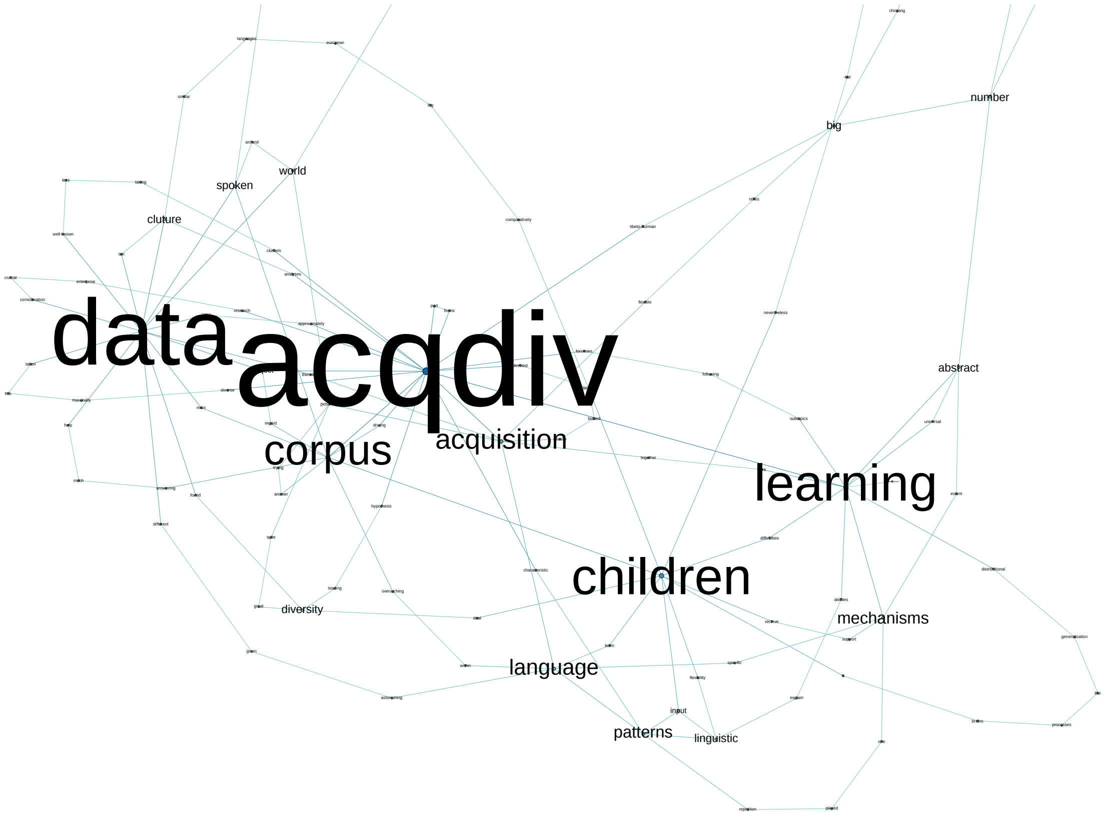


ACQDIV Corpus Database User Manual

Steven Moran, Robert Schikowski, and Sabine Stoll
December 20, 2019

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 Database". 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 Database, and Chapter 4 describes the format and content of the corpus in greater detail. Since the corpus is dynamically generated from several subcorpora,[1] the following Chapter 5 describes the original data and how they are recast into the target structure. Readers with a technical interest may consult Chapter 6 to learn more about the individual steps involved in this procedure. Finally, ?? 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 4 starts with the conditions of access and extension of the corpus in Section 4.1. 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 4.4 and lists of standardized values (e.g. for glosses and parts of speech) in Section 4.5.

Chapter 5 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 Database.

Chapter 6 deals with the steps involved in building the ACQDIV Corpus Database 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.

[1] Throughout this document we refer to the various child language acquisition corpora – that are stand-alone entities in their own right – as *subcorpora* of the ACQDIV Corpus Database.

