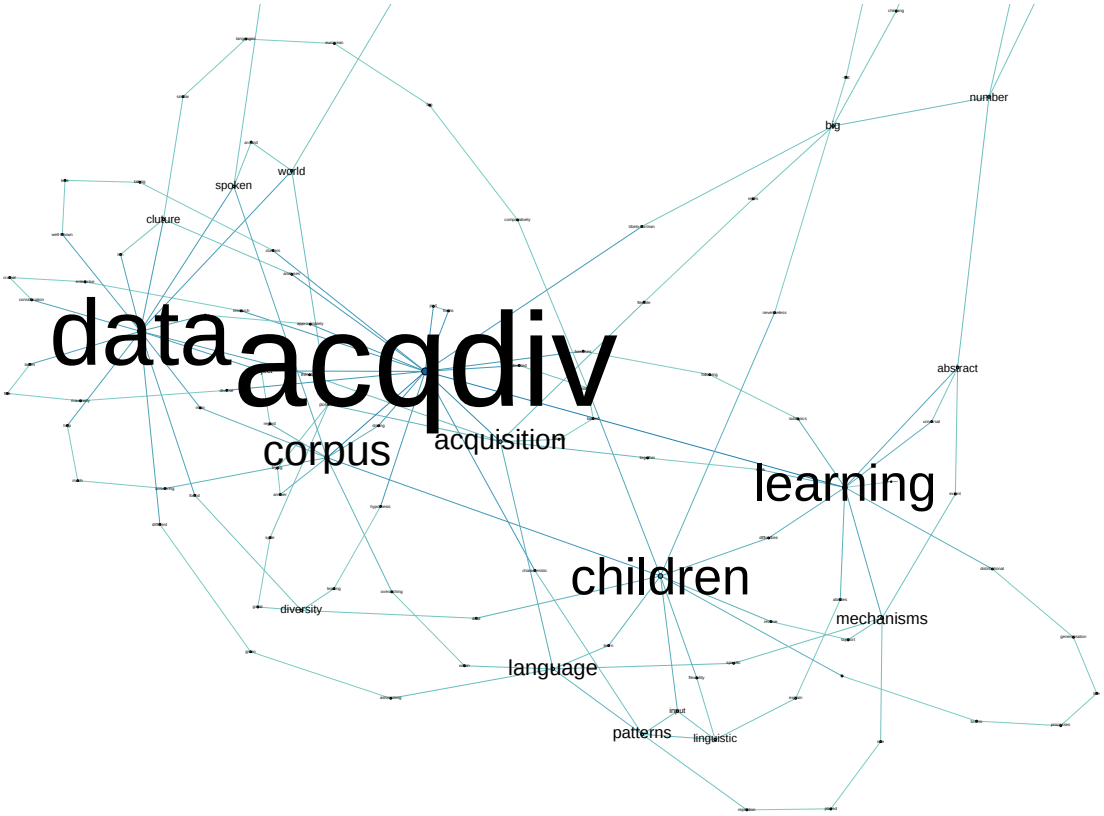

Manual for the ACQDIV Corpus

Robert Schikowski, Steven Moran and Sabine Stoll

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Purpose and structure of this document

This manual describes the corpora used and compiled in the ERC project "Acquisition processes in maximally diverse languages: min(d)ing the ambient language" (ACQDIV, grant no. 615988, 01/09/2014 - 31/08/2019, PI Sabine Stoll) – in short, the "ACQDIV Corpus". Also see http://www.acqdiv.uzh.ch and http://www.psycholinguistics.uzh.ch for the latest information.

The remainder of the manual is divided into five chapters. Chapter 2 gives an overview of the data contained in the ACQDIV Corpus, and Chapter 3 describes the format and content of the corpus in greater detail. Since the corpus is dynamically generated from several subcorpora, the following Chapter 4 describes the original data and how they are recast into the target structure. Readers with a technical interest may consult Chapter 5 to learn more about the individual steps involved in this procedure. Finally, Chapter 6 provides information for developers interested in extending the described architecture and methods to other resources.

Chapter 2 starts with a brief introduction to the ACQDIV languages, including examples for their diversity, shows an overview of the size of the subcorpora (given as the number of utterances, words, and morphemes in each), and summarizes differences between the subcorpora regarding the sampling of speakers and recording periods. This chapter also sketches the available annotation layers as well as notable data gaps in individual subcorpora.

Chapter 3 starts with the conditions of access and extension of the corpus in Section 3.1. It introduces the conceptual architecture of the corpus and the formats in which this is implemented. This is followed by detailed information on the tables and fields of the corpus database in Section 3.4 and lists of standardized values (e.g. for glosses and parts of speech) in Section 3.5.

Chapter 4 first gives an overview of the subcorpora's original corpus formats, i.e. CHAT, TalkBank XML, and Toolbox. It then deals with the subcorpora in alphabetical order: Chintang, Cree, Indonesian, Inuktitut, Japanese MiiPro, Japanese Miyata, Russian, Sesotho, Turkish, and Yucatec. Each section deals with the same recurring aspects: accessibility of the data, recording schemes, file systems and formats, and corpus formats. The subsection on corpus formats also describes how source structures are mapped to target structures in the ACQDIV Corpus.

Chapter 5 deals with the steps involved in building the ACQDIV Corpus from the subcorpora in roughly chronological order. The first two steps mainly apply to corpora which were initially not available in an accepted input format (TalkBank XML or Toolbox). These corpora required automatic and manual cleaning of files (including file systems, file names, and encodings) and corpus formats (typically broken CHAT). The following steps apply to all corpora: they are parsed and read into the dynamically generated database. The data are then postprocessed for the last finish.

Chapter 6 contains a very brief overview of the Python architecture behind the cleaning and the generation of the database and provides contacts for further information and access to the ACQDIV repository on GitHub.

1.2 Contributions

The ACQDIV Corpus is the result of one and a half years of collaborative work. The main contributors are:

- **Sabine Stoll** provided the idea, vision, and concept for the project
- **Robert Schikowski** devised the conceptual architecture of the corpus and supervised its realization
- **Steven Moran** designed and built the IT infrastructure for the database and was responsible for its implementation
- **Cazim Hysi** wrote the metadata parsers and refactored the data parsers
- **Danica Pajović** helped to clean the corpora and to write the parsers

We would also like to thank the following people:

- **Laura Canedo** helped to clean the Yucatec corpus
- **John Gamboa** helped to clean the Inuktitut corpus
- **Andreas Gerster** helped to clean the CHAT corpora and to test the data parsers and worked on gloss and POS unification
- **Irene Ma** helped with role unification
- **Jekaterina Mažara** created the graphics in Section 2.1 and provided expertise on Russian
- **Süleyman Sabri Taşçı** helped to clean the Turkish corpus
- **Melanie Trüssel** helped with gloss and POS unification

Further, this project would not have been possible without the data provided by our external collaborators:

- Shanley Allen for Inuktitut
- Julie Brittain for Cree
- Katherine Demuth for Sesotho
- Gaby Hermon for Indonesian
- Aylin Küntay for Turkish
- Barbara Pfeiler for Yucatec
- Yvan Rose for Cree
- Hannah Sarvasy for Nungon

Even more people were involved in the creation of the original corpora. See Chapter 4 (subsection "Publication, accessibility, documentation" in each corpus-specific section) for detailed information on corpus authors and citation.

Chapter 2

The dataset

2.1 The language sample

The ACQDIV Corpus is a longitudinal language acquisition corpus that currently features ten diverse languages. The languages and the eleven corpora by which they are represented are shown below.

Language	ISO 639-2	Corpora	Acronym
Cree	crl	Corpus of the Chisasibi Child Language Acquisition Study	CCLAS
Chintang	ctn	Chintang Language Corpus (Language Acquisition subcorpus)	CLC
Indonesian	ind	MPI-EVA Jakarta Child Language Database	JCLD
Inuktitut	iku	Allen Inuktitut Child Language Corpus	AIC
Japanese	jpn	MiiPro Japanese Corpus	MPJC
		Miyata Japanese Corpus	MYJC
Nungon	yuw	Sarvasy Nungon Corpus	SNC
Russian	rus	Stoll Russian Corpus	StRuC
Sesotho	sot	Demuth Sesotho Corpus	DSC
Turkish	tur	Koç University Longitudinal Language Development Database	KULLDD
Yucatec	yua	Pfeiler Yucatec Child Language Corpus	PYC

Table 2.1: ACQDIV languages and corpora

The initial set of languages was selected from five clusters calculated via maximum diversity sampling (Stoll & Bickel 2013) on the AUTOTYP database and from the World Atlas of Language Structures. This guarantees maximal diversity with respect to a number of central typological parameters:

- presence and nature of agreement and case marking
- word order
- degree of synthesis
- polyexponence and inflectional compactness of categories
- syncretism
- inflectional classes

Below some examples are given to illustrate the diversity of the ACQDIV languages with respect to these parameters.

Verbs in Japanese (1a) do not agree with any arguments, whereas Russian verbs (1b) agree with an S/A argument and Sesotho verbs (1c) agree with S or both A and P:

(1) a. *Okaa-san ga ue kara kore o otos-u.*
mother-HON NOM above ABL PROX ACC drop-NPST
'Mummy drops this from above.' (MPJC, tom20010518.u1806)

b. *Kak ty mam-u obnima-eš'?*
how 2SG.NOM mother-ACC embrace.IPFV-PRS.2SG.S/A
'How do you embrace mummy?' (StRuC, A00410909_594)

c. *Mme o-e-hlatsw-its-e.*
mother(I) NC.I.S/A-NC.IX.P-wash-PRF-IND
'Mother washed it.' (DSC, tiid.u143)

Sesotho (2a) does not have case marking for core arguments. By contrast, Inuktitut always marks at least one argument in a transitive scenario, be it the A as in (2b) or the P as in (2c).

(2) a. *Fusi a-s-a-di-kh-il-e di-perekisi.*
F. NC.I.S/A-still-NC.I.S/A-NC.X.P-pick-PRF-IND NC.X-peach
'Fusi has already picked the peaches.' (DSC, tviid.u207)

b. *Anaana-ngata aarqi-rataa-kainna-tanga.*
mother-POSS.3SG>3SG.ERG repair-RES-PST.RECENT-IND.3SG>3SG
'His mother has just fixed it.' (AIC, JUP92WM.u1427)

c. *Himmi-mi taku-lau-llu?*
dog-INS see-POL-IMP.1DL.S
'Shall we see the dog?' (AIC, SUP51WM.u733)

Another aspect in which the ACQDIV languages is synthesis. Indonesian (3a) is an example of a language with a fairly low degree of synthesis, whereas Cree (3b) belongs to one of the most genuinely polysynthetic languages of the world, featuring noun incorporation and polypartite stems:

(3) a. *O, Ei lagi minum susu.*
oh E. more drink milk
'Oh, Ei is drinking more milk.' (JCLD, HIZ-1999-05-20.0556)

b. *Chi-wâp-iht-â-n â*
2-light-by.head-TR.INAN.NON3-2SG>0 Q
kâ-pushch-ishk-iw-â-t.
PVB.CONJ-put.on-by.foot-STEM-TR.ANIM-3SG>4SG
'You see? She was putting it on.' (CCLAS, 19-A1-2006-08-16ms.u289)

Word orders differ radically between the ACQDIV languages. The most common word order, SVO, is e.g. found in Russian (4a). Another common word order, SOV, is found in Turkish (4b). Yucatec features (among other orders) the much less common VOS (4c).

(4) a. *Ja ne xoč-u salat!*
1SG.NOM NEG want.IPFV-NPST.1SG.S/A salad
'I don't want salad!' (StRuC, A05021006.68)

b. *Abla çay-ın-ı iç-sin.*
sister tea-POSS.3SG-ACC drink-OPT.3SG.S/A
'Let sister have her tea.' (KULLDD, irem32_02sep03_02-00-16.u1825)

c. *T-u-náach in-k'ab le Osita-o.*
PFV-3.A-bite POSS.1SG-hand DET O.-DIST
'That Osita bit my hand.' (PYC, SAN-1996-06-14.u181)

Russian has inflectional classes both in the nominal and verbal domains and often expresses a large number of categories by a single morpheme. The examples in (5a) and (5b) show the same bundle of grammatical functions (PL.GEN) expressed by very different morphs due to nominal inflection classes. By contrast, Chintang does not feature any inflectional classes, has less compact grammatical morphemes, and may even express a single function several times within a single word, as shown by the complex verb form in (5c).

(5) a. *Skol'ko produkt-ov papa nam privez?*
How.many product-PL.GEN dad.NOM 1PL.DAT bring.PFV.PST.M.SG.S/A
'How many products has dad brought us?' (StRuC, A06830304.1293)

b. *Im mnogo konfet-Ø togda ne da-eš'.*
3SG.DAT much sweet-PL.GEN then NEG give.IPFV-NPST.2SG.S/A
'Don't give him too many sweets then.' (StRuC, A06930318.523)

c. *Athom u-patt-a-ŋ-s-a-ŋ-ni-ŋ=kha.*
before 3A-call-PST-1sP-PRF-PST-1sP-3p=NMLZ
'They had called me before.' (CLC, CLDLCh2R02S01b.415)

The ACQDIV languages also feature very different kinds of syncretism. For instance, even though both Chintang and Inuktitut have an ergative that is used to mark agents in (6a) and (7a), the Chintang ergative also serves (among others) to mark causes (6b), whereas the Inuktitut ergative is also (again among others) used as a genitive (7b):

(6) a. *U-madum-ŋa=ta khur-u-gond-o-ko.*
POSS.3SG-aunt-ERG=FOC carry-3[s]P-around-3[s]P-IND.NPST[.3sA]
'His aunt carries her around.' (CLC, CLDLCh3R03S04.0496)

b. *Kok-ŋa=ta me?-no=kha=lo na.*
rice-ERG=FOC be.big-IND.NPST=NMLZ=SURP TOP
'He's so big because of the rice.' (CLC, CLDLCh2R04S04.438)

(7) a. *Ii, nuka-pi-ppit atu-ruma-mmauk.*
no younger_same_sex_sibling-DIM-POSS.2SG>3SG.ERG use-want-CAUS.3SG>3SG
'No, (it's because) your sister wants to use it.' (AIC, MAE14WM.u206)

b. *Ataata-ppit kami-alu-alu-ni sanarvat-ti-gia-lau-rit.*
father-POSS.2SG>3SG.ERG boot-big-big-INS put-CAUS-INCEP-POL-IMP.2SG.S
'Put your father's big, big boots somewhere.' (AIC, JUP51WM.0593)

2.2 Amount of data

The subcorpora of the ACQDIV Corpus vary considerably in size. Figure 2.1 shows how much utterances, words, and morphemes there are in each.

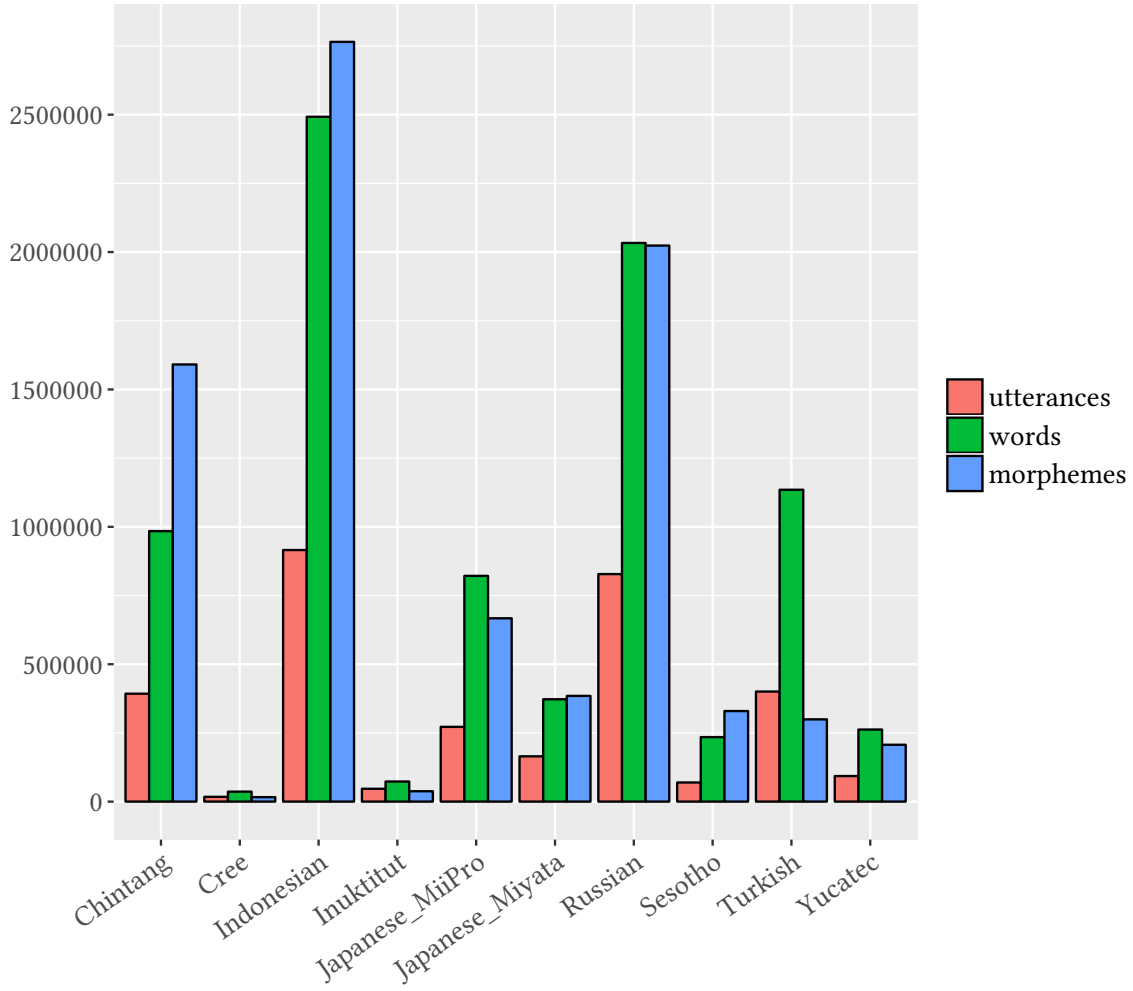

Figure 2.1: Amount of data in the ACQDIV subcorpora

2.3 Sampling for speakers and periods

The ACQDIV corpus focuses on the acquisition period from the beginning of the 2nd to the end of the 3rd year, and this is the period where the most linguistically diverse data are available. However, some subcorpora start at a much younger age (the lower boundary being some Chintang and Turkish children where recordings startet around half a year) and end considerably later (the extreme here is Indonesian, where the recordings for one child start at around 4;6 and end around 8;8).

The subcorpora also vary with regard to the number of target children that were recorded. The Cree subcorpus only features a single target child (and a single session for one other child), whereas the Indonesian and the Turkish corpus both feature eight target children.

The differences between the corpora are shown in summary fashion in Figure 2.2.

