linguistics Colloquium (MCLC) 4/10/2005.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.

MacWhinney, B. (2015a). Tools for analyzing talk, electronic edition, part 1: The CHAT transcription format. Technical report, Carnegie Mellon University, Pittsburg, PA, April27. `http://childes.psy.cmu.edu/manuals/CHAT.pdf`.

MacWhinney, B. (2015b). Tools for analyzing talk, electronic edition, part 2: The CLAN programs. Technical report, Carnegie Mellon University, Pittsburg, PA, February23. `http://childes.psy.cmu.edu/manuals/CLAN.pdf`.

MacWhinney, B. (2017). Tools for analyzing talk, electronic edition, part 3: Morphosyntactic analysis. Technical report, Carnegie Mellon University, Pittsburg, PA, March16. `http://childes.psy.cmu.edu/manuals/MOR.pdf`.

Jan Odijk et al., editors. (2017). *CLARIN in the Low Countries*. Ubiquity Press, London, UK. DOI: `http://dx.doi.org/10.5334/bbi`. License: CC-BY 4.0.

Odijk, J., van Noord, G., Kleiweg, P., and Tjong Kim Sang, E. (2017). The parse and query (PaQu) application. In Jan Odijk et al., editors, *CLARIN in the Low Countries*, chapter 23, pages 281–297. Ubiquity, London, UK. DOI: `http://dx.doi.org/10.5334/bbi.23`. License: CC-BY 4.0.

Odijk, J., van der Klis, M., and Spoel, S. (2018). Extensions to the GrETEL treebank query application. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 46–55, Prague, Czech Republic, January 23-24. `http://aclweb.org/anthology/W/W17/W17-7608.pdf`.

Odijk, J., van der Veen, R., and Otten, M. (in preparation). Mapping D-COI tags to MOR-style tags. AnnCor technical report, Utrecht University.

Odijk, J. (2015). Linguistic research with PaQu. *Computational Linguistics in the Netherlands Journal*, 5:3–14, December.

Odijk, J. (2016a). A Use case for Linguistic Research on Dutch with CLARIN. In Koenraad De Smedt, editor, *Selected Papers from the CLARIN Annual Conference 2015, October 14-16, 2015, Wroclaw, Poland*, number 123 in Linköping Electronic Conference Proceedings, pages 45–61, Linköping, Sweden. CLARIN, Linköping University Electronic Press. `http://www.ep.liu.se/ecp/article.asp?issue=123&article=004`, `http://dspace.library.uu.nl/handle/1874/339492`.

Odijk, J. (2016b). Linguistic research using CLARIN. *Lingua*, 178:1 – 4. Linguistic Research in the CLARIN Infrastructure,`http://dspace.library.uu.nl/handle/1874/339377`.

Oostdijk, N., Goedertier, W., Eynde, F. V., Boves, L., Martens, J., Moortgat, M., and Baayen, H. (2002). Experiences from the Spoken Dutch Corpus project. In M. González Rodriguez et al., editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 340–347. ELRA, Las Palmas.

Otten, M., van der Veen, R., van Engeland, J., Meertens, E., and Wijsman, S. (2018). Anncor documentation. Utrecht University AnnCor Project Document, January.

Pajas, P. and Štěpánek, J. (2004). Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Com-*

putational Linguistics, pages 673–680, Manchester, UK.

Pearl, L. and Sprouse, J. (2013a). Computational models of acquisition for islands. In J. Sprouse et al., editors, *Experimental Syntax and Islands Effects*, pages 109–113. Cambridge University Press, Cambridge, UK.

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

Sagae, K., Lavie, A., and MacWhinney, B. (2001). Parsing the CHILDES database: Methodology and lessons learned. In *Proceedings of the Seventh International Workshop on Parsing Technologies (IWPT-2001), 17-19 October 2001, Beijing, China*. Tsinghua University Press.

Sagae, K., Lavie, A., , and MacWhinney, B. (2004). Adding syntactic annotations to transcripts of parent-child dialogs. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1815–1818, Lisbon, Portugal. ELRA, ELRA.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., and Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, CACLA '07, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

van Gompel, M. and Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, 12/2013.

van Gompel, M., van der Sloot, K., Reynaert, M., and van den Bosch, A. (2017). FoLiA in practice: The infrastructure of a linguistic annotation format. In Jan Odijk et al., editors, *CLARIN in the Low Countries*, chapter 6, pages 71–81. Ubiquity, London, UK. DOI: `http://dx.doi.org/10.5334/bbi.6`. License: CC-BY 4.0.

van Noord, G., Schuurman, I., and Bouma, G. (2011). Lassy syntactische annotatie (revision 19455). Lassy report, RU Groningen, Groningen, September 7. `https://www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf`.

van Noord, G., Bouma, G., Van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Tjong Kim Sang, E., and Vandeghinste, V. (2013). Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns et al., editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.

Wörner, K. (2009). *Werkzeuge zur flachen und hierarchischen Annotation von Transkriptionen gesprochener Sprache*. Phd, University of Bielefeld, Bielefeld. urn:nbn:de:hbz:361-16696, `https://pub.uni-bielefeld.de/publication/2301935`.

Yimam, S. M., Biemann, C., Eckart de Castilho, R., and Gurevych, I. (2014). Automatic annotation suggestions and custom annotation layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 91–96, Baltimore, Maryland, June. Association for Computational Linguistics.