?? 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 Database 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
- **Steven Moran** designed and built the IT infrastructure for the database and was responsible for its programming implementation
- **Robert Schikowski** oversaw the data analysis, devised the cross-linguistic label scheme, and wrote much of the documentation
- **Anna Jancso** wrote the CHAT parser and refactored the pipeline for public access
- **Cazim Hysi** wrote the metadata parsers and refactored the data parsers
- **Danica Pajović** helped to clean the corpora and to write the parsers
- **Janis Goldzycher** was a programmer on the project

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
- **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

This project would not have been possible without the data provided by our external collaborators:

- **Shanley Allen** for Inuktitut
- **Julie Brittain** and **Yvan Rose** for Cree
- **Katherine Demuth** for Sesotho
- **Gaby Hermon** for Indonesian
- **Aylin Küntay** for Turkish
- **Barbara Pfeiler** for Yucatec
- **Hannah Sarvasy** for Nungon
- **Alan Rumsey** for Ku Waru
- **Birgit Hellwig** for Qaqet
- **Geraldine Walther** and **Jekaterina Mažara** for Tuatschin

Other researchers were also involved in the creation of the original corpora. See Chapter 5 (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 Database is a longitudinal language acquisition corpus that currently features more than ten diverse languages. The languages and the eleven corpora by which they are represented are shown below.

Language	ISO 639-2	Corpora	Acronym
Chintang	ctn	Chintang Language Corpus (Language Acquisition subcorpus)	CLC
Cree	crl	Corpus of the Chisasibi Child Language Acquisition Study	CCLAS
English Manchester	eng	English Manchester Corpus	EMC
Indonesian	ind	MPI-EVA Jakarta Child Language Database	JCLD
Inuktitut	iku	Allen Inuktitut Child Language Corpus	AIC
Japanese	jpn	MiiPro Japanese Corpus	MPJC
Japanese	jpn	Miyata Japanese Corpus	MYJC
Ku Waru	mux	Ku Waru Child Language Socialization Study	KWCLSS
Nungon	yuw	Sarvasy Nungon Corpus	SNC
Qaqet	byx	Qaqet Child Language Documentation	QCLD
Russian	rus	Stoll Russian Corpus	StRuC
Sesotho	sot	Demuth Sesotho Corpus	DSC
Tuatschin	roh	Tuatschin Corpus	TC
Turkish	tur	Koç University Longitudinal Language Development Database	KULLDD
Yucatec	yua	Pfeiler Yucatec Child Language Corpus	PYC

Table 2.1: ACQDIV corpora

Corpus	ISO 639-3	Glottocode	# Sessions	# Words	# Morphemes	Status
Chintang	ctn	chhi1245	477	987673	1589827	definitely endangered
Cree	cre	cree1272	25	44751	11686	vulnerable
English_Manchester1	eng	stan1293	804	2016043	2098914	safe
Indonesian	ind	indo1316	997	2489329	2725605	safe
Inuktitut	ike	east2534	77	71191	91685	vulnerable
Japanese_MiiPro	jpn	nucl1643	192	1011670	1009599	safe
Japanese_Miyata	jpn	nucl1643	213	373021	372495	safe
Ku_Waru	mux	boun1245	9	65723	92438	safe
Nungon	yuw	yaum1237	4	19659	19262	safe
Qaqet	byx	qaqe1238	106	56239	105165	definitely endangered
Russian	rus	russ1263	450	2029704	1	safe
Sesotho	sot	seso1234	69	177963	330009	safe
Tuatschin	roh	roma1326	51	118310	1	vulnerable
Turkish	tur	nucl1301	373	1120077	215822	safe
Yucatec	yua	yuca1254	234	262382	171633	safe

Table 2.2: 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, including:

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

Next, we give some examples 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)

Ku Waru verbs agree with their subject (S/A argument), but mark TAM only if they appear in final position within an independent clause or 'clause chain', as described in Section 5.9. In (2a), the two non-final (NF) verbs are not marked for TAM and are underspecified for person and number, but their NF marking shows that their subject is the same as that of the final verb. In clause chains where the subject of a non-final verb differs from that of the following verb, this is indicated by a 'switch reference' form. Ku Waru has two series of switch reference forms, which are morphologically identical with corresponding optative and subjunctive forms, but take on a switch-reference meaning when they are used non-finally, as in (2b).

(2) a. *pu-k li-k me-k o-a*
go-NF.2/3 get-NF.2/3 carry-NF.2/3 come-IMP.2SG
'Go and bring it over here.' (KWCLSS, 20130815 line 120)

b. *to-lkumela paul te-kum=al*
hit-SBJV:2/3PL wrong do-PROG:3SG=DEF2
'They hit you and looks like you are paining.' (KWCLSS, 20140218 line 2388)

Sesotho (3a) 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 (3b) or the P as in (3c).