There is less variation in the intervals between recordings. In most corpora the recordings for one child took place every other week or once a month, and only two of the corpora have an even higher frequency rhythm with weekly recordings. The sessions vary in length both within and

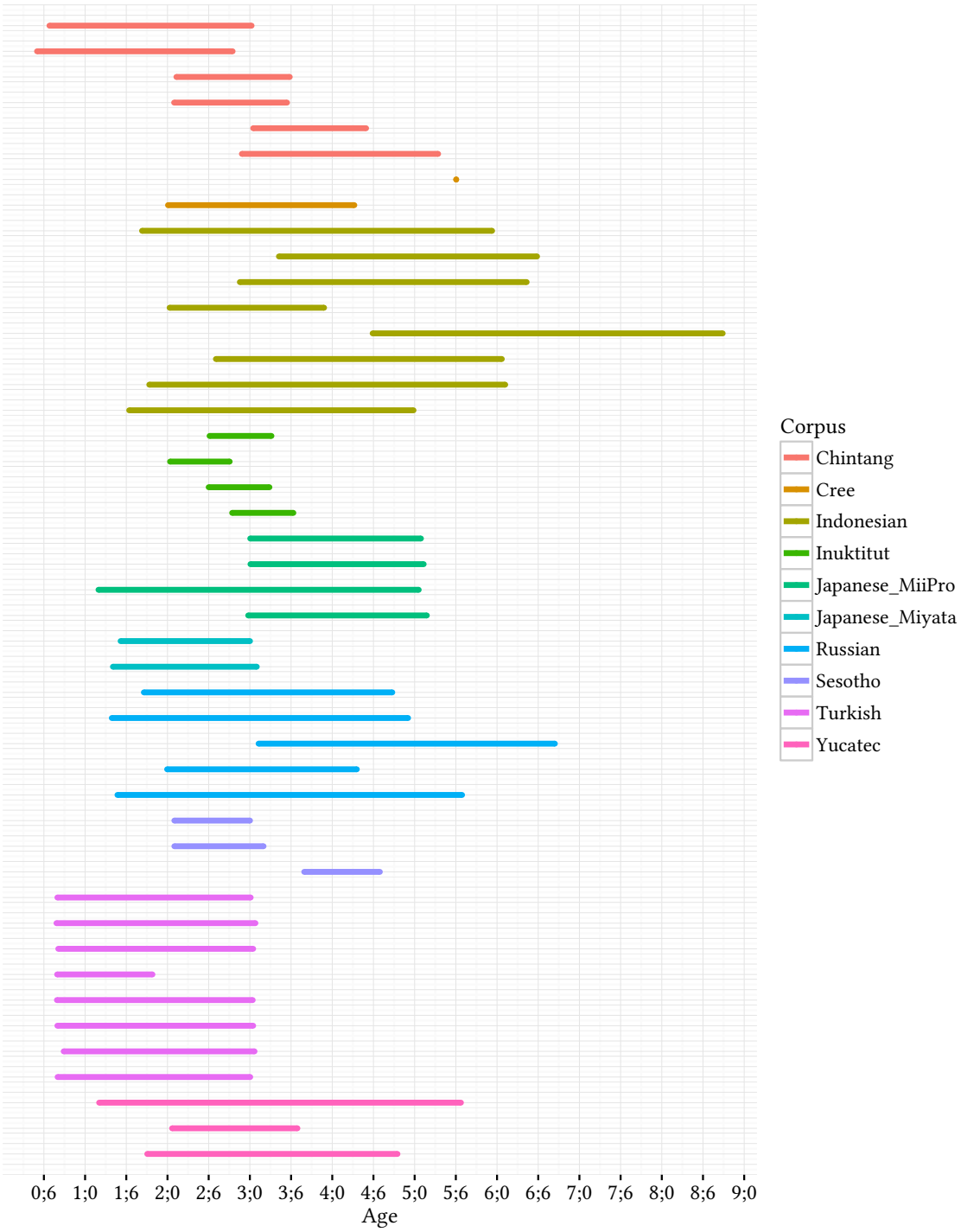

Figure 2.2: Children and recording periods in the ACQDIV Corpus

across corpora, ranging from half an hour to four hours.

More details on temporal sampling can be found in the corpus-specific sections of Chapter 4.

2.4 Annotation layers and data gaps

The ACQDIV Corpus is richly annotated. Each of the three principal levels – utterances, words, and morphemes – has dedicated additional annotations in addition to a transcription. The list below only shows a few frequently used and widely implemented types of annotations; for details see the section on the structure of the corpus.

- **utterances**: speaker, addressee, translation (usually into English), time stamps for start end end in associated media
- **words**: actual and target word, part of speech of the stem
- **morphemes**: gloss (original or unified across corpora), part of speech (original or unified)

These data are associated with metadata, the two principal levels here being sessions and speakers:

- **sessions**: recording date, media file
- **speakers**: label, name, age (as Y;M.D or in days), gender, role

Note that the only thing that all subcorpora have in common is that all sessions have been transcribed and that morphological analyses (including glosses) are at least available for some sessions or utterances. All other annotation layers mentioned above are widespread but not always available. The most important gaps can be summarized as follows:

- One corpus, Japanese Miyata, does not have systematic **transcriptions** for utterances by the mother, which present the overwhelming majority of non-target-child speech. This corpus is therefore not suitable for the study of child-surrounding speech.

- Both Japanese corpora and the Russian corpus have not been **translated** into any language. For Yucatec only Spanish translations are available.

- Almost half of the corpora do not specify addressees: this is the case for Cree, Indonesian, Sesotho, and Yucatec. Chintang features addressee coding only in a subset of the complete corpus.

- Turkish and Yucatec do not have any **time stamps**. The Russian corpus only has time stamps in a few sessions (2% of the Toolbox files which are incorporated into the ACQDIV Corpus; 14% in a parallel set of ELAN files which is currently not part of the ACQDIV Corpus). The Japanese Miyata corpus also has considerable gaps – the roughly 36% of linked files all stem from a single target child (which they cover completely). Indonesian and Inuktitut are comprehensively time-linked (with a few gaps in Inuktitut, around 87% of linked sessions) but only mark the beginning and not the end of utterances, so durations cannot be calculated. Only Chintang, Cree, Japanese MiiPro, and Sesotho have complete time stamps for both utterance boundaries.

- Some corpora contain considerable gaps with respect to **segmentation**, **glosses**, and **parts of speech**. For Cree, only the Ani subcorpus has been morphologically analyzed, and even there analyses are mainly available for the child's utterances. Likewise, Inuktitut completely lacks analyses for some sessions; moreover, many adult utterances in other sessions have not been analysed. The Turkish corpus has complete analyses for all participants in the sessions

of three children but almost nothing for the remaining five children. The corpus team is presently exploring the possibility of using an automatic parser. The situation is similar in Yucatec, although there are no plans for automatic analysis in this case. In Chintang, a small part of the data (about 80 sessions) have been analyzed automatically and thus have lower overall glossing quality. The majority of the Chintang sessions; however, have been analyzed manually.

- While all corpora have glosses, some are of limited use because they comply with **CHAT glossing conventions** where stems are only given in their phonological form (without a functional label) and affixes are only given as glosses (without specifying the phonological form). Thus, a word like German *Tage* is not segmented to *Tag -e* and then assigned two labels ("day -PL") but is glossed as "Tag -PL". This makes it difficult to infer the meaning of a word form from the glosses and makes it impossible to distinguish automatically between homophonous stems or affixes with the same label. Conventions of this kind are fully implemented in the two Japanese corpora and in Turkish. In Yucatec, the phonological form is given for all types of morphemes but there are still no functional labels for stems.

- The Russian corpus does not feature **segmentation**. Glosses cover all functional aspects of word forms but are concatenated into a single string. Accordingly, the `morphemes` table does not contain real morphemes but full word forms for Russian.

- Indonesian does not contain **part-of-speech tags**. Dummy tags are inserted during parsing to differentiate between stems and prefixes/suffixes, but more specific information is not available.

For more details on which layers are available for which corpus, also see the tables in the sections on the database tables `utterances`, `words`, and `morphemes`.

Chapter 3

The corpus

3.1 Getting access and adding data

The ACQDIV Corpus may be described as semi-open. Access may be gained by contributing data (for which see below) or by collaborating with the ACQDIV project. The detailed access regulations are described in the Terms of Use, which are available online at the ACQDIV website. The core points can be summarized as follows:

- The ACQDIV Corpus is a resource to be kept separate from the original data it builds on since it incorporates extensive efforts to clean, unify, and enrich the original data.
- The ACQDIV PI (Sabine Stoll, UZH) decides about access to and distribution of the data in the ACQDIV Corpus. On the other hand, the owners of the original data keep all their rights to these data.
- All resources used in a publication within the ACQDIV framework (including original data) must be properly cited.
- In addition, the developers of the ACQDIV Corpus as well as of any non-public corpora included therein must be asked if they want to become co-authors of publications in which these corpora are used. The contribution of each author (e.g. resource development vs. active contribution to research) must be specified.

The ACQDIV Corpus should be cited as Moran et al. (2016):

> Moran, Steven, Robert Schikowski, Danica Pajović, Cazim Hysi and Sabine Stoll. The ACQ- DIV Database: Min(d)ing the Ambient Language. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 4423–4429. May 23– 28, Portorož, Slovenia. Online: http://www.lrec-conf.org/proceedings/lrec2016/pdf/1198_Paper.pdf

The publication year should correspond to the cited release. New releases will be published after major changes at intervals yet to be fixed.

Currently the master version of the corpus is stored on the server of the Department of Comparative Linguistics at UZH, where the ACQDIV project is based. New corpora can be added at any time by request to the PI.

3.2 Conceptual architecture

Conceptually, the ACQDIV Corpus is a tree with five levels below the root:

- subcorpus
- session
- utterance
- word
- morpheme

A session is defined as a continuous stretch of time which contains spoken communication and whose boundaries are set by the applied recording scheme. Sessions may be instantiated by various types of files such as media, transcripts, or metadata files in the original subcorpora. While the original subcorpora consist of several sessions, where each in turn may or may not be instantiated by several files, all subcorpora and all their session-related data are contained in a single file in the ACQDIV Corpus.

Each level has one or several properties that can be searched for. To name a few examples, subcorpora have a language, sessions have recording dates, utterances may have a phonetic transcription, words may have an actual and a target form, and morphemes may have a gloss. These properties will henceforth be called tiers. Each tier is described in detail in Section 3.4 below.

In addition to the corpus tree, there are two metadata tables (one for session-level metadata, one for participant-level metadata). These tables are linked to the corpus via session IDs and participant codes, respectively.

3.3 Format

The abstract structure sketched above is currently implemented as an SQLite database. The database can be mapped to various output formats as required. Currently, the data are regularly exported as an R data object (R Core Team 2015), whose dataframes largely mirror the tables of the database.

There are many database GUIs that can be used to conveniently interact with the SQLite version. One that the ACQDIV team has made good experiences with and that is free to download is the DB Browser for SQLite, available from http://sqlitebrowser.org/. R is freely available from https://www.r-project.org/. Note that in either environment the corpus may take some time to load, depending on your system and computer. We recommend opening the database locally to save working memory.

The data sources for the subcorpora are encoded in diverse formats – see Chapter 4 for details.

Note that the original subcorpora also contain media files (audio and/or video, mostly digitized). The ACQDIV Corpus does not include these files to protect the children's privacy – sensitive information is much harder to remove or anonymize in media files than in text files. However, the names of the original media files are provided in the `sessions` metadata table.

3.4 Structure of the corpus

3.4.1 Overview and ERD

As a relational database, the ACQDIV Corpus is constituted by several tables and fields (also called columns below). The tables correspond roughly to the corpus levels described above:

- `sessions`: session-level metadata
- `speakers`: speaker-level metadata as given in individual sessions (i.e. one row = one speaker-session tuple)
- `uniquespeakers`: speaker-level metadata that can be specified independently of sessions
- `utterances`: utterances with their annotations, linkable to `sessions` and `speakers`
- `words`: words with their annotations, linkable to `utterances`
- `morphemes`: morphemes with their annotations, linkable to `utterances`

Each table has several fields, which correspond to what would be called a tier in a format more oriented towards running text. The names and detailed contents of the fields are described in the sections below. Two naming conventions are used across tables:

- Foreign keys have the suffix "_fk".
- The database often contains both the original data and a postprocessed version in separate columns. In such cases, the field containing the original data is marked by the suffix "_raw" (e.g. `gloss` vs. `gloss_raw`).

Joining information from several tables requires a key field with shared values. Currently, composite keys have to be used in all relevant cases, i.e. the link between tables has to be established by several fields as follows:

- The tables `utterances` and `sessions` are linked via the combination of the fields `corpus` and `session_id(_fk)`. The composite key is required because session IDs may not be unique across corpora.
- The tables `utterances` and `speakers` are linked via the combination of the fields `corpus`, `session_id(_fk)`, and `speaker_label`. The composite key is required because speaker labels may not be unique across corpora and because age and roles can change with the session.
- The tables `words` (or `morphemes`) and `utterances` are linked via the combination of the fields `corpus` and `utterance_id(_fk)`. The composite key is required because utterance IDs may not be unique across corpora.

Figure 3.1 shows an ERD of the database.[1]

[1] A few columns added later on in the project might be missing, but the basic picture and especially the relations between tables have remained the same.

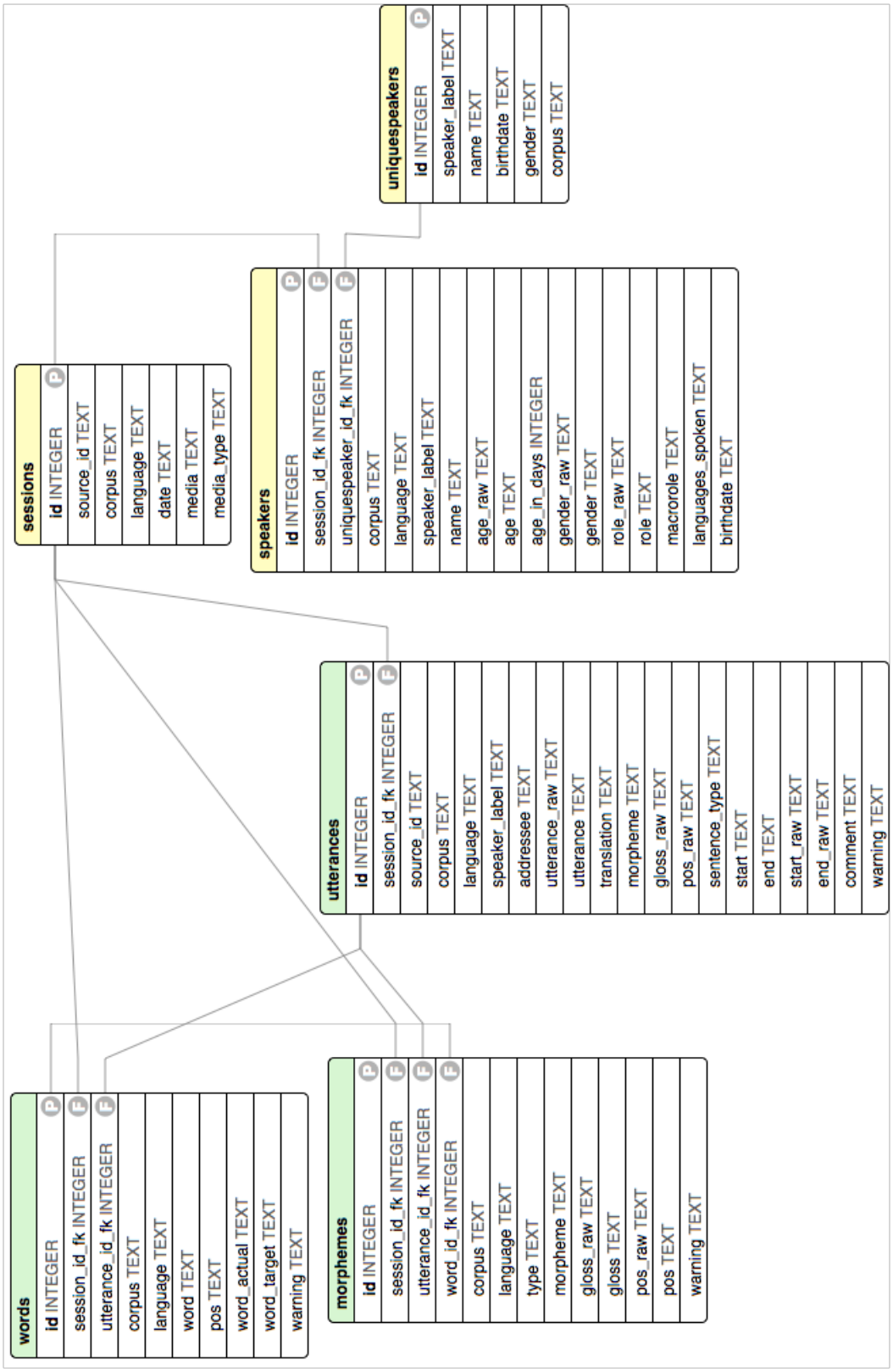

Figure 3.1: Entity-relationship diagram of the ACQDIV Corpus

3.4.2 Table sessions

Column	Content	Origin
id	an automatically generated numeric ID for the session	postprocessing
source_id	the name of the transcript file associated with the session. Note that source IDs are sometimes not unique across all corpora, so they are of limited use for identifying sessions.	data
corpus	the name of the corpus the session belongs to	data
language	the language of the corpus. While normally one language corresponds to one corpus, some languages may be represented by several corpora	postprocessing
date	the recording date for the session	data
target_child_fk	the unique ID of the target child of the session, linking to the table uniquespeakers	postprocessing

Table 3.1: Columns of the table session

3.4.3 Table speakers

Column	Content	Origin
id	an automatically generated numeric ID for the speaker-session combination	postprocessing
uniquespeaker_id_fk	the unique ID of the speaker independently of sessions, linking to the table uniquespeakers	postprocessing
session_id_fk	the ID of the session the speaker appeared in, linking to the table session	data
corpus	the name of the corpus the utterance belongs to	data
language	the language of the corpus	postprocessing
speaker_label	a code used to identify the speaker within the current corpus	data
name	the full name of the speaker	data
age	the age of the speaker at the time when the current session was recorded. The age may be given in years or (especially for children) in the formats Y;M.D or Y;M	postprocessing
age_in_days	the equivalent of the age in days	postprocessing
age_raw	the age as given in the original data. This may be formally slightly different from the standardized form given in age	data
gender	the gender of the speaker. The only allowed values are “female” and “male”.	postprocessing
gender_raw	the gender as given in the original data, sometimes slightly different from the processed form	data

Table 3.2: Columns of the table speakers

Column	Content	Origin
`role`	the role performed by the speaker in the present recording. Because of the diversity of the ACQDIV corpora, this concept covers both kinship terms (given in relation to the target child) and roles related to the setting (e.g. speaker, recorder, assistant). See below for a list of the possible values of this field.	postprocessing
`macrorole`	this column only allows four values: Target_Child, Child (any other children younger than or 12 years old), Adult (older than 12 years), Unknown. Its purpose is to make the most basic age- and role-related information available for all speakers, even when the precise age and/or role are not known.	postprocessing
`role_raw`	the role as given in the original data. This is often slightly different from the standardized form in `role` because of terminological differences ("target child" vs. "focus child" etc.)	data
`languages_spoken`	a space-separated list of all languages the speaker is able speak, given in the form of ISO 639-2 codes	data
`birthdate`	the birthdate of the speaker in the format YYYY-MM-DD	data

Table 3.2: Columns of the table `speakers`

3.4.4 Table `uniquespeakers`

The corpora themselves do not always make it clear which of the speaker labels they use are unique, so this table requires some additional explanation. In the CHAT-based corpora, different speakers with identical speaker labels occur regularly because (different) target children always have the code CHI and their mothers are always referred to as MOT. Thus, speaker labels alone are not sufficient for identifying unique speakers. The `uniquespeakers` table therefore uses unique combinations of speaker labels, full names, and birthdates (if available) to achieve this.

On the other hand, there is also the less frequent case of a single speaker being referred to by different labels (and/or names and birthdates) because of gaps or mistakes in the metadata. These case are currently ignored, i.e. these cases will appear as different speakers in the `uniquespeakers` table.