(3) 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 (4a) is an example of a language with a fairly low degree of synthesis, whereas Cree (4b) belongs to one of the most genuinely polysynthetic languages of the world, featuring noun incorporation and polypartite stems:

(4) 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 (5a). Another common word order, SOV, is found in Turkish (5b). Yucatec features (among other orders) the much less common VOS (5c).

(5) 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 (6a) and (6b) 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 (6c).

(6) 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-nɨ-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 (7a) and (8a), the Chintang ergative also serves (among others) to mark causes (7b), whereas the Inuktitut ergative is also (again among others) used as a genitive (8b):

(7) 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 meʔ-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)

(8) 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)

In Qaqet, verbs have up to four different aspectual stems that usually (but not always) differ in their initial consonant, e.g., non-continuous *rek* vs. continuous *tek* 'hold/put'. The language has a quite complex noun class system (marked on the noun and on dependent elements), having 8 classes and 3 numbers. It also shows a considerable amount of cliticization, resulting in long phonological words, at least in adult-directed speech.

Most distinctly, Qaqet has two possibilities for expressing arguments entailed by the verb: an argument can either be formally unmarked (as *giagel* 'another one of your cut ones' in (9a) or else introduced by a preposition as *ngua* 'me' in (9b) and *nget* 'it' in (9c). In fact, many verbs are semantically general and allow for more than one combination with concomitant changes in meaning.

For example, the verb root *rek tek* 'hold/put' is attested in different types of events, including (but not restricted to) events of 'putting up, erecting' (with an unmarked argument, as in (9a), 'holding' (with an argument introduced by the preposition *pet* 'on/under', as in (9b) or 'pouring' (with an argument introduced by the preposition *ne* 'from/with', as in (9c).

While the verb itself does not distinguish between event types, its co-occurrence with prepositions unambiguously determines the type. Many such verb-preposition combinations have lexicalized, resulting in complex verbs consisting of a verb root and a verb particle or suffix, e.g., *rekmet tekmet* 'do/act' is composed of *rek tek* 'hold/put' plus the preposition met 'in'. The combining elements can often (but not necessarily) be identified and their contribution to the overall meaning can be sketched out, i.e., there is some degree of compositionality and semantic transparency. Nevertheless, the resulting meanings are never fully compositional, exhibiting idiosyncratic meaning changes; and there are unpredictable limits to the productivity of any given combination.

Such complex verbs can be compared to the prefix and particle verbs in better described West Germanic languages (e.g., English *break in*, *break out* or *break down* or their German equivalents *einbrechen*, *ausbrechen* or *zusammenbrechen*).

(9) a. *nyirek giagel paapit*
nyi=**rek** gi-ia-gel pe=pit

2sg.sbj.npst=hold/put.ncont 2sg.poss-other-sg.exc place=up
'put another one of your cut ones up there' (LongZDL20160104_1 712)

b. *mama, gimga qatek prangua*
mama, gi-uim-ka ka=**tek** **pet**-ngua
mama 2sg.poss-child-sg.m 3sg.m.sbj=hold/put.cont on/under-1sg
'mama, your child is holding me' (LongZDL20160304_1 291)

c. *saqi uantek nanget*
saqi uan=**tek** **ne**-nget
again 2du.sbj-hold/put.cont from/with-3n
'pour it out again' (LongYDS20150516_1 107)

Tuatschin, as the other Sursilvan varieties, displays four forms of the adjective, which are used in the following way, with *alv* 'white' as an example in **??**

Agreement	Form	Representation
m.sg.	attributive	*alv*
m.sg.	predicative	*alvs*
m.pl.	attributive and predicative	*alvs*
f.sg.	attributive and predicative	*alva*
f.pl.	attributive and predicative	*alvas*

Table 2.3: Adjective forms in Tuatschin

The adjective agrees with the noun as described:

(10) a. *Quaj péz è* ***alvs****.*
dem peak cop.prs.3sg white m.sg.pred
'This mountain peak is white.'

b. *Quèls pézs èn* ***alvs****.*
dem peak cop.prs.3pl white.m.pl.pred
'These mountain peaks are white.'

c. *Quaj è in péz* ***alv****.*
dem cop.prs.3sg indef.art.m.sg peak white.m.sg.attr
'This is a white mountain peak.'

The masculine singular attributive form has a further function: it is used with subjects that are neither masculine nor feminine, as for instance the demonstrative *quaj* 'this':

Quaj è **alv**. 'This is white.'

This form of the adjective is not simply the predicative form which loses its -s (alvs ⟶ alv), but a genuine form, which is best shown by the adjectives with stem alterations, as the adjective for 'good', which has the forms *bian* M.SG.ATTR., *buns* M.SG and PL.PRED, *buna* F.SG., and *bunas* F.PL., both attributive and predicative, shown in **??**.

If the adjective precedes the noun without forming a constituent, there is no agreement and the masculine attributive form is used.

Tuatschin	English
Quaj cùdasch è **buns**.	'This book is good.'
Quaj è in **bian** cùdesch.	'This is a good book.'
Quaj è **bian**.	'This is good.'
Mazá è bégja **bian**.	'To kill is not good.'

Table 2.4: Genuine adjective forms in Tuatschin

(11) a. *Parquaj* *ṣè* ***impurtònt la gramática*** *tga* *té* *fas.*
therefore cop.prs.3sg important.m.attr the.f grammar rel you.sg do.prs.2sg
'Therefore the grammar you write is important.'

2.2 Amount of data

2.2.1 Total number of utterances, words, and morphemes

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

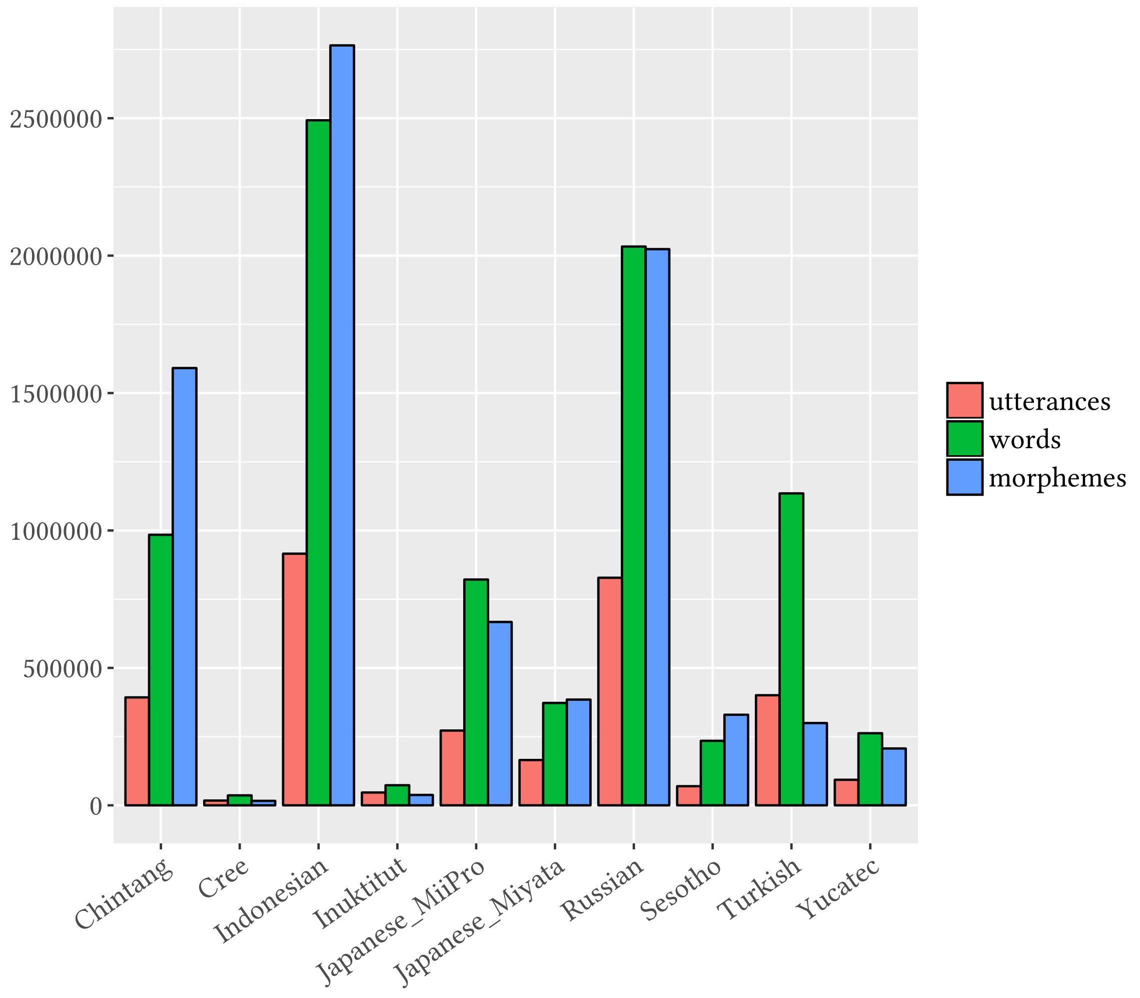


Figure 2.1: Amount of data in the ACQDIV subcorpora