Column	Content	Origin
`id`	an automatically generated numeric ID for the speaker	postprocessing
`speaker_label`	a code used to identify the speaker within the associated corpus	data
`corpus`	the corpus the speaker appears in	data
`name`	the full name of the speaker	data
`birthdate`	the birthdate of the speaker in the format YYYY-MM-DD	data

Table 3.3: Columns of the table `uniquespeaker`

Column	Content	Origin
gender	the gender of the speaker	postprocessing

Table 3.3: Columns of the table uniquespeaker

3.4.5 Table utterances

Column	Content	Origin
id	an automatically generated numeric ID for the utterance	postprocessing
session_id_fk	the ID of the session the utterance belongs to, linking to the table session	data
source_id	the ID of the utterance in the original data	data
corpus	the name of the corpus the utterance belongs to	data
language	the language of the corpus	postprocessing
speaker_id_fk	the ID of the speaker who produced the utterance, linking to the table speakers	postprocessing
uniquespeaker_id_fk	the ID of the unique speaker who produced the utterance, linking to the table uniquespeakers	postprocessing
speaker_label	a code which uniquely identifies the speaker of an utterance within a corpus and presents a link to the tables speaker and uniquespeaker	data
addressee	a code for the participant adressed by the speaker of an utterance	data
utterance	an orthographic representation of the utterance (created by concatenating the single words if no separate representation is available; cleaned of punctuation marks)	postprocessing
utterance_raw	the original orthographic representation of an utterance (created by concatenating the single words if no separate representation is available	
translation	a free translation of the utterance (mostly English but Spanish for Yucatec)	data
sentence_type	broad sentence types, the most frequent values being default, question and exclamation. This may be taken directly from the data or inferred on the base of sentence delimiters.	data or postprocessing
childdirected	1 for utterances directed to a target child, 0 for all others (including unknown addressees)	data or postprocessing
start	the point in time in an associated media file where the utterance starts; format HH:MM:SS.	postprocessing
start_raw	like start but not unified to HH:MM:SS	data
end	the point in time in an associated media file where the utterance ends; format HH:MM:SS.	postprocessing
end_raw	like end but not unified to HH:MM:SS	data

Table 3.4: Columns of the table utterance

Column	Content	Origin
`morpheme`	all morphemes contained in an utterance, separated by spaces	postprocessing (concatenated morphemes)
`gloss_raw`	all glosses contained in an utterance, separated by spaces	postprocessing (concatenated morphemes)
`pos_raw`	all part-of-speech tags contained in an utterance, separated by spaces	postprocessing (concatenated morphemes)
`comment`	any comments on the utterance. This tier merges several tiers that are separated in some of the subcorpora but mostly overlap due to inconsistent usage: actions accompanying an utterance, background situation, ethnographic comments, comments on grammar, generic comments.	data
`warning`	warnings about formal problems on the utterance level	parsing

Table 3.4: Columns of the table `utterance`

The table below shows which of the utterance columns are regularly filled in which of the corpora.

tier	CLC (ctn)	CCLAS (crl)	JCLD (ind)	AIC (ike)	MPJC (jpn)	MYJC (jpn)	StRuC (rus)	DSC (sot)	KULLDD (tur)	PYC (yua)	SNC (yuw)
addressee	(+)	-	-	+	(+)	(+)	-	-	+	-	-
childdirected	(+)	-	-	+	+	+	-	-	+	-	-
comment	+	+	+	+	+	+	+	+	+	+	+
corpus	+	+	+	+	+	+	+	+	+	+	+
end	+	+	-	-	+	(+)	(+)	+	-	-	+
end_raw	+	+	-	-	+	(+)	(+)	+	-	-	+
id	+	+	+	+	+	+	+	+	+	+	+
language	+	+	+	+	+	+	+	+	+	+	+
sentence_type	-	+	-	+	+	+	-	+	+	+	+
speaker_label	+	+	+	+	+	+	+	+	+	+	+
session_id_fk	+	+	+	+	+	+	+	+	+	+	+
start	+	+	+	+	+	(+)	(+)	+	-	-	+
start_raw	+	+	+	+	+	(+)	(+)	+	-	-	+
translation	+	+	+	+	-	-	-	+	+	+	+
uniquespeaker_id_fk	+	+	+	+	+	+	+	+	+	+	+
utterance	+	+	+	+	+	+	+	+	+	+	+
utterance_id	+	+	+	+	+	+	+	+	+	+	+
utterance_raw	+	+	+	+	+	+	+	+	+	+	+
warning	+	+	+	+	+	+	+	+	+	+	+

Table 3.5: Presence of columns in the table `utterances`

3.4.6 Table `words`

Column	Content	Origin
`id`	an automatically generated numeric ID for the word	postprocessing
`utterance_id_fk`	the ID of the utterance the word belongs to, linking to the table `utterances`	data
`session_id_fk`	the ID of the session the word belongs to, linking to the table `sessions`	data
`corpus`	the name of the corpus the word belongs to	data
`language`	the language of the corpus	postprocessing
`word_language`	the language of the stem of a word; equals the corpus language by default	data
`word`	an orthographic representation of a word. When both actual and target forms are available (see Section 3.4.9), this is the actual word; otherwise it is the only available form. See `word_actual` and `word_target` for more precisely specified (but often empty) word forms.	data
`word_actual`	the word form the speaker actually produced; may be empty when only the target form is known	data
`word_target`	the word form the speaker intended to produce; may be empty when only the actual form is known	data
`pos`	the standardized part-of-speech tag of the stem of the word	postprocessing
`pos_ud`	the universal part-of-speech tag[2] of the stem of the word	postprocessing
`warning`	warnings about formal problems on the word level, e.g. missing or broken glosses	parsing

Table 3.6: Columns of the table `words`

The table below shows which of these columns are regularly filled in which of the corpora.

tier	CLC (ctn)	CCLAS (crl)	JCLD (ind)	AIC (ike)	MPJC (jpn)	MYJC (jpn)	StRuC (rus)	DSC (sot)	KULLDD (tur)	PYC (yua)	SNC (yuw)
corpus	+	+	+	+	+	+	+	+	+	+	+
id	+	+	+	+	+	+	+	+	+	+	+
language	+	+	+	+	+	+	+	+	+	+	+
session_id_fk	+	+	+	+	+	+	+	+	+	+	+
utterance_id_fk	+	+	+	+	+	+	+	+	+	+	+

Table 3.7: Presence of columns in the table `words`

[2] http://universaldependencies.org/u/pos/

tier	CLC (ctn)	CCLAS (crl)	JCLD (ind)	AIC (ike)	MPJC (jpn)	MYJC (jpn)	StRuC (rus)	DSC (sot)	KULLDD (tur)	PYC (yua)	SNC (yuw)
warning	+	+	+	+	+	+	+	+	+	+	+
word	+	+	+	+	+	+	+	+	+	+	+
word_actual	+	+	+	+	+	+	+	+	+	(+)	-
word_target	-	+	+	+	+	+	-	+	+	+	+
pos	+	+	(+)	+	+	+	+	+	+	+	+
pos_ud	+	+	(+)	+	+	+	+	+	+	+	+

Table 3.7: Presence of columns in the table `words`

3.4.7 Table `morphemes`

Column	Content	Origin
`id`	an automatically generated numeric ID for the morpheme	postprocessing
`word_id_fk`	the ID of the word the morpheme belongs to, linking to the table `words`	data
`utterance_id_fk`	the ID of the utterance the morpheme belongs to, linking to the table `utterances`	data
`session_id_fk`	the ID of the session the morpheme belongs to, linking to the table `sessions`	data
`corpus`	the name of the corpus the morpheme belongs to	data
`language`	the dominant language of the corpus	data
`morpheme_language`	the language of an individual morpheme; equals the corpus language by default	data
`type`	the morpheme type (actual vs. target, (see Section 3.4.9). Because most corpora only specify either the actual or the target morpheme most of the time (differently from the word level, where contrasting forms are often given), only this one form is taken over and the type is specified in this column.	data
`morpheme`	an orthographic representation of a morpheme (often in its underlying shape). Mostly this is the only form available, but in the rare case where both an actual and a target form are given only the actual form is taken over.	

Table 3.8: Columns of the table `morphemes`

Column	Content	Origin
`gloss`	a standardized label indicating the function of grammatical morphemes. The Leipzig Glossing Rules form the base for standardization and additional labels are drawn from a project-internal vocabulary given in Section 3.5.3. Morphemes whose original form cannot be assigned to a standard appear as `NULL` in this column. This also includes all lexical morphemes – there are too many different types in this partition to create a standardized vocabulary, and there are no simple automatizable rules for distinguishing them from grammatical morphemes.	data/postprocessing
`gloss_raw`	the original gloss (before standardization). Depending on the corpus, this column may contain glosses for both grammatical and lexical morphemes (differently from `gloss`, where only standardized grammatical labels appear).	data
`pos`	a part-of-speech tag. Parts of speech are also standardized. The project-internal set of tags is given in Section 3.5.4	data/postprocessing
`pos_raw`	the original part-of-speech tag (before standardization)	data
`warning`	warnings about formal problems on the morpheme level	parsing

Table 3.8: Columns of the table `morphemes`

The table below shows which of these columns are regularly filled in which of the corpora.

tier	CLC (ctn)	crl (CCLAS)	JCLD (ind)	AIC (ike)	MPJC (jpn)	MYJC (jpn)	StRuC (rus)	DSC (sot)	KULLDD (tur)	PYC (yua)	SNC (yuw)
corpus	+	+	+	+	+	+	+	+	+	+	+
language	+	+	+	+	+	+	+	+	+	+	+
gloss	+	+	+	+	(+)	(+)	+	+	(+)	(+)	+
gloss_raw	+	+	+	+	(+)	(+)	+	+	+	+	+
id	+	+	+	+	+	+	+	+	+	+	+
language	+	+	+	+	+	+	+	+	+	+	+
morpheme_language	+	+	-	-	+	+	+	-	+	-	+
morpheme	+	+	+	+	(+)	(+)	+	+	(+)	+	+
pos	+	+	(+)	+	+	+	+	+	+	+	+
pos_raw	+	+	-	+	+	+	+	+	+	+	+
session_id_fk	+	+	+	+	+	+	+	+	+	+	+
type	+	+	+	+	+	+	+	+	+	+	+

Table 3.9: Presence of columns in the table `morphemes`

tier	CLC (ctn)	crl (CCLAS)	JCLD (ind)	AIC (ike)	MPJC (jpn)	MYJC (jpn)	StRuC (rus)	DSC (sot)	KULLDD (tur)	PYC (yua)	SNC (yuw)
utterance_id_fk	+	+	+	+	+	+	+	+	+	+	+
warning	+	+	+	+	+	+	+	+	+	+	+

Table 3.9: Presence of columns in the table `morphemes`

3.4.8 Table `all_data`

This table only exists in the R object and brings together information from all tables in one big flat object. IDs and foreign keys on which the merger is performed, duplicated columns, and a few less often used columns are omitted. Some columns are renamed in order to make clear which table they originate from. The table below shows the correspondences between original columns and columns in `all_data`.

original table	old column	new column
sessions	id	session_id
sessions	source_id	session_id_source
sessions	corpus	corpus
sessions	language	language
sessions	date	date
speakers	id	speaker_id
speakers	uniquespeaker_id_fk	-
speakers	session_id	-
speakers	corpus	corpus
speakers	language	language
speakers	speaker_label	speaker_label
speakers	name	name
speakers	age	age
speakers	age_in_days	age_in_days
speakers	age_raw	age_raw
speakers	gender	gender
speakers	gender_raw	gender_raw
speakers	role	role
speakers	macrorole	macrorole
speakers	role_raw	role_raw
speakers	languages_spoken	-
speakers	birthdate	birthdate
uniquespeakers	id	-
uniquespeakers	speaker_label	-
uniquespeakers	corpus	-
uniquespeakers	name	-
uniquespeakers	birthdate	-
uniquespeakers	gender	-

Table 3.10: Columns in the merged table

original table	old column	new column
utterances	id	utterance_id
utterances	session_id_fk	-
utterances	source_id	utterance_id_source
utterances	corpus	corpus
utterances	language	language
utterances	speaker_label	speaker_label
utterances	addressee	addressee
utterances	utterance	utterance
utterances	utterance_raw	utterance_raw
utterances	translation	translation
utterances	sentence_type	sentence_type
utterances	start	start
utterances	start_raw	start_raw
utterances	end	end
utterances	end_raw	end_raw
utterances	morpheme	utterance_morphemes
utterances	gloss_raw	utterance_glosses_raw
utterances	pos_raw	utterance_poses_raw
utterances	comment	comment
utterances	warning	-
words	id	word_id
words	utterance_id_fk	-
words	session_id_fk	-
words	corpus	corpus
words	language	language
words	word	word
words	word_actual	word_actual
words	word_target	word_target
words	pos	pos_word_stem
words	warning	-
morphemes	id	morpheme_id
morphemes	word_id_fk	-
morphemes	utterance_id_fk	-
morphemes	session_id_fk	-
morphemes	corpus	corpus
morphemes	language	language
morphemes	type	morpheme_type
morphemes	morpheme	morpheme
morphemes	gloss	gloss
morphemes	gloss_raw	gloss_raw
morphemes	pos	pos
morphemes	pos_raw	pos_raw
morphemes	warning	-

Table 3.10: Columns in the merged table

3.4.9 Actual and target fields

All of the original subcorpora make a distinction between what a child actually said and what the adult target form would have been. Although none of the corpora carries this distinction through on all tiers of all levels, all of them incorporate it at least implicitly and many have separate tiers for the actual and target versions of at least overarching tiers. The table below shows for each corpus if the main tiers of each level always belong to one type ("a(ctual)" or "t(arget)"), if the types are distinguished using separate tiers ("a vs. t"), or if both types are mixed on a single tier without making a clear distinction ("a/t").

subcorpus	`words` `word`	`morphemes` `morpheme`	`morphemes` `gloss`	`morphemes` `pos`
CLC (Chintang)	a	t	t	t
CCLAS (Cree)	a vs. t	a vs. t	t	t
JCLD (Indonesian)	a vs. t	t	t	-
AIC (Inuktitut)	a vs. t	t/a	t	t
MPJC (Japanese)	a vs. t	t/a	t/a	t/a
MYJC (Japanese)	a vs. t	t	t	t
StRuC (Russian)	a	t	t	t
DSC (Sesotho)	a vs. t	t	t	t
KULLDD (Turkish)	a vs. t	t/a	t/a	t/a
PYC (Yucatec)	t/a	t	t	t
SNC (Nungon)	t	t	t	t

Table 3.11: Actual and target tiers in the original subcorpora

Since there is not a single corpus which consistently codes the actual/target distinction on all tiers of all levels and the overall emerging picture is rather chaotic, the following rules for simplification were applied:

- The distinction is most relevant and most frequently coded on the word level. Therefore, the `words` table of the ACQDIV Corpus features three columns: `word_actual`, `word_target`, and `word`. The latter is intended for easy searches regardless of the actual/target distinction. It contains the actual word form by default but may contain the target word form in the rare case that the actual word form is not available.

- On the morpheme level, the actual/target distinction is less relevant and less consistently coded. The `morphemes` table therefore only gives three default columns (`morpheme`, `gloss`, `pos`) and an additional column `type` that specifies if the values normally correspond to actual or to target forms. This representation glosses over inconsistencies (many of the subcorpora do not have a clear guideline for the distinction on the morpheme level so that both actual and target forms are found) and ignores any differences that might exist between the three fields.

- Finally, only two corpora (Cree and Indonesian) makes a distinction on the utterance level. This distinction is therefore completely ignored in the ACQDIV Corpus.

The current implementation for `word_actual` and `word_target` is encoded with a binary value in the *consistent_actual_target* field in the corpus-specific configuration files. The reasoning behind this decision is that if a corpus has a high congruence between actual and target in child speech, it is likely, that the transcribers/annotators did not make the distinction consistently. In other words, "inconsistently coded" means there are some cases where an actual/target distinction has been coded

but they are so few that it's very unlikely that the coding is consistent. There are two corpora that are currently tagged *consistent_actual_target==no*. These are Yucatec and Nungon. In Yucatec, the transcription comes on three lines (Pfeiler, p.c.):

1. the expression according to the norm
2. phon: what the child actually said
3. eng/esp: Translation into Spanish.

In our data processing, we did not include the CHAT *%(x)pho* tier, it being located on the utterance level (i.e. alignment to words is not guaranteed). All in all we had too few corpora with phonetic transcriptions on that level to introduce an column like *utterances.phonetic*. Perhaps in Yucatec, however, the alignment is even good enough to pull at least the Yucatec tier to the word level – evaluation pending.

Regarding Nungon, the transcriber aims to write down the actual word spoken by the child (Sarvasy, p.c.). If the spoken form by the child diverges phonetically or phonologically from the adult enough to be noticed by the transcriber, this should be documented on the first tier. The second tier then reflects the target form, as understood by the transcriber. Large differences are noted by the transcriber, but minor ones may be missed (e.g. the child doesn't articulate a word medial syllable carefully). Hence Nungon given our criteria for inconsistent encoding is marked as "no", but this is due to the very few data points that we have and this issue should be revisited in the future when more data is put into the pipeline.

Finally, there is largely no difference in *consistent_actual_target==yes* languages in the adult forms. That is, most transcriptions of words by adults are the same between actual spoken utterance and the target form. This seems a bit confusing from a database point-of-view, i.e. a lot of duplicated between two or three fields in the same table (word, word_actual, word_target). But the reasoning here is that each column can be taken separately depending on the analysis.

3.5 Conventions

3.5.1 Transcription conventions

The transcription conventions used in the ACQDIV Corpus have been greatly simplified compared to the original subcorpora, especially those that were initially coded as CHAT or TalkBank XML. This was necessary in order to ensure maximal comparability.

While the conventions for representing segmental material have not been touched, the following changes were applied with respect to additional symbols:

- All punctuation has been removed. The information contained in utterance delimiters (including CHAT's Special Utterance Terminators) was transferred to the newly introduced tier `utterances.sentence_type` (for instance, a utterance-final question mark now corresponds to the sentence type "question").
- All special CHAT codes such as postcodes, Satellite Markers, tone and prosody markers, quotation markers, Utterance Linkers, and overlap markers have been removed without replacement.
- Likewise, CHAT's Special Form Markers (codes starting with "@" attached to words) have been deleted.
- CHAT's Local Events have been transferred to the comment tier (concatenating them to any pre-existing material). Where the utterance only consists of an event, the sentence type has

been set to "action". Pauses, which are also classified as a special type of Local Event by the CHAT manual, have been removed without traces.

- All types of codes for untranscribed material have been replaced by `NULL/NA` in isolation and by "???" when embedded into a string. This includes the CHAT codes "xxx", "yyy", "www", so the the difference between unintelligible words, words with a clear phonetic shape but unclear phonology, and words not transcribed for other reasons is lost. Where more detailed comments are available on what could not be transcribed for which reason, they are transferred to the relevant `warnings` field.
- Morpheme separators (mainly given on the morphology tiers, but sometimes also elsewhere) have been deleted. The information contained in them has been transferred to the field `morphemes.pos`, where all prefixes and suffixes get the dummy tags "pfx" and "sfx", respectively.
- A few corpora have explicit coding for compounds. This has been simplified (see the description of the original data), leaving only "=" as the separator between the compound elements ("apple=tree").

This leaves "???" (untranscribed element within string) and "=" (compound separator) as the only metalinguistic elements on the object language tiers.

3.5.2 Roles and macroroles

The ACQDIV Corpus currently allows the following values in the `speakers.role` field. This list is the result of a simplification of the values found in the original data, which are diverse both because of terminological differences (e.g. "target child" vs. "focus child") and spelling mistakes ("Garndmother" vs. "Grandmother"). While some subcorpora distinguish between kinship terms ("mother", "son"), age groups ("child", "adult"), and other roles ("caretaker", "playmate") most of the corpora do not, so these categories also appear as one in the ACQDIV Corpus.