2.2.2 Total number of words per speaker role per corpus

Table 2.5 shows the number of words and their percentages of the corpus per speaker macrorole (null transcribed words have been removed). There are four values macroroles: "Target_Child", "Adult", "Child" (i.e. any individual in the recording session that is not the target child and under the age of 12), and "Unknown" (see Section 4.5.2 for more details).

Corpus	Words	Target_Child	Adult	Children	Total
Chintang	987668	161630 0.17	459186 0.47	329746 0.34	950562 0.97
Cree	65447	25540 0.4	39453 0.61	425 0.01	65418 1
Indonesian	2386454	761822 0.32	1210636 0.507294924	301611 0.13	2274069 0.96
Inuktitut	70295	33357 0.48	22453 0.32	14385 0.21	70195 1
Japanese MiiPro	810899	287606 0.36	509903 0.63	1282 0.01	798791 0.99
Japanese Miyata	368751	136689 0.38	231721 0.63	0 0	368410 1
Russian	2012645	580971 0.29	1312118 0.66	80132 0.04	1973221 0.99
Sesotho	234559	85078 0.37	82899 0.36	66579 0.29	234556 1
Turkish	1085814	164319 0.16	915754 0.85	4976 0.01	1085049 1
Yucatec	253572	122520 0.49	90338 0.36	39771 0.16	252629 1

Table 2.5: Amount of words per speaker role in each corpus

2.3 Sampling for speakers and periods

The ACQDIV Corpus Database 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 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 5.