Adult	Family_Friend	Male	Target_Child
Aunt	Father	Mother	Teacher
Babysitter	Female	Neighbour	Teenager
Boy	Friend	Niece	Toy
Brother	Girl	Playmate	Twin_Brother
Caller	Grandfather	Research_Team	Uncle
Caretaker	Grandmother	Sister	Unknown
Child	Great-Grandmother	Speaker	Visitor
Cousin	Host	Student	
Daughter	Housekeeper	Subject	

The field `speakers.macrorole`, which is created during postprocessing, is the result of mapping these roles to the four values "Child", "Target_Child", "Adult", and "Unknown". Differently from the `role` field, macroroles also include inference based on age (younger than 12 years = "Child") and IDs (e.g. CHI = "Child"; other ID-based mappings depend on the individual corpora).

3.5.3 Grammatical glosses

The ACQDIV Corpus uses a standardized set of grammatical glosses in the column `morphemes.gloss`. The value used in the original data is given in `morphemes.gloss_raw`. The standardizet set consists of all glosses proposed in the Leipzig Glossing Rules plus additional values as needed (marked with an asterisk in the list below). Less frequent values were directly taken over from the original data in order to fill all rows but are not documented below.

*0	non-person
1	first person
2	second person
3	third person
4	fourth person (in switch reference or direct/inverse systems)
4SYL	tetrasyllabifier
A	agent-like argument of canonical transitive verb
ABIL	abilitative
ABL	ablative
ABS	absolutive
ACC	accusative
ACROSS	distal horizontal deixis
ACT	active
ADESS	adessive
ADJ	adjective
ADJZ	adjectivizer
ADN	adnominal
ADV	adverb(ial)
ADVZ	adverbializer
AFF	affirmative
AGT	agentive
AGR	agreement
ALL	allative
ALT	alternating tense
AMBUL	ambulative
ANIM	animate
ANTIP	antipassive
AOR	aorist
APPL	applicative
ART	article
ASP	unspecified aspect marker
ASS	assertive
ASSOC	associative
ATTN	attention
AUTOBEN	autobenefactive
AUX	auxiliary
AV	actor voice
BABBLE	babbling
BEN	benefactive
CAUS	causative
CHOS	change of state
CLF	classifier
CLIT	clitic with unspecified function
CM	compound marker
COLL	collective
COM	comitative
COMP	complementizer
COMPAR	comparative
COMPL	completive
CONC	concessive
COND	conditional
CONJ	conjunction
CONJ	conjugation marker
CON	conative
CONT	continuous
CONTEMP	contemporative mood
CONTING	contingent mood
CONTR	contrastive
COP	copula
CVB	converb
DAT	dative
DECL	declarative
DEF	definite
DEICT	deictics (other than demonstratives)
DEM	demonstrative
DEP	dependent (mood or other form)
DEPR	deprivative
DESID	desiderative
DESTR	destructive
DET	determiner
DETR	detransitivization
DIFF.SBJ	different subject
DIM	diminutive
DIR	directional case
DIR	direction (in direct/inverse systems)
DIST	distal
DISTR	distributive
DOWN	distal deixis pointing down
DU	dual
DUB	dubitative
DUR	durative
DYN	dynamic
ECHO	echo word
EMPH	emphatic
EQU	equative
ERG	ergative
EVID	evidential
EXCL	exclusive
EXCLA	exclamation
EXIST	existential copula
EXT	extensional
F	feminine
FILLER	filler
FOC	focus

FUT	future
GEN	genitive
HAB	habitual
HES	hesitative
HHON	high honorific
HON	honorific
HORT	hortative
IDEOPH	ideophone
IMIT	imitative
IMNT	imminent
IMP	imperative
IMPERS	impersonal
INAL	inalienable possession
INAN	inanimate
INCEP	inceptive
INCH	inchoative
INCL	inclusive
INCOMPL	incompletive
IND	indicative
INDF	indefinite
INDIR	indirect
INF	infinitve
INS	instrumental
INSIST	insistive
INSIST	intensifier
INTJ	interjection
INTR	intransitive
INTRG	interrogative
INV	inverse
IPFV	imperfective
IRR	irrealis
LNK	linker
LOC	locative
M	masculine
MED	medial (deixis)
MHON	mid honorific
MIR	mirative
MOD	modal
MOOD	unspecified mood marker
MV	middle voice
N	neuter
N	noun
N	non- (e.g. NSG, NPST…)
NAG	nomen agentis
NAME	person's name
NC	noun classes, e.g. NC.I, NC.II, NC.III…
NEG	negative
NICKNAMER	suffix for forming nicknames
NMLZ	nominalizer
NOM	nominative
NPST	nonpast
NSG	non-singular
NSPEC	non-specific
NTVZ	nativizer
NUM	numeral
OBJ	object
OBJVZ	objectivizer
OBL	oblique
OBLIG	obligative
OBV	obviative
ONOM	onomoatopoeia
OPT	optative
ORD	ordinal
P	patient-like argument of canonical transitive verb
PARTIT	partitive
PASS	passive
PEJ	pejorative
PERL	perlative
PERMIS	permissive
PERSIST	persistive
PFV	perfective
PL	plural
POL	polite
POSS	possessive
POT	potential
PRAG	pragmatic marker
PRED	predicate/predicative
PREDADJ	predicative adjective
PREP	preposition
PREP	preopositional case
PRF	perfect
PRO	pronoun
PROB	probabilitive
PROG	progressive
PROH	prohibitive
PROP	proper noun
PROX	proximal
PRS	present
PST	past
PTCL	particle
PTCP	participle
PURP	purposive
PV	patient voice
PVB	preverb
Q	question
QUANT	quantifier

QUOT	quotative
RECENT	recent past tense
RECNF	reconfirmative
RECP	reciprocal
REF	referential
REFL	reflexive
REL	relative
REM	remote (past/future)
REP	reportative
RES	resultative
RESTR	restrictive
REVERS	reversive
S	single argument of canonical intransitive verb
SAME.SBJ	same subject
SBJ	subject
SBJV	subjunctive
SEQ	sequential
SG	singular
SIM	simultaneous
SOC	sociative
SPEC	specific
STAT	stative
STEM	stem (esp. in languages with multipartite stems)
SUPERL	superlative
SURP	surprise
TEASER	form for teasing people
TEL	telic
TEMP	temporal
TENSE	unspecified tense marker
TERM	terminative
TOP	topic
TR	transitive
UP	distal deixis pointing up
V	verb
V2	vector verb with unspecified function
V.AUX	verbal auxiliary
V.CAUS	causative verb
V.IMP	imperative verb
V.ITR	intransitive verb
V.PASS	passive verb
V.POS	positional verb
V.TR	transitive verb
VBZ	verbalizer
VOICE	voice marker with unspecified function
VN	verbal noun
VOC	vocative
VOL	volitional
WH	wh-word

The following characters have special meanings:

- . joins several functions expressed by a single morpheme, e.g. “IND.PST”
- / joins alternative functions of a morpheme for which no common label is available, e.g. “1/2” (= 1st or 2nd person)
- _ joins several metalanguage words coding a single object language function, e.g. “put_on”
- > agent acting on patient; possessor and possessum

3.5.4 Part-of-speech tags

The ACQDIV Corpus uses a standardized set of part-of-speech tags in the column `morphemes.pos`. The set was deliberately kept small in order to make broad comparisons across languages possible. The original tags are maintained in the column `morphemes.pos_raw`. `NULL/NA` is inserted when the part of speech is unknown. Tags not contained in the Leipzig Glossing Rules are again marked by an asterisk in the list below.

ADJ	adjective
ADV	adverb
ART	article
AUX	auxiliary
CLF	numeral classifier
CONJ*	conjunction
IDEOPH*	ideophone
INTJ*	interjection

N*	noun	PVB*	preverb
NUM*	numeral	QUANT*	non-numeral quantifier
pfx*	prefix	sfx*	suffix
POST*	postposition	stem*	stem
PREP*	preposition	V*	verb
PRODEM*	pronouns/demonstratives		
PTCL*	particle		

The Universal Dependency (UD) part-of-speech tag is added to the column `words.pos_ud`. It is derived from the raw POS label and not from the standardized ACQDIV tag, i.e. every corpus has a separate mapping which is defined in the corresponding configuration file. The reason for this is that the UD tags are more specific in some cases. For instance, the UD tag-set distinguishes between determiners (DET) and pronouns (PRON) whereas the ACQDIV tag-set conflates them to PRODEM. This would lead to arbitrary mappings like 'PRODEM=PRON' which would bias the UD label distribution. There are also numerous cases where the raw tags are less specific than the UD tags. In these cases, we map them to the most common equivalent. All cases are listed in Table 3.12.

Corpus	**Mapping**	**Comment**
Chintang	gm = PART	Some of these are ADP. In Nepali, CCONJ and SCONJ are also possible.
Chintang	n = NOUN	PROPN are not marked.
Chintang	pro = PRON	There is no lexical distinction between referential and adnominal pronouns in Chintang, but in UD they would probably be tagged as DET even when simply used adnominally in syntax.
Cree	p,conjn = CCONJ	It seems like there is no difference between subordinating and coordinating conjunctions in Cree, and all conjunctions in our corpus have glosses that one would rather associate with a coordinating function. However, there are very clear (verbal inflectional) markers of subordination with which these conjunctions can co-occur. Thus, UD might require that they be tagged as SCONJ in such cases.
Cree	pro,* = PRON	This is a whole class of tags, some of which might also be DET in the UD framework.
Inuktitut	DEM, DM, DR, LR = PRON	These are demonstrative stems whose translation and classification depends a lot on case. In the ABS or ERG, they correspond to English pronouns, in the various LOC cases to (pronominal) adverbs such as 'here', 'there', which in UD would be tagged ADV.
Japanese MiiPro/Miyata	conj = CCONJ	Some of these would probably be counted as SCONJ under the UD definition, but most are CCONJ.
Japanese MiiPro/Miyata	n:deic:pr(e)s = PRON	The personal pronouns behave like ordinary nouns in Japanese, but this classification is probably more in the comparative spirit of UD.

Table 3.12: Problematic mappings of raw to UD POS tags.

Corpus	Mapping	Comment
Japanese MiiPro/Miyata	ptl:conj = PART	This is a heterogeneous class, some of whose members would rather correspond to ADP or SCONJ, depending on their use in syntax.
Nungon	d, dem = PRON	Some of these can probably be used adnominally, i.e. they would be DET depending on use.
Nungon	n = NOUN	PROPN is not distinguished.
Russian	CONJ = CCONJ	Some of these are definitely SCONJ, but the two types that cover 70% of all tokens ('a' and 'i') are CCONJ.
Russian	NA = PRON	This also includes some potential DET.
Russian	PRO-DEF, PRO-DEM, PRO-INTERROG, PRO-REFL = PRON	This could also be DET. Note, though, that in most cases Russian consistently distinguishes between adnominal and referential use, e.g. PRO-DEM-ADJ vs. PRO-DEM-NOUN. Thus, this is much less of a problem than in the other corpora.
Sesotho	cj = CCONJ	Most of these seem to be CCONJ, but it is not excluded that there are also SCONJ. The grammatical situation is similar to Cree, i.e. subordination is mainly expressed via verbal inflections.
Sesotho	d = PRON	These words can also be used adnominally (DET).
Sesotho	ps = DET	Possessives are adnominal by default, but they can also refer (e.g. 'Whose is this?').
Turkish	CON* = CCONJ	The two most frequent types, which cover 80% of all tokens ('dA', 'ama'), are clearly CCONJ, but others might be SCONJ.
Yucatec	CONJ = CCONJ	The two most frequent types are CCONJ and cover 75% of all tokens ('kux', 'pero'). Others might be SCONJ.
Yucatec	DET = DET	The Yucatec tag also covers forms that can be used referentially. It is not clear what criteria the use of DET in the corpus was based on (it does not seem to be syntax).
Yucatec	QUANT = PRON	This includes many forms that can also be used adnominally (DET), but only syntactic annotations would help us.

Table 3.12: Problematic mappings of raw to UD POS tags.

Chapter 4

Data sources

This chapter describes the structure of the input data and how it is mapped to the target structure found in the ACQDIV Corpus. The `warnings` tiers found on the utterance, word, and morpheme levels are inserted during postprocessing and are therefore ignored below.

Note that the input data used for the ACQDIV Corpus are a subset of the original data. Tiers that were not present in the majority of corpora were generally ignored, as were parts of the subcorpora whose target children were out of the core age range (2;0.0-3;12.31) during the whole recording period. For this reason this chapter cannot be a complete documentation of the original data and may often diverge from the original documentation (which is linked below, if existing).

There were two valid input formats, TalkBank XML and Toolbox. A third important format is CHAT: most corpora were originally formatted as CHAT and still contain traces of it. These were converted from CHAT to a valid input format either by the respective corpus teams or by the ACQDIV core team. Section 4.1 gives a brief introduction to the three formats. For details on conversion work done by the ACQDIV core team, see Chapter 5. The remaining sections in this chapter deal with the particularities of the individual subcorpora.

The locations of tiers in XML corpora are specified using XPath. Toolbox corpora are flatter, so it is sufficient to give the tier name here.

4.1 Original corpus formats

4.1.1 CHAT

CHAT is the original format of most subcorpora: Indonesian, Inuktitut, Japanese MiiPro and Miyata, Russian, Sesotho, Turkish, and Yucatec. The Indonesian and Russian corpora had already been converted to Toolbox by the corpus teams at the time they were added to the ACQDIV Corpus, so the input format was Toolbox. Similarly, the two Japanese corpora and Sesotho had been converted to TalkBank XML using the existing parser Chatter, so only Inuktitut, Turkish, and Yucatec were still only available as CHAT at the beginning of the project. Nevertheless, traces of CHAT are omnipresent in all corpora listed above, be it because of imperfect conversion routines or because of deliberate exceptions.

CHAT is the format associated with the CHILDES online archive of child language acquisition corpora (MacWhinney 2000). The full specification can be found at http://childes.psy.cmu.edu/manuals/CHAT.pdf. The most important characteristics of CHAT are as follows.

One corpus file corresponds to one recording session (or sometimes to a smaller stretch corresponding to the length of a tape). Each file contains the metadata for the session and all speakers in its head and the primary data (transcriptions and all annotations) in its body. Corpus-level metadata are stored in separate text-based files with the extension cdc. The body part of corpus files is divided into utterance blocks, where each utterance block in turn consists of one or several lines

corresponding to different tiers. The first line in an utterance block is the main transcription tier and all following lines are annotations associated with it. An example for the first few lines of a CHAT corpus file is shown in the screenshot in Figure 4.1. The file was opened in CLAN, the editor associated with the format.

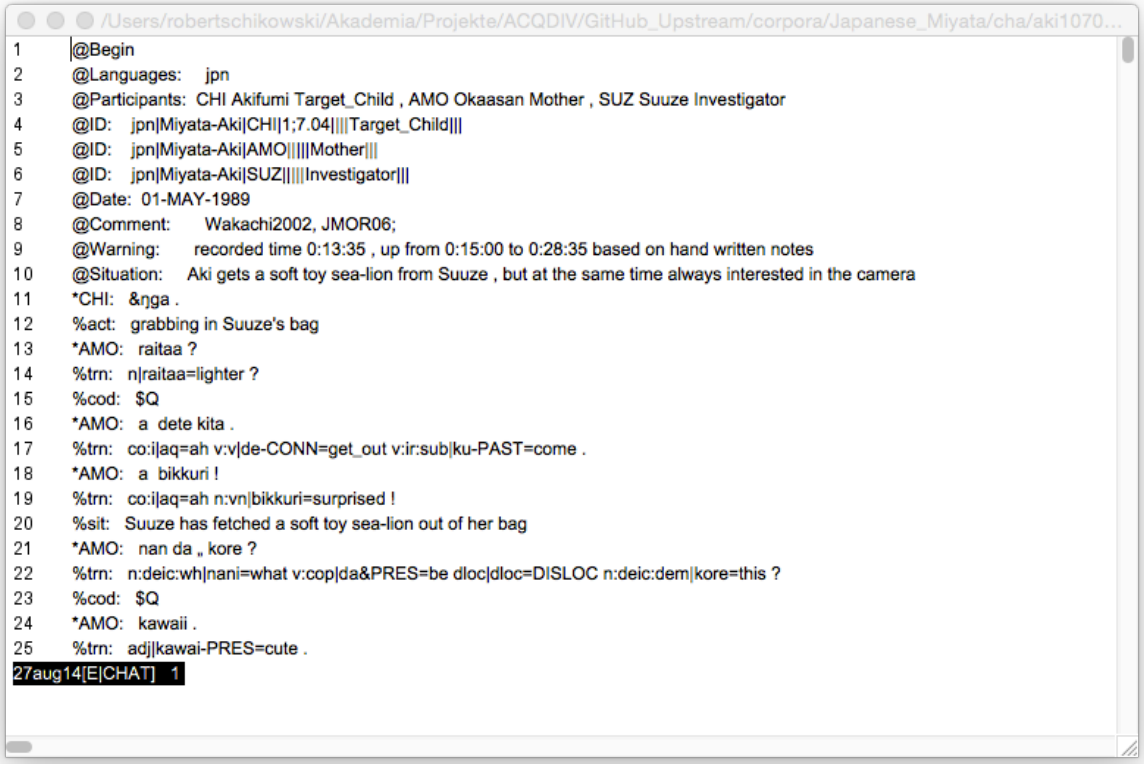

Figure 4.1: The first lines of a typical CHAT file, opened in CLAN

A peculiarity of CHAT, which makes it as difficult to keep it consistent as it is to parse it, is that logical tiers are often not kept separate in the syntax (i.e. information belonging to different tiers may be coded on a single line) and that a multitude of special characters in various combinations is used to accommodate such "dislocated" annotations. For instance, error coding, coding for action accompanying speech, comments on the language and register of individual words, prosodic and/or pragmatic markers, and free comments may all be inserted on the main transcription tier using various kinds of brackets, asterisks, equal signs, at signs combined with single-letter codes, and various combinations of punctuation markers.

Take the string "dashiyo(o)@n [= dasoo] [*] ka ?" (taken from Japanese MiiPro, als19990706.cha) as an example. Here, what the child said is *dashiyo ka.* "(o)" means the transcriber assumes this form has been shortened (the target form would have had a long [o:]), "@n" indicates that the same word is a neologism, "[= dasoo]" gives the adult target form, "[*]" marks *dashiyo* as an error, and "?" marks the whole utterance as a question.

This mismatch between corpus syntax and semantics also was the reason why CHAT was not accepted as an input format for the ACQDIV Corpus. The three corpora that initially were only available as CHAT were cleaned and converted to TalkBank XML as described in Chapter 5. Nevertheless, CHAT is present in fragments in almost all input data, especially in the morphology tier `%mor:`, whose syntax is so complicated and inconsistent that it turned out to be easier to take it over

without changes into TalkBank XML and parse it from there than to first make it conform to CHAT and then convert it.

The following tools are associated with CHAT:

- CLAN can be used for editing and validating CHAT and can perform basic statistics. It can be downloaded from http://childes.psy.cmu.edu/clan/. Documentation can be found at http://childes.psy.cmu.edu/manuals/clan.pdf.
- Chatter is a parser that can transform CHAT to TalkBank XML. It can be downloaded from http://talkbank.org/software/chatter.html. There is no comprehensive documentation available.