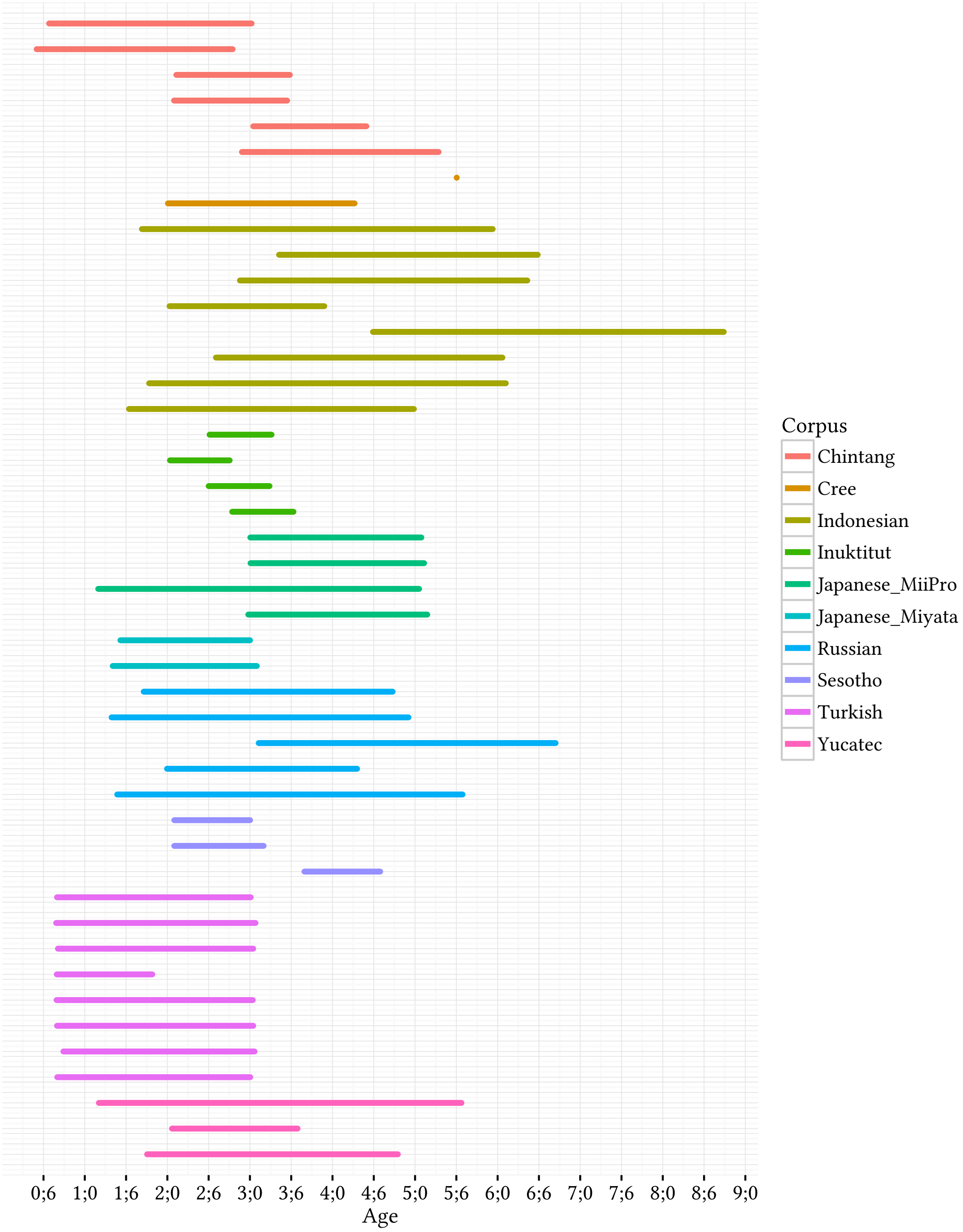


Figure 2.2: Children and recording periods in the ACQDIV Corpus Database

Chapter 3

Annotation layers, data gaps, and things to watch out for (!)

The ACQDIV Corpus Database 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 Database; 14% in a parallel set of ELAN files which is currently not part of the ACQDIV Corpus Database). 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 their 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`.

A few more critical points of note when using the ACQDIV Corpus Database for research purposes:

- When using the `all_data` view, it is critical to understand that many of the `morpheme` and `word` fields are `NULL`. This is due to a decision by project leaders to keep both the word and morpheme tiers **even when they do not align correctly in the input data**. This allows users to use the word, morpheme, gloss, etc., tiers independently without throwing out all data that does not align. Unfortunately, the input corpora may contain many misalignments.

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

- File names in some corpora (including at least Japanese MiiPro and Sesotho) are not unique. We use file names as the `source_id` in the `sessions` table, so in the corpora that do not use unique file names across different recording sessions (e.g. same file names for different target children, which reside in different file folders), we concatenate the CHILDES folder to the filename, e.g. "HIZ" in Indonesian "HIZ-1999-05-20".

- Please refer to Table 4.12 for the problematic mappings of POS-labels from the input corpora to the ACQDIV Corpus Database and Universal Dependency tags. See also Section 4.5.5 for additional remarks on the differences between POS tag sets.

- 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**.

- Please refer to Section 6.1 regarding issues of data cleaning and input file formats.

- Chintang sometimes has outdated language codes in `speakers.languages_spoken`, e.g. x-sil-BAP.

- Some session durations could not be identified due to missing (or no, e.g. English Manchester) media files, corrupt data formats, mismatching media IDs and filenames, etc. See also Section 6.5.

- The Russian and Tuatschin corpora do not have morphological segmentation.

Chapter 4

The ACQDIV Corpus Database

4.1 Getting access and adding data

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

> Moran, Steven, Robert Schikowski, Danica Pajović, Cazim Hysi and Sabine Stoll. The ACQDIV 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

We release versions of the ACQDIV Corpus Database pipeline on PYPI and we archive them in Zenodo, which provides a Digital Object Identifier (DOI) for reference. This allows users to cite particular versions of the pipeline and the database for scientific replicability. For specific references to database releases, refer to the Zenodo DOI of the public-facing database.