4.1.2 TalkBank XML

TalkBank XML is an XML format closely associated with CHILDES and CHAT (and a few other formats and archives, see http://talkbank.org/). There is a parser from CHAT to TalkBank XML (Chatter) that can deal with all standard CHAT constructs. Currently there are two different yet closely related schemas describing the structure of TalkBank XML:

- http://talkbank.org/talkbank.xsd
- https://talkbank.org/software/talkbank.xsd

TalkBank XML keeps the basic structure of CHAT, coding all data and most metadata associated with a session in a single XML file with a head and a body section. Within the body section, the utterance and word levels are marked by the nested tags `<u>` and `<w>`, respectively. While the CHAT primary transcription tier is split up into words in TalkBank XML, all other tiers are taken over *en bloc* and appear directly under `<u>` as `<a>` with various attributes marking the tier type.

The frequent mismatches between corpus syntax and semantics that are characteristic of CHAT carry over to TalkBank XML, where they are variously coded as attributes of words or groups of words (`<g>` between `<u>` and `<w>`), as tags nested in `<w>`, or as tags on the same level as `<w>` (grouped together with it via `<g>`).

Only one of the XML corpora in the ACQDIV Corpus (Japanese Miyata) has explicit XML coding for morphemes. In the other corpora, morphology is coded in less explicit, often very dense and idiosyncratic formats and is structurally located on the utterance level.

This is also the main reason why TalkBank XML is not trivial to process. The most important problematic constructs are briefly described below. All of them have the following properties in common:

- They feature a contrast between an actual and a target form (cf. Section 3.4.9).
- They are coded by a complicated syntax that makes it hard to process them (especially when nested).
- Words containing them are frequently not glossed, giving rise to alignment problems between the word and morpheme levels.

Table 4.1 shows an overview of the existing constructs and their distribution in the ACQDIV original data. "-" indicates absence, "+" presence; where a construction is present, "+g" indicates that it is normally glossed, "+n" that it is normally not glossed.

Untranscribed words

Words can be untranscribed for various reasons but mostly are because they are unintelligible. This is coded in TalkBank XML as `<w untranscribed="unintelligible">xxx </w>`. In the ACQDIV Corpus, the actual and target form for such words is `NULL/NA` (= 'unknown').

	CCLAS crl	AIC ike	MPJC jpn	MYJC jpn	DSC sot	KULLDD tur	PYC yua
untranscribed	+g	+g	+n	+n	+g	+n	+n
fragments	-	+n	+n	+n	-	-	-
omissions	+?	-	-	-	-	-	-
replacements	-	-	+g	+g	-	+g	-
shortenings	+?	+g	+g	+g	+g	+g	+g
repetitions	-	+g	+n	-	-	+n	-
retracings	-	+g	+n	+n	-	+g	-

Table 4.1: Distribution of actual/target constructs in the ACQDIV original data

Fragments

An actual word with no clear target is called a fragment. TalkBank XML codes this as `<w type="fragment">...</w>`, where the "..." part marks the transcribed actual word. The target is set to `NULL/NA` in the ACQDIV Corpus.

Omissions

Rarely, no actual word is present but the target syntax suggests there should have been. In this case the `<w>` tag contains the target form and is marked as an omission by the type attribute: `<w type="omission">...</w>`. Omissions are completely ignored in the ACQDIV Corpus (i.e. the omitted target word is not represented at all) because they are rare and considered to be speculative.

Replacements

Children often replace a target word by another, similar actual word. TalkBank XML codes this in a rather complicated manner: all involved words appear within the group tag `<g>`. The actual form is the text of an ordinary `<w>` tag, which is the parent of a `<replacement>` tag that can itself contain an arbitrary number of `<w>` tags having the target form(s) as their text. The schema in short is `<g><w>...<replacement><w>...</w></replacement></w></g>`. The actual/target contrast is taken over into the ACQDIV Corpus. Where a single actual word corresponds to several target words, new empty actual words are inserted, creating a 1:1 correspondence between the two groups of words.

Shortenings

Shortenings feature a contrast between a full target string and a contracted actual string. The target string may once more consist of several words, so the group tag `<g>` is used: `<g><w>...<shortening>...</shortening>...</w></g>`. The substring that was omitted is the text of the nested `<shortening>` tag. The full target string is the text of the parent `<w>` before *and* after `<shortening>`. Shortenings are mapped similarly to replacements.

In some corpora, shortenings may appear within replacements, causing particularly convoluted constructions. For instance, Japanese MiiPro (aprm19990722.xml) contains the following XML string: `<w>kitenee<replacement><w>kite</w><w><shortening>i</shortening>nai</w></replacement></w>`. This basically assumes two layers of target strings: the ultimate target *kite inai* (two words) first gets shortened to *kite nai*, which is then "replaced" by *kitenee* (one word).

Repetitions

Repetitions are specially marked when they diverge from the target syntax (where repetitions are sometimes expected, e.g. where full reduplication is a means of grammatical marking). TalkBank XML does this by embedding a special tag `<r>` with an attribute specifying the number of repetitions under `<w>`: `<g><w>...</w><r times="..."></g>`. In the ACQDIV Corpus repetitions are simply spelt out (n times the same word). The corresponding glosses are repeated, too.

Retracings

Retracing in conversation analysis is the action of canceling an utterance at a given point in order to restart it or switch to a different utterance. TalkBank XML groups the complete group of words assumed to be part of a canceled utterance by `<g>` and uses the special tag `<k>` with a type attribute to mark what is happening: `<g><w>...</w><w>...</w><k type="retracing"/> </g>`. Retracings are not marked specially in the ACQDIV Corpus. However, when there are no glosses for the cancelled portion, the parser tries to suggest a gloss from similar parts of the rest of the utterance.

4.1.3 Toolbox

Toolbox is a textual format that is associated with the software of the same name and has been developed by SIL international. General documentation and links to downloads can be found at http://www-01.sil.org/computing/toolbox/.

Typical Toolbox corpus files code sessions as trees where the three central levels are utterance, word, and morpheme, very much as in CHAT and TalkBank XML. However, differently from these, the syntactic coding of this structure is highly implicit. The syntactic unit corresponding to the utterance level is the record. Records are delimited by a record ID at the top and a double linebreak at the end. Each record may have several tiers consisting of a so-called field marker, which starts with a backslash and indicates the type of content (e.g. "\ps" for parts of speech), and of the content itself (e.g. "adj"). The association of annotations with the three levels (utterance, word, morpheme) is not explicitly coded.

All elements on a tier (words or morphemes) are separated by spaces. Alignment across tiers works via indices: the first element on one tier (e.g. a segment) is associated with the first element on another (e.g. a gloss), the second with the second, and so on. The various other fields listed above are all on separate tiers in Toolbox.

Dependent morphemes are marked by morpheme separators on one side (e.g. "un-" for prefixes, "-able" for suffixes). These separators make it possible to reconstruct word boundaries on a tier focussing on morphemes. Sequences of the types stem-stem, stem-prefix, and suffix-stem can be inferred to belong to different words, whereas stem-suffix, suffix-suffix, prefix-stem, and prefix-prefix must belong to the same word. A "floating separator" (morpheme separator with spaces on both sides) can be used to indicate that two stems belong to the same word (e.g. in the case of compounds: "apple - tree -s").

Figure 4.2 shows an example for one record in a typical Toolbox file.

4.2 Chintang

4.2.1 Publication, accessibility, documentation

The Chintang Language Corpus (Stoll et al. 2015) was compiled between 2004 and 2015 in the course of several research projects now summarized as the Chintang Language Research Program (CLRP). It is documented in the Conventions for the linguistic analysis of Chintang (Schikowski 2015). The standard citation for the language acquisition subcorpus, which is the portion included in the ACQDIV Corpus, is as follows:

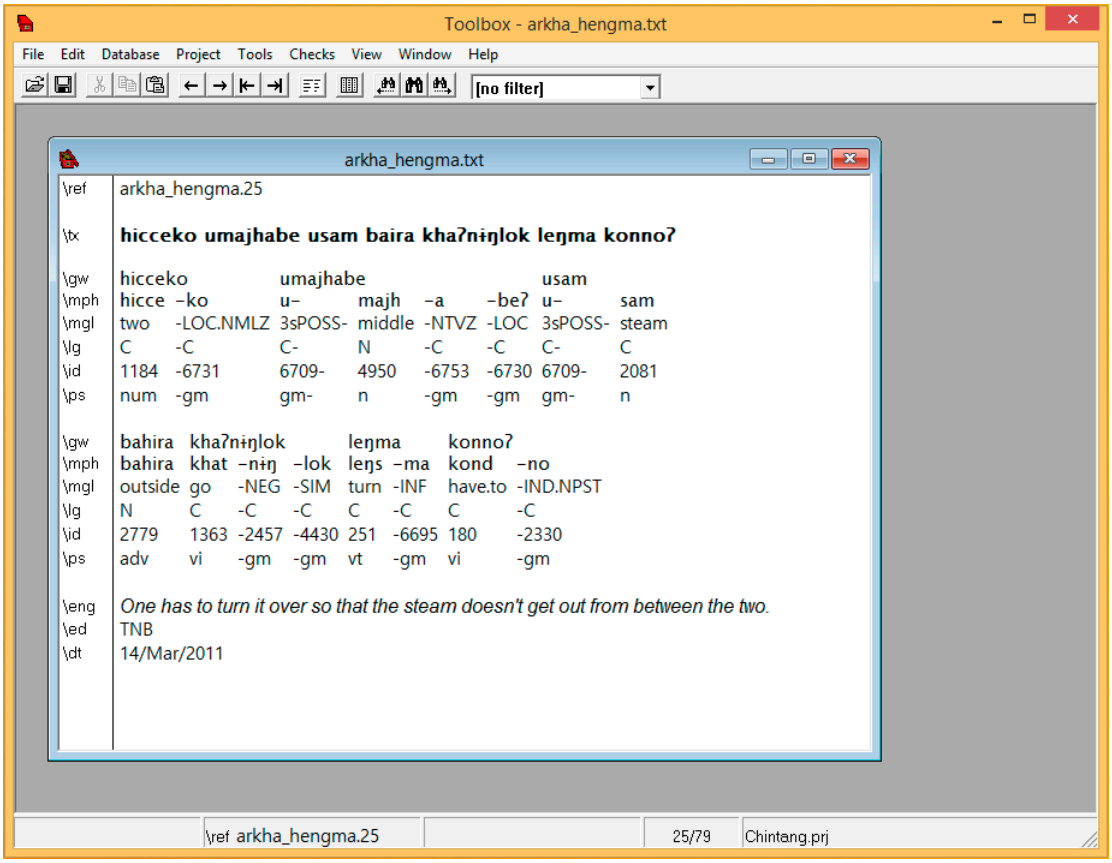

Figure 4.2: The first lines of a typical Toolbox file, opened in the Toolbox program

> Stoll, Sabine, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Manoj Rai, Novel Kishor Rai, Ichchha P. Rai, Taras Zakharko, Robert Schikowski & Balthasar Bickel. 2015. *Audiovisual corpus on the acquisition of Chintang by six children.*

An older version of the corpus was published in the DoBeS archive at the MPI Nijmegen. This version is now outdated and the publication guidelines are under revision.

4.2.2 Recording scheme

number of children	7 (one canceled early)
age ranges	0;7.23-0;7.25, 0;6.30-1;11.12, 0;6.12-1;9.20, 2;1.9-3;5.25, 2;0.29-3;5.13, 3;0.14-4;4.25, 2;11.2-4;3.14
recording rhythm	4h per month (taken during several sessions within a single week)
recording environment	mainly outside, close to home
other speakers	relatives, other children, passers-by
other languages	Nepali, Bantawa

Table 4.2: Recording scheme for the Chintang corpus

4.2.3 File system and formats

All files are located in a single folder. Files in the language acquisition subcorpus follow a naming scheme that is best understood on the base of examples such as "CLDLCh2R03S10" and "CLLDCh1-R10S01". The detailed rules are as follows:

- "CL" ("Child Language") is prefixed to all file names.
- "DL" ("Days in the Life of") is prefixed to baby sessions (range 0;6-2;0), "LD" ("Linguistic Development") to sessions with older target children.
- "Ch" combined with the following number indicates the speaker code of the target child.
- "R" and "S" (each with following numbers) indicate the recording cycle (= the number of the month, "01" being the first month in which recordings were taken for a child) and the number of the session within that month.

All corpus files are encoded as UTF-8 text. Tiers containing Chintang words frequently feature the special characters ⟨ŋ⟩, ⟨ɨ⟩, ⟨ʔ⟩, ⟨ṽ⟩ (Tilde on vowels, U+0303). Nepali translations contain Devanagari letters and punctuation.

4.2.4 Corpus format

The input format used for the ACQDIV Corpus is Toolbox. Table Table 4.3 shows how the fields in the ACQDIV Corpus are related to tiers in the input.

target table	target field	source
sessions	session_id_fk	file name
utterances	utterance_id	\ref
utterances	start	\ELANBegin
utterances	end	\ELANEnd
utterances	speaker_label	\ELANParticipant
utterances	addressee	\tos
utterances	childdirected	\tos
utterances	sentence_type	utterance delimiter on \nep
utterances	utterance_raw	\gw
utterances	translation	\eng
utterances	comment	\comment
words	word	\gw
morphemes	morpheme	\mph
morphemes	gloss_raw	\mgl
morphemes	pos_raw	\ps
morphemes	morpheme_language	\lg

Table 4.3: Chintang tiers

Morphology in the Chintang corpus is coded in the regular Toolbox format.

4.3 Cree

4.3.1 Publication, accessibility, documentation

The Cree corpus (Brittain 2015) is associated with the Chisasibi Child Language Acquisition Study (CCLAS), which started in 2004 and will continue until 2018. It should be cited as:

Brittain, Julie. Corpus of the Chisasibi Child Language Acquisition Study (CCLAS). http://childes.psy.cmu.edu/.

A fully anonymized version of a small subcorpus is freely available from CHILDES. This is also the subcorpus incorporated into the ACQDIV Corpus.

Some documentation for the CCLAS corpus can be found in the Cree Auto-Parser Guide (Acton 2013).

4.3.2 Recording scheme

The following information holds for the subcorpus included in the ACQDIV Corpus ("Ani corpus"):

number of children	1
age ranges	2;1.14-3;8.24
recording rhythm	30-40 min every 2-3 weeks
recording environment	indoors at home
other speakers	mainly mother
other languages	English

Table 4.4: Recording scheme for the Cree corpus

4.3.3 File system and formats

Cree file names are composed of an ascending number (for files within one subcorpus), a code for the target child, and the recording date in the format YYYY-MM-DD, e.g. "09-A1-2005-10-17". Some files have an undocumented suffix "ms" behind the date, e.g. "10-A1-2005-11-21ms".

All files are encoded as UTF-8 text. Tiers containing Cree words frequently feature vowels from the ASCII set with an additional circumflex, e.g. ⟨â⟩ (U+00E2). There are phonetic transcriptions which feature a bigger set of IPA characters.

4.3.4 Corpus format

The input format used for the ACQDIV Corpus is TalkBank XML (converted from Phon by the Cree team). Table Table 4.5 shows how the tiers in the ACQDIV Corpus are related to tiers in the input. The following peculiarities exist in the Cree input:

- The phonetic transcription is taken from `//u/actual`. However, this node always has children `ph` and `ss`, each of which contains a single segment or suprasegmental, respectively. These single values are concatenated for the representation in the ACQDIV Corpus.
- `//u/w` may contain the string `missingortho` when it is empty. This is redundant and therefore removed.
- `//u/w` contains "_" as a morpheme separator. Since this is neither part of the orthography used in this tier nor required for parsing the morphology, it is removed.
- All tiers containing transcriptions (including morphology tiers) may contain semantically redundant square brackets at their edges – these are removed.

The Cree morphology tiers are structured as follows:

- Words are separated by spaces, morphemes are separated by "~".

target table	target field	source tier
sessions	session_id_fk	/CHAT@Id
utterances	utterance_id	//u@uID
utterances	start_raw	//media@start
utterances	end_raw	//media@end
utterances	speaker_label	//u@who
utterances	addressee	-
utterances	sentence_type	//u/t
utterances	utterance_raw	//u//w
utterances	translation	//u/a[@type="english translation"]
utterances	comments	//u/a[@type="comments"]
words	word	//u//w
morphemes	morpheme	//u/a[@type="extension" and @flavor="actmor"]
morphemes	gloss_raw	//u/a[@type="extension" and @flavor="mormea"]
morphemes	pos_raw	//u/a[@type="extension" and @flavor="mortyp"]
morphemes	morpheme_language	//u/a[@type="extension" and @flavor="mormea"], special gloss `Eng`

Table 4.5: Cree tiers

- "%%%" indicates untranscribed words, "#" is for unglossed elements. Both are replaced by `NULL/NA` (in isolation) or "???" (within strings).
- "?" is used instead of a gloss when the meaning of a morpheme is not clear. It may be isolated (e.g. "=?", unclear suffix) or follow a form (e.g. "=h?", might be suffix *-h*). This, too, is replaced by `NULL/NA`.
- "*" marks an element on the actmor or tarmor tier that does not correspond to an element on the other tier. It is replaced by `NULL/NA`.
- "." and "+" connect two glosses to one. "," adds an additional specification to a gloss, e.g. "p,quest" (question particle). "+" and "," are replaced by the more standard ".".
- Transitive agreement in the gloss tiers is marked by numbers connected by spaces and a greater than sign, e.g. "2 > 1". The spaces are removed.
- Brackets indicate covert grammatical categories in the mortyp tier. In tarmor, they are used around abstract morphemes with no overt morphological shape in order to make morpheme numbers match across tiers. The meaning of the individual abstract morphemes is not clear. They are uppercased by the parser in order to emphasize their grammatical status.
- "/" seems to mark semantic underspecification, e.g. "yellow/green".
- "Eng" stands for any English word in the gloss tiers. The gloss is replaced by the English word itself.

4.4 Indonesian

4.4.1 Publication, accessibility, documentation

The Indonesian corpus (Gil & Tadmor 2007) was collected at the Jakarta Field Station of the Max Planck Institute for Evolutionary Anthropology between 1999 and 2004. It is officially cited as:

Gil, David & Uri Tadmor. 2007. *The MPI-EVA Jakarta Child Language Database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.* https://jakarta.shh.mpg.de/acquisition.php.

An earlier release of the full corpus is freely available at CHILDES. However, the most recent data can now be downloaded from the website of the MPI Jakarta Field Station, https://jakarta.shh.mpg.de/acquisition.php. Documentation is available on the same website and also in the CHILDES manual for East Asian languages.

4.4.2 Recording scheme

number of children	10
age ranges	1;6.15-4;11.29, 1;8.14-5;11.14, 1;9.15-6;1.5, 2;0.11-3;10.29, 2;10.20-6;4.9, 2;7.4-6;0.20, 3;4.20-6;5.27, 4;6.0-8;9.29
recording rhythm	45-60 min every 7-10 days
recording environment	mostly indoors at home, sometimes outdoors
other speakers	variety of children and adults
other languages	Javanese, Traditional Betawi, Toba Batak, others

Table 4.6: Recording scheme for the Indonesian corpus

4.4.3 File system and formats

The Indonesian corpus has several subfolders containing subcorpora for each target child and named after their code. Within the folders, Toolbox files are named as "COD-DDMMYY" (where "COD" represents the speaker code of the target child) and XML files are named as "YYYY-MM-DD" with no indication of the target child. Note that this makes XML file names potentially ambiguous when taken out of their folders.

Indonesian files are encoded as UTF-8 text and do not contain any non-ASCII characters.

4.4.4 Corpus format

The corpus data for this corpus are taken from the Toolbox version while the metadata are based on the corresponding TalkBank XML files. Both formats have been converted from CHAT by the Indonesian team. Table Table 4.7 shows how the tiers in the ACQDIV Corpus are related to tiers in the input.
Some peculiarities should be noted in the Indonesian input:

- The Indonesian Toolbox data contain CHAT constructs in the transcriptions (e.g. truncations as "(ba)nana"), which are dealt with as the corresponding structures in the TalkBank XML corpora (see Section 4.1.2).
- The first two records of many Toolbox file contain metadata imported from CHAT and dummy markers which do not convey any information. These contents have been put on ordinary Toolbox tiers, which have a different meaning in the body of Toolbox files (see Table 4.7 above). Where this is the case these tiers are ignored. The following tiers may be affected:
- Two different formats are used for speaker codes. The base format is found in CHAT and XML and consists of three uppercase letters. Codes in this format are not unique – for instance, "CHI" can stand for any of the eight target children and "MOT" for any of their mothers. In

target table	target field	source tier
sessions	session_id_fk	file name
utterances	utterance_id	\ref
utterances	start_raw	\begin
utterances	end_raw	-
utterances	speaker_label	\sp
utterances	addressee	-
utterances	sentence_type	utterance delimiter on \tx
utterances	utterance_raw	\tx
utterances	translation	\ft
utterances	comment	\nt
words	word	\tx
morphemes	morpheme	\mb
morphemes	gloss_raw	\ge
morphemes	pos_raw	-

Table 4.7: Indonesian tiers

tier	divergent content
\sp	dummy marker @PAR (first record) or @Begin (second record)
\tx	speaker codes (CHAT @Participants, first record) or dummy marker @Begin (second record)
\pho	associated media file (CHAT @Filename)
\ft	duration of media file (CHAT @Duration)
\nt	comments on recording situation (CHAT @Situation)

Table 4.8: Indonesian tiers with differing contents in the first two Toolbox records

order to make codes unique, there is another, extended format where either a code for the target child is suffixed to the base code (e.g. “CHIHIZ”, “MOTHIZ” = child and mother in the subcorpus for the target child HIZ) or, in the case of researchers, “EXP” is prefixed to the base code (e.g. “EXPBET”). This format is used in Toolbox and in a flat Excel metadata table. The ACQDIV Corpus uses the extended format throughout.

- Indonesian is the only corpus without any part-of-speech annotation.

Morphology in the Indonesian corpus is coded in the regular Toolbox format (see Section 4.1.3).

4.5 Inuktitut

4.5.1 Publication, accessibility, documentation

The Inuktitut Corpus (Allen Unpublished) was compiled for the work in Allen (1996), which also contains some documentation. The corpus itself has not been published and is cited as:

Allen, Shanley. Unpublished. Allen Inuktitut Child Language Corpus.

number of children	4
age ranges	2;6.6-3;3.2, 2;0.11-2;9.5, 2;6.2-3;2.26, 2;9.16-3;6.12
recording rhythm	4h every month
recording environment	indoors at home
other speakers	relatives, friends
other languages	(little) English

Table 4.9: Recording scheme for the Inuktitut corpus

4.5.2 Recording scheme

4.5.3 File system and formats

Recording sessions in the Inuktitut corpus may correspond to a single file or to a folder containing several transcripts associated with successive tape portions. Folders are named according to the scheme "COD"+"DDMMM" (where "COD" is the speaker code of the target child and "MMM" are three-letter month abbreviations, hence e.g. "ALI2APR"). Files within folders are named as "COD"+"DD", an ascending number and/or letter for tape portions, and the suffix "TF", e.g. "ALI71TF" in the folder "ALI7SEP". Single files not placed in folders also start with "COD" but apart from that do not have clear naming conventions (e.g. "SUP11WM").

The original Inuktitut files come with a variety of file extensions (.XXS, .XXX, .NAC) which, however, amount to plain text (structured as CHAT). They are mostly encoded as ISO-Latin and contain a number of unexpected special characters (Inuktitut itself does not use non-ASCII characters, nor should any of the annotations).

4.5.4 Corpus format

The input format for this corpus is TalkBank XML (cleaned and converted from CHAT by the ACQDIV core team). Table Table 4.10 shows how the tiers in the ACQDIV Corpus are related to tiers in the input.

target table	target field	source tier
sessions	session_id_fk	/CHAT@Id
utterances	utterance_id	//u@uID
utterances	start_raw	//u/a[@type="timestamp"]
utterances	end_raw	-
utterances	speaker_label	//u@who
utterances	addressee	//u/a[@type="addressee"]
utterances	childdirected	//u/a[@type="addressee"]
utterances	sentence_type	//u/t, //u/e
utterances	utterance_raw	//u//w, //u/ga[@type="alternative"]
utterances	translation	//u/a[@type="english translation"]
utterances	comment	//u/a[@type="comments"], //u/a[@type="situation"]
words	word	//u//w, //u/ga[@type="alternative"]
morphemes	morpheme	//u/a[@type="extension" and @flavor="mor"]
morphemes	gloss_raw	//u/a[@type="extension" and @flavor="mor"]
morphemes	pos_raw	//u/a[@type="extension" and @flavor="mor"]

Table 4.10: Inuktitut tiers

Several things should be noted for the Inuktitut input:

- While actual words are normally found in `//u/w`, occasionally `w` occurs under `//u/g` together with another tag `ga` with the attribute `type="alternative"`. When this is done, `ga` contains the actual word and `w` contains the target word.
- The corresponding construction on the morphology tier is an actual word followed by "[=? ...]", where the brackets contain the guessed target form. Occasionally there may be several target forms and several corresponding glosses, e.g. `<g><w>qaungatillaunga</w><ga type="alternative">maungatillaunga</ga><ga type="alternative">paungatillaunga</ga></g>` In this case only the first target form is used.
- There are some pseudo-compounds where one element is always "xxx" e.g. `<w>ski-doo<wk type="cmp"/>xxx</w>`, gloss `VR|sikituuq^ride_snowmobile+xxx`. This also includes pseudo-compounds of the type "cli" (clitic), which is likewise meaningless. The "xxx" stands for an unidentified morpheme. This type of compound is treated like a normal word without any internal boundaries on the word level; on the morpheme level, "xxx" is treated as an unknown morpheme.
- Inuktitut has a special sentence type "broken for coding", which indicates that there is no gloss but does not say anything about the sentence type. A warning "not glossed" is inserted into the corpus object and the sentence type is set to "default".

The Inuktitut morphology tier (taken over from CHAT) has the following internal structure:

- Words are separated by spaces, morphemes by "+".
- Each morpheme consists of three components (identical for lexical and grammatical morphemes):
  - The core element is a phonological form.
  - A POS tag is prefixed to this form, using "|" as the separator. Sometimes there are several POS tags, all separated by "|" (e.g. "NN|DIM|apik"). Labels further to the right are interpreted as subcategories.
  - A gloss is suffixed to the form, using "^" as the separator.
- The following special characters are found within glosses:
  - "_" connects several words that form a single gloss (e.g. "look_for"). This remains unchanged.
  - "&" (sic) connects a stem gloss with a grammatical gloss (e.g. "here&SG_ST"). This is replaced by more standard ".".
  - "@e" marks English words and is deleted.
  - Utterance terminators such as "." or "?" are redundant on the morphology tier and therefore deleted.
  - "<", ">" (sic) mark annotation groups in CHAT. They are ignored by the parser together with any associated annotations.
- Codes in square brackets are often found at the end of the morphology tier. Some of these are generic CHAT, others are specific to Inuktitut and have been documented in Allen (1996). All of these codes are removed because they do not directly affect the interpretation of the morphology tier. The only exception is "[?]", which indicates insecure glosses and is converted to a warning.
- The glossed form normally is associated with the target form, although glosses of the actual form are also found. Additional information on the relation between actual and target form

is given in <a type="errcoding">, but the format is inconsistent, so it is impossible to exploit this tier.

- Untranscribed words are found as "xxx" on the morphology tier.

4.6 Japanese MiiPro

4.6.1 Publication, accessibility, documentation

The Japanese MiiPro Corpus (Miyata & Nisisawa 2009, Nisisawa & Miyata 2009, Miyata & Nisisawa 2010, Nisisawa & Miyata 2010, Miyata 2012) was compiled between 1997 and 2010. The four subcorpora are cited as:

Miyata, Susanne & Hiro Yuki Nisisawa. 2009. *MiiPro – Asato Corpus.* Pittsburgh, PA: TalkBank.
Miyata, Susanne & Hiro Yuki Nisisawa. 2010. *MiiPro – Tomito Corpus.* Pittsburgh, PA: TalkBank.
Nisisawa, Hiro Yuki & Susanne Miyata. 2009. *MiiPro – Nanami Corpus.* Pittsburgh, PA: TalkBank.
Nisisawa, Hiro Yuki & Susanne Miyata. 2010. *MiiPro – ArikaM Corpus.* Pittsburgh, PA: TalkBank.
Miyata, Susanne. 2012. *Japanese CHILDES: The 2012 CHILDES manual for Japanese.* Available online at http://www2.aasa.ac.jp/people/smiyata/CHILDESmanual/chapter01.html.

It is comprehensively documented in the CHILDES manual for Japanese (in Japanese) and the CHILDES manual for East Asian languages (in English).

4.6.2 Recording scheme

number of children	4
age ranges	2;11.27-5;1.23, 2;11.28-5;0.17 (×2), 3;0.1-5;0.27
recording rhythm	70 min per session, every week from 1;2 to 3;0, later every 1 or 2 months
recording environment	indoors at home in limited area
other speakers	mainly mother
other languages	none

Table 4.11: Recording scheme for the Japanese MiiPro corpus

4.6.3 File system and formats

MiiPro files are composed of the code of the target child and the recording date as "YYYY MM DD" but without any separators, e.g. "aprm19990515". The files are located in folders named after the target children.

All files are encoded as UTF-8 text. The orthography tiers contain CJK characters but are not taken over into the ACQDIV Corpus. All tiers included in the ACQDIV Corpus only contain ASCII characters.

target table	target field	source tier
session	session_id_fk	/CHAT@Id
utterance	utterance_id	//u@uID
utterance	start_raw	//u/a[type="time stamp"], //u/media@start
utterance	end_raw	//u/media@end
utterance	speaker_label	//u@who
utterance	addressee	//u/a[@type="addressee"]
utterance	childdirected	//u/a[@type="addressee"]
utterance	sentence_type	//u/t, //u/e
utterance	utterance_raw	//u/w
utterance	translation	-
utterance	comment	//u/a[@type="situation"], //u/a[@type="actions"], //u/a[@type="comments"
word	word	//u//w
morpheme	morpheme	//u/a[@type="extension" and @flavor="trn"]
morpheme	gloss_raw	//u/a[@type="extension" and @flavor="trn"]
morpheme	pos_raw	//u/a[@type="extension" and @flavor="trn"]
morpheme	morpheme_language	//u/w/langs

Table 4.12: Japanese MiiPro tiers

4.6.4 Corpus format

The input format for the ACQDIV Corpus is TalkBank XML (converted from CHAT by the MiiPro team). Table Table 4.12 shows how the tiers in the ACQDIV Corpus are related to tiers in the input.

The MiiPro morphology tier has been taken over from CHAT without changes and has the following structure in the input:

- Words are separated by spaces. There are no unique morpheme separators but various types of boundary markers.
- If there are prefixes, they are always on the left edge of a word and separated from it by a "#". The prefix string consists of the phonological shape of the prefix without a gloss.
- If there is a gloss for the stem, it is always on the right edge of the word and separated from it by a "=".
- Apart from these special markers, words consist of one or (in the case of compounds) several blocks separated by "+".
- Each block in turn consists of a POS tag, a stem (phonological shape only), and optional suffixes (gloss only, no phonological shape).
- An example for a minimal gloss is "v|mi-PST", which is a verb with the stem shape *mi* and a suffix with the function 'past'. In standard glossing the word form would be *mi-ta* and the glosses would be "see-PST". Since the MiiPro corpus leaves the meaning of the stem (and prefixes) and the shape of suffixes open, the value `NULL/NA` is filled in in the corresponding columns.
- Compounds may have an additional POS tag for the complete compound. In this case, the POS is prefixed in the usual form (xxx|) but there is no stem that follows.

4.7 Japanese Miyata

4.7.1 Publication, accessibility, documentation

The Japanese Miyata Corpus (Miyata 2004a,b,c, 2012) was collected between 1986 and 2004. The three subcorpora are cited as:

> Miyata, Susanne. 2004. *Aki Corpus.* Pittsburgh, PA: TalkBank. 1-59642-055-3.
> Miyata, Susanne. 2004. *Ryo Corpus.* Pittsburgh, PA: TalkBank. 1-59642-056-1.
> Miyata, Susanne. 2004. *Tai Corpus.* Pittsburgh, PA: TalkBank. 1-59642-057-X.
> Miyata, Susanne. 2012. *Japanese CHILDES: The 2012 CHILDES manual for Japanese.* Available online at http://www2.aasa.ac.jp/people/smiyata/CHILDESmanual/chapter01.html.

Contentwise this corpus is closely related to the Japanese MiiPro Corpus. It is documented in the same resources, the CHILDES manual for Japanese (in Japanese) and the CHILDES manual for East Asian languages (in English).

4.7.2 Recording scheme

number of children	3
age ranges	1;5.7-3;0.0, 1;4.3-3;0.30, 1;5.20-3;1.29
recording rhythm	40-60 min every week
recording environment	indoors at home
other speakers	mainly mother
other languages	none

Table 4.13: Recording scheme for the Japanese Miyata corpus

4.7.3 File system and formats

The Miyata corpus as published on CHILDES contains several files for every session. These files code the same content and are therefore doublets (or triplets), which seem to represent different workflow stages. The files come in three folders named after the target children and with partially diverging file naming conventions:

- The folder "Aki" contains three files per session. Series 1 is named as "aki" and the age of the child at the time of recording in the format "YMMDD", e.g. "aki10507" (= 1 year, 5 months, 7 days). Series 2 is named as "aki" combined with ascending numbers ("aki01" to "aki56"). Series 3 combines ascending numbers and age but does not contain the code of the child, e.g. "50_21020".
- The folder "Ryo" contains four files per session. The names for all files contain the age of Ryo in the same format as for Aki. Series 1 has the prefix "ryo" ("ryo10303"), series 2 has "yo", series 3 has "r", and series 4 has no prefix at all.
- The folder "Tai" contains four files per session. The file names of series 1 and 2 are composed of "tai" and "t", respectively, and the recording date as "YYMMDD" ("tai931125", "t931125"). Series 3 has "tai" combined with the age in the format already described ("tai21114"), and series 4 has ascending numbers combined with age ("36_20220").

All files are encoded as UTF-8 text. The orthography tiers contain CJK characters but are not taken over into the ACQDIV Corpus. All tiers included in the ACQDIV Corpus only contain ASCII characters.

4.7.4 Corpus format

The Japanese Miyata corpus is semantically very similar to the MiiPro corpus, with which it shares the author. However, the Miyata corpus does not contain any traces of CHAT but is completely formatted as TalkBank XML, which also serves as the input format for the ACQDIV Corpus. It is documented in the CHILDES manual for East Asian languages. Table Table 4.14 shows the associations between tiers in the input and in the ACQDIV Corpus.

target table	target field	source tier
sessions	session_id_fk	/CHAT@Id
utterances	utterance_id	//u@uID
utterances	start_raw	//u/a[type="time stamp"], //u/media@start
utterances	end_raw	//u/media@end
utterances	speaker_label	//u@who
utterances	addressee	//u/a[@type="addressee"]
utterances	childdirected	//u/a[@type="addressee"]
utterances	sentence_type	//u/t, //u/e
utterances	utterance_raw	//u/w
utterances	translation	-
utterances	comment	//u/a[@type="situation"], //u/a[@type="actions"], //u/a[@type="comments"
words	word	//u/w
morphemes	morpheme	//u/w/mor//mpfx, //u/w/mor//stem, //u/w/mor//mk
morphemes	gloss_raw	//u/w/mor//menx
morphemes	pos_raw	//u/w/mor//pos/c, //u/w/mor//pos/s
morpheme	morpheme_language	//u/w/langs

Table 4.14: Japanese Miyata tiers

The transcription tier in the Japanese Miyata Corpus is incomplete in that utterances of the mother have often been omitted. These omissions are not marked, so the Miyata data are not suitable for studying child-surrounding speech or adult language in general.

The Miyata input also has some peculiarities in its morphology coding:

- Different morphological components have their own tags: prefixes are coded by <mpfx> (under the morphological word <mw>) or under the compound group <mwc>), stems by <stem> (under <mw>), and suffixes by <mk> (under <mw>). Glosses are only given for stems and are coded by <menx> (under <mw> or <mwc>).
- Some clitics (e.g. honorifics) are regularly glossed, but the glosses appear in <menx> rather than <mk>. These glosses are moved to the right place by the parser.
- Some suffixes have a type attribute "fused". These are suffixes with no clear phonological shape which are fused with their stem. The glosses of such suffixes are joined to that of the preceding stem using the conventional separator ".".
- Part-of-speech tags are not given directly in <pos> but in the child nodes <c> "category" and <s> "subcategory".
- Compounds are coded for on the morphology tier. When there is a compound, the node <mwc> appears directly under <mor> with its own part-of-speech group. Prefixes and morphological words are also under <mwc> in this case. The ACQDIV Corpus ignores compounding, so the stems are concatenated using "=". The POS tag is taken from the top level rather than from the individual words.

- The glosses for words containing replacements are given *within* the `<replacement>` tag.

4.8 Nungon

4.8.1 Publication, accessibility, documentation

The Sarvasy Nungon Corpus (Sarvasy 2017b,a) was compiled between 2015 and 2017. Two resources should be cited:

> Sarvasy, Hannah. Sarvasy Nungon Corpus. Available online at http://childes.talkbank.org.
> Sarvasy, Hannah. 2017. *A Grammar of Nungon: A Papuan Language of Northeast New Guinea*. Leiden: Brill.

Some documentation is available on the CHILDES website.

4.8.2 Recording scheme

number of children	5
age ranges	2;1-4;1, 2;10-4;10, 3;5-5;5, 3;8-5;8, 1;2-2;3
recording rhythm	1 continuous hour per month
recording environment	natural environment
other speakers	various
other languages	Tok Pisin

Table 4.15: Recording scheme for the Nungon corpus

4.8.3 File system and formats

So far files are not yet systematically named, but the schema "code-age" is emerging where both code and age refer to the target child, e.g. "TowetOe-020310". The folder structure is not final yet.

All files are encoded as UTF-8 text. There is a single but frequent non-ASCII character ⟨ö⟩.

4.8.4 Corpus format

The input format for the ACQDIV Corpus is TalkBank XML (converted from CHAT). Table Table 4.16 shows how the tiers in the ACQDIV Corpus are related to tiers in the input.

The morphology tiers in the Sesotho input are structured as follows:

- Words on the target gloss tier are separated by spaces, morphemes are separated by hyphens. Since prefixes and suffixes have the same separators and there are no spaces between them and stems, stem boundaries can only be reconstructed from comparison with the cod(ing) tier. Clitics are separated by "=" and correspond to independent words in `//u/w`. They are therefore split off in parsing.

- On the coding tier the same rules apply. Within complex words, the stem is the morpheme to which the POS tag is prefixed (e.g. "n^", "v^").