Due to the nature of the data and the dissemination of child language corpora from very different cultures, some of the corpora in the full ACQDIV sample are not open source (as discussed in Section 4.1. Note that there is restricted access to Chintang (Stoll et al. Unpublished), Inuktitut (Allen Unpublished), Russian (Stoll & Meyer 2008), Tuatschin, Turkish (Küntay et al. Unpublished), and Yucatec (Pfeiler Unpublished). Access is made available via the ACQDIV corpus database terms of agreement.[1]

In accordance with the TalkBank's code of conduct,[2] corpora published in CHILDES must be released under the CC BY-NC-SA 3.0 license.[3] In the ACQDIV corpus database, these corpora include: Cree (Brittain 2015), English Manchester (Theakston et al. 2001), Japanese MiiPro (Miyata & Nisisawa 2009, 2010, Nisisawa & Miyata 2009, 2010), Japanese Miyata (Miyata 2004a,b,c, 2012), Ku Waru (Rumsey et al. 2019),[4] Nungon (Sarvasy 2017b), and Sesotho (Demuth 2015). The ACQDIV Corpus Database (public version) is available on Zenodo (Moran et al. 2019a).

4.2 Conceptual architecture

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

- corpora
- sessions
- utterances
- words
- morphemes

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 Database.

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 4.4 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.

4.3 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 5 for details.

[1] https://www.acqdiv.uzh.ch/en/resources.html

[2] https://talkbank.org/share/rules.html

[3] Creative Commons CC BY-NC-SA 3.0 license

[4] Forthcoming.

Note that the original subcorpora also contain media files (audio and/or video, mostly digitized). The ACQDIV Corpus Database 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.

4.4 Structure of the corpus

4.4.1 Overview and ERD

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

- `corpora`: metadata table consisting of information about each subcorpus
- `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. Figure 4.1 shows an ERD diagram of the database. Two naming conventions are used across tables:

- Foreign keys have the suffix "_fk".
- The database often contains both the original data and a processed 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`).

The following subsections explain the ACQDIV Corpus Database's tables' contents and where the data in these fields originate, i.e. from the original input corpus data ("data") or in our ACQDIV Corpus Database Pipeline aggregation tool ("postprocessing").[5] For more detail information on data aggregation, see Section 6.6).

4.4.2 Table `corpora`

Column	Content	Origin
`id`	an ID for the corpus	postprocessing
`language`	language name	postprocessing
`iso_639_3`	ISO 639-3 language name identifier	postprocessing
`glottolog_code`	Glottolog glottocode	postprocessing
`owner`	name of the corpus owner	postprocessing
`acronym`	acronym for the corpus	postprocessing
`name`	name of the corpus	postprocessing