- Untranscribed words are coded as "xxx" on the word tier and are not morphologically coded.

target table	target field	source tier
session	session_id_fk	file name
utterance	utterance_id	//u@uID
utterance	start_raw	//u/media@start
utterance	end_raw	//u/media@end
utterance	speaker_label	//u@who
utterance	sentence_type	//u/t
utterance	utterance_raw	//u/w
utterance	translation	//u/a[@type="english translation"]
utterance	comment	//u/a[@type="comments"]
word	word	//u//w
morpheme	morpheme	//u/a[@type="target gloss"]
morpheme	gloss_raw	//u/a[@type="extension" and @flavor="cod"]
morpheme	pos_raw	//u/a[@type="extension" and @flavor="cod"]
morpheme	morpheme_language	//u/a[@type="extension" and @flavor="cod"]

Table 4.16: Nungon tiers

4.9 Russian

4.9.1 Publication, accessibility, documentation

The Russian Corpus (Stoll & Meyer 2008) was compiled for the work in Stoll (2001) but was only finished later. The corpus itself has not been published and is cited as:

> Stoll, Sabine & Roland Meyer. 2008. Audio-visual longitudinal corpus on the acquisition of Russian by 5 children.

The corpus is also known as the "Stoll Russian Corpus" (hence the acronym StRuC used in this document). There is no official documentation available.

4.9.2 Recording scheme

number of children	5
age ranges	1;3.26-4;11.0, 1;4.22-5;6.26, 1;6.10-5;4.18, 1;11.28-4;3.14, 3;1.8-6;8.12
recording rhythm	1h every week
recording environment	indoors at home
other speakers	mother and relatives
other languages	none

Table 4.17: Recording scheme for the Russian corpus

4.9.3 File system and formats

The Russian corpus consists of several parallel versions in separate numbered folders. While most of the folders build on each other (for instance, "4a_tbx_lemma_separated_timecodes_lgr" takes over all information from "4_tbx_lemma_separated_timecodes_lgr" but adds glosses modified according to the Leipzig Glossing Rules), some also contain conflicting information (for instance, "6_elan_coded_pointing" contains annotations which are missing from the mentioned folders but at the same time

does not have LGR glosses). As indicated by the folder names, the versions are distinguished by varying formats and annotation layers.

Within that folder, files are named as "code + session number + age", where code is the first letter of the target child code, session number is a three-digit ascending number, and age is the age of the target child given as "YMMDD", e.g. "A03120419".

Most files were encoded as UTF-8 text with some exceptions in ISO-Latin. For the ACQDIV Corpus the original data were all reencoded to UTF-8. The Russian data do not contain any non-ASCII characters.

4.9.4 Corpus format

The input format of the Russian corpus is hybrid Toolbox/CHAT (converted from CHAT by the Russian team). All files contain CHAT-style metadata headers and Toolbox-like bodies with frequent traces of CHAT on the transcription tier and elsewhere. Table Table 4.18 shows how the tiers in the ACQDIV Corpus are related to tiers in the input.

target table	target field	source tier
sessions	session_id_fk	file name
utterances	utterance_id	\ref
utterances	start_raw	\ELANBegin
utterances	end_raw	\ELANEnd
utterances	speaker_label	\EUDICOp
utterances	addressee	\add
utterances	utterance_raw	\text
utterances	sentence_type	utterance delimiter on \text
utterances	comment	\act, \com, \ct, \err, \sit
words	word	\text
morphemes	morpheme	\lem
morphemes	gloss_raw	\mor
morphemes	pos_raw	\mor
morphemes	morpheme_language	\mor, special gloss FOREIGN

Table 4.18: Russian tiers

Several points to be noted concern the morphology tiers in the input:

- There is no segmentation. Instead, words are analyzed on `\mor` using long strings of concatenated glosses. Presently the lemmatization tier `\lem` is interpreted as if it contained segments for the sake of uniformity across the ACQDIV subcorpora.
- The elements on `\mor` are separated by spaces and contain both glosses and POS, which are in turn separated by "-" or ":" according to the following rules:
  - Sub-POS are always separated by "-" (e.g. `PRO-DEM-NOUN`), subglosses are always separated by ":" (e.g. `PST:SG:F`). What varies is the character that separates POS from glosses in the word string.
  - If the POS is `V` ('verb') or `ADJ` ('adjective'), the glosses start behind the first "-", e.g. `V-PST:SG:F:IRREFL:IPFV` → POS `V`, gloss `PST.SG.F.IRREFL.IPFV`.
  - For all other POS, the glosses start behind the first ":", e.g. `PRO-DEM-NOUN:NOM:SG` → POS `PRO.DEM.NOUN`, gloss `NOM.SG`.
  - If there is no ":" in a word string, gloss and POS are identical (most frequently the case with `PCL` 'particle').

Also note that overlaps are regularly transcribed twice in the Russian corpus (once in the interrupted utterance, then once again in a separate record with the right speaker). This could not be corrected in the ACQDIV representation.

4.10 Sesotho

4.10.1 Publication, accessibility, documentation

The Sesotho corpus (Demuth 1992, 2015) was compiled between 1980 and 1990. Citations should mention the corpus and one following paper:

> Demuth, Katherine. Demuth Sesotho Corpus. http://childes.psy.cmu.edu/.
> Demuth, Katherine. 1992. Acquisition of Sesotho. In Dan Slobin (ed.), *The Cross-Linguistic Study of Language Acquisition*, vol. 3, 557-638. Hillsdale, N.J.: Lawrence Erlbaum Associates.

4.10.2 Recording scheme

number of children	4
age ranges	2;1-3;0, 2;1-3;2, 2;4-3;3, 3;8-4;7
recording rhythm	3-4 hours every month
recording environment	home and neighborhood
other speakers	relatives, other children, passers-by
other languages	none

Table 4.19: Recording scheme for the Sesotho corpus

4.10.3 File system and formats

Examples for file names in the published corpus are “hiib” and “tvie”. These names are composed of three elements:

- the first letter of the target child
- an ascending roman number (counting sessions within that child)
- an ascending lowercase letter indicating several recording sessions which come from the same period of intensive recording within a month but may or may not be adjacent. Sessions of this type also correspond to separate media files.

All files are encoded as UTF-8 text and only contain ASCII characters.

4.10.4 Corpus format

The input format of the Sesotho corpus is TalkBank XML (converted from CHAT by the Sesotho team). Table Table 4.20 shows how the tiers in the ACQDIV Corpus are related to tiers in the input.

The morphology tiers in the Sesotho input are structured as follows:

- Words on the target gloss tier are separated by spaces, morphemes are separated by hyphens. Since prefixes and suffixes have the same separators and there are no spaces between them and stems, stem boundaries can only be reconstructed from comparison with the coding tier.

target table	target field	source tier
sessions	session_id_fk	/CHAT@Id
utterances	utterance_id	//u@uID
utterances	start_raw	//u/media@start
utterances	end_raw	//u/media@end
utterances	speaker_label	//u@who
utterances	addressee	-
utterances	sentence_type	//u/t
utterances	utterance_raw	//u/w
utterances	translation	//u/a[@type="translation"]
utterances	comment	//u/a[@type="situation"]
words	word	//u/w
morphemes	morpheme	//u/a[@type="target gloss"]
morphemes	gloss_raw	//u/a[@type="coding"]
morphemes	pos_raw	//u/a[@type="coding"]

Table 4.20: Sesotho tiers

- On the coding tier the same rules apply; however, many glosses (esp. for noun classes) contain brackets within which spaces do not count as word separators. Any orthographic word with only one morpheme is a stem. Within complex words, the stem is the morpheme starting with "n^", "v^", or "id^", or ending with "aj", "nm", or "ps" and a sequence of digits.
- "_" connects two glosses to one, e.g. "come_out", "t^..._v^".
- Brackets after a noun most frequently indicate noun classes. There are always two noun classes (possibly corresponding to singular and plural) separated by a comma, and sometimes several such pairs may appear separated by semicolons. Noun classes are kept with their brackets, but spaces within the brackets are deleted.
- Contracted forms which do not leave any traces at the surface also appear in brackets – these are completely removed with their contents.
- Finally, morphemes may occur in brackets without a documented meaning. In this last case only the brackets are removed and the content is kept.
- Parts of speech are incorporated into the coding tier in various ways. Verbs and ideophones have prefixes "v^" and "id^", respectively. Nouns can be recognized by their noun class brackets. For all other parts of speech the gloss itself is the part of speech and a true gloss reflecting the semantics is missing.
- Noun class prefixes are given in the form "n^" followed by a sequence of digits.
- Proper nouns are given as "n^" followed by "name", "place", "game", or "song".
- Untranscribed words are found as "xxx" on the morphology tier.

4.11 Turkish

4.11.1 Publication, accessibility, documentation

The Turkish corpus (Küntay et al. Unpublished) has not been published. It should be cited as

Küntay, Aylin Copty, Dilara Koçbaş, Süleyman Sabri Taşçı. Unpublished. Koç University Longitudinal Language Development Database on language acquisition of 8 children from 8 to 36 months of age.

There is no official documentation available.

4.11.2 Recording scheme

number of children	8
age ranges	1;0.2-3;0.3, 0;7.28-3;0.24, 0;8.6-3;0.14, 0;8.1-1;9.28, 0;8.0-2;4.20, 0;8.2-3;0.14, 0;8.30-3;0.20, 0;9.27-2;9.13
recording rhythm	1h every 2 weeks
recording environment	indoors at home
other speakers	variety of children and adults
other languages	none

Table 4.21: Recording scheme for the Turkish corpus

4.11.3 File system and formats

File names consist of the code of the target child, an ascending number for counting sessions within that child, the recording date (DDMMMYY), and the age at that time (YY-MM-DD), e.g. "burcu45_10apr04_02-06-20". Files are located in folders named after the target children.

The original Turkish CHAT files come with mixed encodings, most prominently UTF-8 and ISO-Latin, and contain a plethora of unintended special characters. The only special characters that are well-formed by the criteria of the Turkish orthography are ⟨ç⟩, ⟨ğ⟩, ⟨ö⟩, ⟨ş⟩, ⟨ü⟩.

4.11.4 Corpus format

The input format of the Turkish KULLDD corpus is TalkBank XML (converted from CHAT by the ACQDIV core team and the KULLDD team). Table Table 4.22 shows how the tiers in the ACQDIV Corpus are related to tiers in the input.
The following peculiarities should be noted in the Turkish KULLDD input:

- Many constructions are present on the word level but have no corresponding elements on the morphology tier. This concerns words with the formtype attributes "interjection", "onomatopoeia", "family-specific" and the letter construction `<g><w>...</w><ga type="explanation">letter</ga></g>`. Repetitions are mostly glossed, retracings mostly not.

- Replacements may occur in `//u/w` or in `//u/g` with a `<w>` sibling. Shortenings are in `//u/w` or in `//u/g/w`.

The input morphology tier is structured as follows:

- Words are separated by spaces; morphemes are separated by "-".

- There are no prefixes, so the first morpheme is always the stem. The stem is preceded by a POS tag, separated from it by "|". Sub-POS can be given using the colon, e.g. "PRO:DEM|bu" (replaced by period).

target table	target field	source tier
sessions	session_id_fk	/CHAT@Id
utterances	utterance_id	//u@uID
utterances	start_raw	//u/a[@type="time stamp"]
utterances	end_raw	-
utterances	speaker_label	//u@speaker
utterances	addressee	//u/a[@type="addressee"]
utterances	childdirected	//u/a[@type="addressee"]
utterances	sentence_type	//u/t, //u/e
utterances	utterance_raw	//u/w
utterances	translation	//u/a[@type="english translation"]
utterances	comment	//u/a[@type="explanation"], //u/a[@type="actions"]
words	word	//u/w
morphemes	morpheme	//u/a[@type="extension" and @flavor="mor"]
morphemes	gloss_raw	//u/a[@type="extension" and @flavor="mor"]
morphemes	pos_raw	//u/a[@type="extension" and @flavor="mor"]
morphemes	morpheme_language	//u/w/langs

Table 4.22: Turkish tiers

- For lexical elements, only the phonological form of the stem is given. By contrast for grammatical elements, only the function of the morpheme is given. This results in "glosses" such as "V|getir-FUT-1S" (= verb with stem "getir", first suffix = future marker, second suffix = 1st person singular), which in standard interlinearization would be *getir-eceğ-im* for the form and "bring-FUT-1SG" for the gloss. In the ACQDIV Corpus, the unknown form of the suffixes and function of the stem are given as `NULL/NA`.

- Grammatical information contained in the stem and subglosses for suffixes are also indicated by ":" (replaced by period).

- Some difficulties are connected to the use of "+" and "_", both of which indicate mismatches between word boundaries as indicated by orthography and morphology and are completely interchangeable (e.g. *bir şey* is considered a single morphological word (*bir+şey*/*bir_şey*) meaning 'something' but is spelt apart in standard orthography). The corresponding words on the orthographic tier may or may not be joined by an underscore.
  When a complex containing these characters is indeed treated as a single morphological word (i.e. the complex shares a single POS tag and suffix chain), the corresponding orthographic words are joined by "_" (if they aren't already). When a complex is treated as two words (i.e. they have separate POS tags and/or suffixes), the corresponding orthographic words are split (if they aren't already separate).

4.12 Yucatec

4.12.1 Publication, accessibility, documentation

The Yucatec corpus (Pfeiler Unpublished) has not been published. It should be cited as

> Pfeiler, Barbara. Unpublished. Pfeiler Yucatec Child Language Corpus.

There is no official documentation available.

4.12.2 Recording scheme

number of children	3
age ranges	1;11.9-3;5.4, 2;0.1-3;0.29, 2;1.5-3;3.11
recording rhythm	30-90 min every 2 weeks
recording environment	indoors and outdoors at home
other speakers	relatives
other languages	Spanish

Table 4.23: Recording scheme for the Yucatec corpus

4.12.3 File system and formats

There are no principled file naming conventions for the Yucatec corpus, although almost all files include the recording date as "MMDDYY" and the code of target children (full or abbreviated) is an additional frequent element. Files are located in a complex folder structure motivated by target children, recording cycles, and steps in the workflow (transcription, glossing). About one third of all files are doublets or triplets.

The original Yucatec CHAT files are formatted as text (structured as CHAT or unstructured) or doc (MS Word). They have highly heterogeneous encodings and a long list of unintended special characters apparently produced by multiple incomplete reencodings. The only characters that naturally appear in Spanish or Yucatec orthography are vowels with acute accents, ⟨ñ⟩, and ⟨ʼ⟩ (= modifier letter apostrophe, U+02BC).

4.12.4 Corpus format

The input format of the Yucatec corpus is TalkBank XML (converted from CHAT by the ACQDIV core team). Table Table 4.24 shows how the tiers in the ACQDIV Corpus are related to tiers in the input.

target table	target field	source tier
sessions	session_id_fk	/CHAT@Id
utterances	utterance_id	//u@uID
utterances	start_raw	-
utterances	end_raw	-
utterances	speaker_label	//u@who
utterances	addressee	-
utterances	sentence_type	//u/t
utterances	utterance_raw	//u/w
utterances	translation	//u/a[@type="english translation"]
utterance	comment	//u/a[@type="explanation"], //u/a[@type="comments"]
word	word	//u//w
morpheme	morpheme	//u/a[@type="extension" and @flavor="mor"]
morpheme	gloss_raw	//u/a[@type="extension" and @flavor="mor"]
morpheme	pos_raw	//u/a[@type="extension" and @flavor="mor"]

Table 4.24: Yucatec tiers

The Yucatec morphology tier is structured as follows in the input:

- Words are separated by spaces, morphemes by "#" (prefixes) or ":" (suffixes).
- The morpheme tier may also contain the symbols "&" and "+", both of which mark clitics. Since these are most often treated as separate orthographic tiers in `<w>`, these symbols are treated like spaces (i.e. as word separators) in the morphology tier.
- Every morpheme block consists of a gloss and a morpheme form, separated by "|". The gloss of stems is a part of speech rather than a functional label. The form of suffixes is preceded by a redundant "-". An example for a word with both prefixes and suffixes is "3ERG|u#VN|ho'ol:POS|-il" (standard interlinearization: form *u-ho'ol-il*, gloss "3ERG-VN-POS"). In the ACQDIV Corpus, `NULL/NA` is inserted when the function of a stem is not known.
- In many glosses ":" is also used to separate one or several subglosses, e.g. "IMP:ABS:SG". This use can be distinguished from the morpheme-separating use by checking the strings to the left and right of the ":" – when they consist of nothing but uppercase letters and digits, they are subglosses; otherwise they belong to different morphemes.
- Sometimes words do not contain any "#" or ":" but do contain "-". In this case "-" represents a morpheme separator. Words with "-" as the morpheme separator only contain morpheme forms but no glosses.

Chapter 5

Generating the corpus

The ACQDIV Corpus is dynamically generated from the original data described in Chapter 4. A new version is generated every time the original data or their interpretation change.

The original data are extremely heterogeneous and often have greater or smaller internal problems. Therefore, creating a single user-friendly corpus from them requires several processing steps:

- clean files of formal issues that hinder automatic processing (e.g. problematic encodings and file formats)
- ensure compliance with applicable corpus standards as far as necessary
- parse the data and metadata, i.e. read the information contained in them and store it in a temporary unified structure
- build the database and map the unified structure into it
- postprocess the data in the database for greater semantic homogeneity (e.g. with regard to glosses or timestamps)

Note that the first two steps, which also involved manual cleaning, were only carried out once during the initial phase of the project. The remaining steps are fully automatised and are repeated every time the corpus is generated. The following section gives an overview of what happens during all steps in which corpora. It reflects the conceptual rather than the technical functioning of corpus generation. For details on the technical side see Chapter 6 and the documentation that comes with the individual scripts involved in each step.

5.1 Cleaning of file formats

The first cleaning step deals with general issues that make files hard to process automatically. Our goal was for every corpus to have a flat collection of text files that had UTF-8 encoding and the intended character set. In addition, one file should correspond to one recording session (in the sense of a contiguous stretch of time) and vice versa. Both files and their names were required to be unique within one corpus. The sections below describe how this goal was achieved.

5.1.1 Non-textual formats

Most corpora were already formatted as structured text (e.g. XML or Toolbox) at the time the ACQDIV project started working on them. However, there were a few exceptions that were dealt with as follows:

- The Yucatec and to a lesser degree the Turkish corpus contained many doc files. The text was extracted using `doc2txt` and saved with the extension txt.

- All files in the Inuktitut corpus were text but had undocumented file extensions (XXS, XXX, NAC). These were converted to txt.
- Some of the Inuktitut corpus documentation had Word Perfect formats (REP, SUB, IKT). These were converted to pdf.

5.1.2 Encodings

All corpora were required to be encoded in UTF-8. This was not the case for most Inuktitut and Yucatec as well as for some Russian and Turkish files, where ISO-Latin and ASCII were found among other, less common encodings. Encodings were determined using the Python library `chardet` and converted to UTF-8.

5.1.3 Character sets

The same three corpora (Inuktitut, Turkish, Yucatec) also had problems with unintended special characters such as letters with accents, letters from foreign alphabets, and non-textual characters such as suns or alien heads. Problems of this kind were least prominent in Inuktitut, whose orthography does not feature any special characters, but very widespread in Turkish (special characters ⟨ç⟩, ⟨ğ⟩, ⟨ö⟩, ⟨ş⟩, ⟨ü⟩) and Yucatec (vowels with acute accents, special characters ⟨ñ⟩, ⟨ʼ⟩ (= modifier letter apostrophe, U+02BC)). Character lists were automatically extracted from all files of these corpora, replacement lists were compiled for all corrupted characters, and automatic replacements were made wherever such a corrupted character uniformly corresponded to a well-formed character.

5.1.4 Folder systems and file names

Many corpora initially had deeply nested folder systems which often obscured the actual structure of the corpus or made it possible for a corpus to contain two or more files with the same name. The following steps were undertaken to overcome these problems:

- Whenever a corpus was available in different formats (e.g. CHAT vs. TalkBank XML), only the strictly required formats were taken over.
- The Indonesian corpus originally had several subfolders containing subcorpora for each target child and named after their code. Within the folders, Toolbox files were originally named as "COD-DDMMYY" (where "COD" represents the speaker code of the target child) and XML files were more simply named as "YYYY-MM-DD" with no indication of the target child. Both formats were unified to "COD-YYYY-MM-DD" in the input data for the ACQDIV Corpus to achieve consistency and unique session names across all folders. All files were put on the same level.
- In the Inuktitut corpus, recording sessions often corresponded to several files within one subfolder. All such files were fused to a single file, keeping the shared string in the beginning of the file name and replacing everything else by "All" (e.g. "JUP21All" instead of "JUP21ATF", "JUP21BTF", "JUP21CTF" etc.). The merged files were put on the same level as the pre-existing other files, thus creating a flat structure.
- The Japanese MiiPro files were originally located in folders named after the target children but were all put on the same level for input in the ACQDIV Corpus.
- The Japanese Miyata subcorpus is a particularly complicated case. Every session is represented at least twice and maximally four times by files with largely identical contents but different file names – see the description of the original data for details. Only the most recent series of files was used for each child (Aki 3, Ryo 4, Tai 4). Since these series did not indicate

the name of the target child in the file name and were thus potentially ambiguous, the codes were prefixed to the file name with an underscore ("aki_34_20629") before putting all files on the same level.

- The Russian corpus consists of several parallel versions with different annotations in separate folders. For the ACQDIV Corpus the folder was used that contained most of all recent annotations and glosses based on the Leipzig Glossing Rules ("4a_tbx_lemma_separated_timecodes_lgr").

- The Turkish corpus contained subfolders for target children. This structure was flattened as in the other corpora.

- Yucatec is another corpus with many doublets and triplets, which in this case becomes possible by a complicated folder structure and competing naming conventions (see the description of the original data for details). Doublets were detected by checking all file names and especially the string of digits contained in them, which turned out to be most indicative of session identity. In the next step, the most recent version in every doublet set was determined based on file size and annotation layers. Older versions were discarded and all files were renamed according to the scheme "COD-YYYY-MM-DD" and put on a single level. Where all versions represented the same level of analysis the version kept was the one which was easiest to process (e.g. because of encodings).

5.2 Cleaning of corpus formats

The two corpus formats that were accepted as input for the ACQDIV Corpus are TalkBank XML and Toolbox. CHAT was deliberately excluded from this list because a good CHAT to XML parser is available with Chatter. However, three corpora – Inuktitut, Turkish and Yucatec – were delivered in broken CHAT that could initially not be parsed by Chatter. Given the choice to either write a new CHAT parser or clean up the CHAT in order to make it convertible, we decided that it would be both easier and more sustainable to go for the latter option.

One of the most frequent parsing problems were broken headers. The header of a CHAT file is an obligatory section at the head of the file that lists all session-level and speaker-level metadata associated with it. Some examples for problems with headers are to missing information, corrupted tier names, or whitespace characters in the wrong places. Cleaning headers required the following steps:

- Specify a set of non-discardable metadata. For the session level these were the recording date (CHAT tier `@Date:`), the recording situation (`@Situation:`), and the name of the associated media file (`@Media`). For the speaker level the non-discardable data were code, name, age, sex, and spoken languages (all coded on the CHAT tier `@ID:`) as well as role (`@Participants:`).

- For each corpus create a table containing all existing information from all metadata tiers. Our collaborators were requested to go through these tables fill in any gaps in the data. In addition, tiers that had been found in the corpus but did not form part of the above-mentioned list were to be deleted by default. The collaborators were asked to go through these tiers and to transfer any contents that they wanted to be retained to another tier from the standardized set. For instance, one frequent pseudo-tier that was not accepted by Chatter was `@Age of CHI:`. The contents of this tier could easily be transferred to a modified or newly created `@ID:` tier for the target child.

- Then clean the finished tables of any remaining clutter and convert them to two simple CSV files per corpus: `ids.csv` for speaker-level metadata and `sessions.csv` for session-level

metadata. These files were then used by the pyacqdiv package (cf. Chapter 6) to create a new metadata header for every file in each corpus. All pre-existing information that had not been captured in the metadata tables was overwritten in this process.

The body data presented different problems and were therefore dealt with separately. While these did not contain any non-systematic gaps, there were many more formatting problems, many of them having to do with whitespace characters and special characters, which are only allowed in specific places in CHAT. These problems were dealt with in the following way:

- Chatter was systematically run over the data. Error message produced by the parser were collected and sorted by frequency in order to be able to deal with the most frequent problems first.
- For all types of problems with a token frequency roughly higher than 50, replacement rules were collected in a file called `code.py` (one per corpus). This served again as input for the pyacqdiv package, which applied the replacement rules to all files in each corpus.
- All problems with lower frequencies were corrected manually, either by the ACQDIV team or by the collaborators responsible for the particular subcorpus. This was mainly done in CLAN, although certain differences between the standards applied in CLAN and Chatter sometimes brought to light additional problems when attempting to parse corrected files in Chatter.

Although all corpora apart from the ones mentioned above superficially did not contain problematic formats, most of them do on a deeper level. The reason for this is always that the corpora were converted from CHAT and still contain traces of it. This applies to the following cases:

- Japanese MiiPro and Sesotho are available as TalkBank XML. However, the morphology tier has no explicit internal structure but has been directly taken over from CHAT as a single string. This CHAT pocket has to be parsed differently from the rest of the files (see Section 5.3 below for details).
- Similarly, Inuktitut, Turkish and Yucatec had morphology tiers with rather different conventions (see the description of corpus-specific conventions in Chapter 4 for details). These differences seem to reflect changes in the CHAT standard during time. Because of this and also because the morphology tier is by far the most complex CHAT tier with the most possibilities for formal mistakes, we decided not to clean it but to transfer it directly to XML and deal with it in the parser, as for Japanese MiiPro and Sesotho.
- The Indonesian Toolbox files contain metadata in CHAT format that have been superficially "masked" as Toolbox (see again Chapter 4 for details).
- Russian contains CHAT codes on some of its Toolbox tiers, especially on the main transcription tier.

All of these problems do not affect the convertibility of files, so they were not dealt with in the cleaning block but directly in the parser.

5.3 Parsing the corpus data

Parsing in a narrow sense concerns the process that transforms one of the accepted input formats (TalkBank XML, Toolbox) into an output format (e.g. a SQLite database, cf. the specifications). This is the most complex process in the corpus pipeline and can only be roughly sketched here. For more details refer to the documentation of the relevant scripts.

5.3.1 TalkBank XML

Parsing a TalkBank XML file involves the following conceptual steps:

- The XML tree is read using a Python library such as ElementTree.
- The content of unproblematic tiers (e.g. translations, timestamps) is transferred as a whole to the relevant target field.
- The problematic tiers are on the word level (represented by `<w>`) and the morpheme level (segmentation, glosses, and POS, mostly on a single tier, sometimes spread over several). These have to be split (especially the morpheme tier, which has no explicit internal structure) and aligned with each other (so that one `<w>` word corresponds to one morphological word and each segment has matching glosses and POS tags).
- The contents of `<w>` are cleaned of any remnants of CHAT or other idiosyncratic conventions. The contrast between actual and target form is built based on the constructs described in Section 4.1.2. For instance, the construct `<g><w><shortening>com</shortening>puter</w> </g>` would give rise to an actual word *puter* and a target word *computer*.
- The same constructs also influence the interpretation of the morphology tier. For instance, fragments are not glossed in many corpora, i.e. there is no element on the morphology tier that corresponds to the word coded by `<w>`. After encountering such a word in `<w>`, the parser adjusts the indices for aligning morphological words.
- The morphology tier(s) are cleaned of small-scale inconsistencies and components which are redundant or not compatible with the other corpora. Morpheme-internal spaces are removed. They are then split into words and morphemes.
- Some corpora occasionally specify distinct actual and target forms for words or morphemes on the morphology tier(s). These are separated and stored before all following actions.
- Segments, glosses, and POS are identified based on the corpus-specific formalism. In some cases the formal coding is in contrast to the real function of the coded element. For instance, labels may sometimes be formally marked as glosses but rather code POS. In this case contents are transferred from one category to another as far as possible.
- The words in `<w>` are now aligned with the morphological words, taking into account what is known about missing glosses.

5.3.2 Toolbox

Toolbox files are somewhat simpler to parse because they have a flatter structure than XML, thus providing less opportunities for drastic mismatches between coding syntax and semantics. Parsing a Toolbox XML file involves the following conceptual steps:

- Files are split into records based on linebreak characters. Each record contains several lines, which at the same time correspond to its tiers.
- All tiers are cleaned of remnants of CHAT or other idiosyncratic conventions.
- The content of unproblematic tiers (e.g. translations, timestamps) is transferred as a whole to the relevant target field.
- Tiers coding words are split into words by spaces.

- Tiers coding morphemes are split into morphemes by spaces. The boundaries of morphological words then have to be reconstructed based on morpheme separators. For instance, given the string *play -ing with* the parser can infer that there is a word boundary between *-ing* and *with* because a suffix cannot be followed by a stem (at least given the definition of these terms in Toolbox). Identifying segments, glosses, and POS is trivial because these are given on separate tiers.
- The last step concerns alignment. Orthographic words are aligned with morphological words, and for every morpheme-level element the parser checks if there are corresponding elements on all three tiers (segments, glosses, POS). Note that alignment in Toolbox is always based on corresponding indices (i.e. the first element of set 1 corresponds to the first element of set 2, etc.).

5.3.3 Intermediate storage

The parsed data are stored in a nested Python structure before inserting them into the database. This structure can be schematically depicted as follows:

```
{
   "corpus": {
      "session": {
         "utterance": {
            [
               ["key", "value"],
               ["key", "value"],
               ["key", "value"]
            ]
         }
      }
   }
}
```

where the `["key", "value"]` tuples give all information contained in one tier irrespective of its level (utterance, word, morpheme). Multiple words or morphemes are passed on as a single string (e.g. `[["morpheme", "Bäum e"], ["gloss", "tree PL"], ["pos", "n sfx"]]`).

An earlier version of the parser used JSON for intermediate storage. The main difference to the new structure is that there used to be additional nested levels below the utterance level. One full example for a record mapped to JSON is given below for legacy reasons.

```
{
   "session_name": [
      {
         "utterance_id": "child1-session17.386",
         "speaker_id": "CHI1",
         "addressee": "MOT2",
         "starts_at": "00:12:53",
         "ends_at": "00:12:55",
         "orthographic": "Atukunai no?",
         "phonetic": "atsunaino:",
```

```
        "phonetic_target": "atsukunai no",
        "english": "Isn't it cold?",
        "nepali": "Garmi chaina?",
        "spanish": "No es caliente?",
        "sentence_type": "question",
        "comments": "Mother has gone out, CHI1 talking to herself.",
        "warnings": "English translation might be wrong",
        "words": [
           {
              "full_word": "atsunai",
              "full_word_target": "atsukunai",
              "warnings": "transcription might be wrong",
              "morphemes": [
                 {
                    "segments": "atsu",
                    "glosses": "hot",
                    "pos": "adj"
                 },
                 {
                    "segments": "na",
                    "segments_target": "kuna",
                    "glosses": "NEG",
                    "glosses_target": "NEG",
                    "pos": "sfx"
                    "pos_target": "sfx"
                 },
                 {
                    "segments": "i",
                    "glosses": "NPST",
                    "pos": "sfx"
                 }
              ]
           },
           {
              "full_word": "no"
              "full_word_target": "no"
              "morphemes": [
                 {
                    "segments": "no",
                    "glosses": "Q",
                    "pos": "ptcl",
                 }
              ]
           }
        ]
     }
  ]
}
```

5.4 Parsing the metadata

Metadata are often stored in separate files and/or structured differently from corpus data. They are therefore parsed separately from the latter.

Metadata are either read from TalkBank XML files, where data and metadata for a session are stored in a single file, or from IMDI XML, a generalized metadata standard that is commonly used in combination with Toolbox data. A special case is presented by Indonesian, where the latest data are formatted as Toolbox but the metadata are still in TalkBank XML, reflecting the origin of the corpus in CHAT.

Compared to the data, the metadata are relatively easy to parse because their syntax is more robust (and thus much less frequently broken), string operations such as replacements, deletions or splits are not necessary, and elements of fields do not have to be connected or moved (in contrast to the multiple alignments described above for the data). The data are therefore simply read and transferred to the relevant target fields in the database. Note that the metadata tables `sessions` and `speakers` are generated independently of the corpus data tables and can only be relinked to them via the relevant keys (`corpus`, `session_id`, `speaker_label`).

5.5 Building the database and postprocessing

Once the corpus data and metadata have been parsed as described above, a database skeleton with the structure described in Section 3.4 is built and the internal data structures are mapped to it. Numeric IDs are automatically generated for all records.

After the tables have been filled, any edits and additions that are best performed over all data simultaneously are carried out. This is referred to as postprocessing. Postprocessing adds the following tables and columns (also cf. again Section 3.4):

- Glosses (`morphemes.gloss_raw`) and parts of speech (`morphemes.pos_raw`) are unified across corpora based on the lists given in Section 3.5.3 and Section 3.5.4. The modified labels are written to `morphemes.gloss` and `morphemes.pos`/`words.pos_ud`, respectively.
- `utterances.utterance_raw` is cleaned of any remaining punctuation and other special markers (cf. the transcription conventions); the new content is put into `utterances.utterance`.
- In the `utterances` table, `start_raw` and `end_raw` are unified to the format HH:MM:SS.
- The `language` field is set in various tables tables based on the `corpus` field.
- The contents of `speakers.age_raw` are converted to the language acquisition format YY;MM.DD in the `speakers.age` field. An additional representation in days is inserted into `speakers.age_in_days`.
- The gender values found in `speakers.gender_raw` are standardized to the two values `Female` and `Male` in the column `speakers.gender`. Where the gender is not known the postprocessor tries to infer it from role values such as `Mother`.
- The many values of `speakers.role_raw` are greatly simplified based on the controlled vocabulary given in Section 3.5.2.
- Based on age, standardized role, and ID, a new column `speakers.macrorole` is inserted, which only allows the three values `Child`, `Target_child`, `Adult`.
- The complete `uniquespeakers` table is generated based on inference from the name, speaker label, and birthdate of speakers as occurring in sessions.

Postprocessing also checks the tables `utterances`, `words`, and `morphemes` for completeness and consistency, editing the field `warnings` in each of the tables in the following way:

table	warning	explanation
utterances	insecure transcription	the transcription for the complete utterance is insecure, e.g. because the utterance is hard to understand
utterances	not glossed	the complete utterance has not been glossed for some reason
utterances	broken word alignment	the number of words coded on the main transcription tier is bigger or smaller than the one coded on the morphology tier so that it's not clear which words are associated
words	broken morpheme alignment	the number of morphemes coded on one of the logical morphology tiers (morphemes, glosses, parts of speech) is bigger or smaller than the one coded on another of these tiers so it's not clear which annotations are associated
words	contains unstructured morpheme	the marked word seems to contain morpheme boundaries but can't be split because the formatting is broken
morphemes	annotation insecure for tier X	the marked morpheme, gloss, or part-of-speech tag is insecure

Table 5.1: Overview of warnings

Chapter 6

Information for developers

The ACQDIV Corpus has been cleaned and is generated with Python 3. For cleaning, a Python package called `pyacqdiv` has been created that is currently still under development and might incorporate database generation in the future. The database is generated by a suite of scripts, the most important ones being the loader, which reads corpus-specific information from ini files, calls the matching data and metadata parsers, and creates the database, and the postprocessor, which checks the contents of database tables and applies rules for normalization. A testing mechanism based on gold standards ensures that changes in the generating and processing scripts do not alter the structure of the database in undesirable ways.

This architecture is the result of team work that has been made possible with the help of GitHub. Currently the GitHub repository at https://github.com/uzling/acqdiv is closed to the public. Developers wishing to get access should contact Steven Moran (tech lead).

Bibliography

Acton, Sara. 2013. CCLAS Auto-Parser Guide. Unpublished manuscript.

Allen, Shanley. Unpublished. Allen Inuktitut Child Language Corpus.

Allen, Shanley E. M. 1996. *Aspects of argument structure acquisition in Inuktitut*. Amsterdam: Benjamins.

Brittain, Julie. 2015. Corpus of the Chisasibi Child Language Acquisition Study (CCLAS). http://phonbank.talkbank.org/access/Other/Cree/CCLAS.html.

Demuth, Katherine. 2015. Demuth Sesotho Corpus. http://childes.talkbank.org/access/Other/Sesotho/Demuth.html.

Demuth, Katherine A. 1992. Acquisition of Sesotho. In Dan Isaac Slobin (ed.), *The crosslinguistic study of language acquisition*, vol. 3, 557–638. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gil, David & Uri Tadmor. 2007. The MPI-EVA Jakarta Child Language Database. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University. https://jakarta.shh.mpg.de/acquisition.php.

Küntay, Aylin C., Dilara Koçbaş & Süleyman Sabri Taşçı. Unpublished. Koç University Longitudinal Language Development Database on language acquisition of 8 children from 8 to 36 months of age.

MacWhinney, Brian. 2000. *The CHILDES project: tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Miyata, Susanne. 2004a. *Aki Corpus*. Pittsburgh, PA: Talkbank.

Miyata, Susanne. 2004b. *Ryo Corpus*. Pittsburgh, PA: Talkbank.

Miyata, Susanne. 2004c. *Tai Corpus*. Pittsburgh, PA: Talkbank.

Miyata, Susanne. 2012. Japanese CHILDES: The 2012 CHILDES manual for Japanese. http://www2.aasa.ac.jp/people/smiyata/CHILDESmanual/chapter01.html.

Miyata, Susanne & Hiro Yuki Nisisawa. 2009. *MiiPro - Asato Corpus*. Pittsburgh, PA: Talkbank.

Miyata, Susanne & Hiro Yuki Nisisawa. 2010. *MiiPro - Tomito Corpus*. Pittsburgh, PA: Talkbank.

Moran, Steven, Robert Schikowski, Danica Pajović, Cazim Hysi & Sabine Stoll. 2016. The ACQDIV Database: Min(d)ing the Ambient Language. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/pdf/1198_Paper.pdf.

Nisisawa, Hiro Yuki & Susanne Miyata. 2009. *MiiPro - Nanami Corpus*. Pittsburgh, PA: Talkbank.

Nisisawa, Hiro Yuki & Susanne Miyata. 2010. *MiiPro - ArikaM Corpus*. Pittsburgh, PA: Talkbank.

Pfeiler, Barbara. Unpublished. Pfeiler Yucatec Child Language Corpus.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org.

Sarvasy, Hannah. 2017a. *A Grammar of Nungon: A Papuan Language of Northeast New Guinea*. Leiden: Brill.

Sarvasy, Hannah. 2017b. Sarvasy Nungon Corpus. http://childes.talkbank.org/access/Other/Nungon/Sarvasy.html.

Schikowski, Robert. 2015. Conventions for the linguistic analysis of Chintang. http://spwarran.uzh.ch/chintangwiki/index.php/Conventions_for_the_linguistic_analysis_of_Chintang.

Stoll, Sabine. 2001. *The acquisition of Russian aspect*. UMI Publications.

Stoll, Sabine & Balthasar Bickel. 2013. Capturing diversity in language acquisition research. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency: studies in honor of Johanna Nichols*, 195–260. Amsterdam: Benjamins. [pre-print available at http://www.psycholinguistics.uzh.ch/stoll/publications/stollbickel.sampling2012rev.pdf].

Stoll, Sabine, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Manoj Rai, Novel Kishor Rai, Ichchha P. Rai, Taras Zakharko, Robert Schikowski & Balthasar Bickel. 2015. Audiovisual corpus on the acquisition of Chintang by six children.

Stoll, Sabine & Roland Meyer. 2008. Audio-visional longitudinal corpus on the acquisition of Russian by 5 children.