Table 4.1: Columns of the table `corpora`

[5] https://github.com/acqdiv/acqdiv

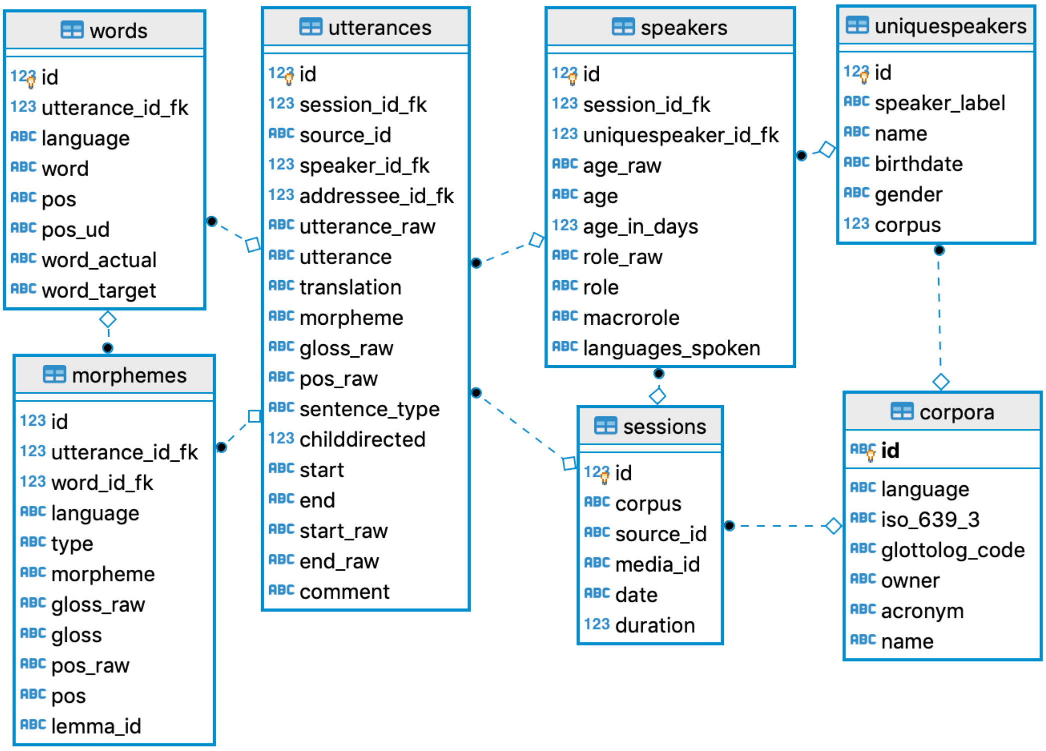


Figure 4.1: Entity-relationship diagram of the ACQDIV Corpus Database

4.4.3 Table `sessions`

Column	Content	Origin
`id`	an automatically generated numeric ID for the session	postprocessing
`corpus`	the name of the corpus the session belongs to	data
`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
`media_id`	the id of the associated media (if available)	data
`date`	the recording date for the session	data
`duration`	the recording sessions duration (if available)	postprocessing

Table 4.2: Columns of the table `session`

4.4.4 Table `speakers`

Column	Content	Origin
`id`	an automatically generated numeric ID for the speaker-session combination	postprocessing
`session_id_fk`	the ID of the session the speaker appeared in, linking to the table `session`	data
`uniquespeaker_id_fk`	the unique ID of the speaker independently of sessions, linking to the table `uniquespeakers`	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

Column	Content	Origin
`languages_spoken`	a space-separated list of all languages the speaker is able speak, given in the form of ISO 639-2 codes	data

Table 4.3: Columns of the table `speakers`

4.4.5 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 may 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
`name`	the full name of the speaker	data
`birthdate`	the birthdate of the speaker in the format YYYY-MM-DD	data
`gender`	the gender of the speaker	postprocessing
`corpus`	the corpus the speaker appears in	data

Table 4.4: Columns of the table `uniquespeaker`

4.4.6 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
`speaker_id_fk`	the ID of the speaker who produced the utterance, linking to the table `speakers`	postprocessing
`addressee_id_fk`	the ID of the addressee (when available)	

Table 4.5: Columns of the table `utterance`

Column	Content	Origin
`utterance_raw`	the original orthographic representation of an utterance (created by concatenating the single words if no separate representation is available	
`utterance`	an orthographic representation of the utterance (created by concatenating the single words if no separate representation is available; cleaned of punctuation marks)	postprocessing
`translation`	a free translation of the utterance (mostly English but Spanish for Yucatec)	data
`morpheme`	all morphemes contained in an utterance, separated by spaces (concatenated morphemes)	postprocessing
`gloss_raw`	all glosses contained in an utterance, separated by spaces (concatenated morphemes)	postprocessing
`pos_raw`	all part-of-speech tags contained in an utterance, separated by spaces (concatenated morphemes)	postprocessing
`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
`end`	the point in time in an associated media file where the utterance ends; format HH:MM:SS.	postprocessing
`start_raw`	like `start` but not unified to HH:MM:SS	data
`end_raw`	like `end` but not unified to HH:MM:SS	data
`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

Table 4.5: 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	+	+	+	+	+	+	+	+	+	+	+

Table 4.6: Presence of columns in the table `utterances`

tier	CLC (ctn)	CCLAS (crl)	JCLD (ind)	AIC (ike)	MPJC (jpn)	MYJC (jpn)	StRuC (rus)	DSC (sot)	KULLDD (tur)	PYC (yua)	SNC (yuw)
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 4.6: Presence of columns in the table `utterances`

4.4.7 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
`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 4.4.10), 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
`pos`	the standardized part-of-speech tag of the stem of the word	postprocessing
`pos_ud`	the universal part-of-speech tag[6] of the stem of the word	postprocessing
`word_actual`	the word form the speaker actually produced; may be empty when only the target form is known	data

Table 4.7: Columns of the table `words`

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

Column	Content	Origin
`word_target`	the word form the speaker intended to produce; may be empty when only the actual form is known	data

Table 4.7: 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	+	+	+	+	+	+	+	+	+	+	+
warning	+	+	+	+	+	+	+	+	+	+	+
word	+	+	+	+	+	+	+	+	+	+	+
word_actual	+	+	+	+	+	+	+	+	+	(+)	-
word_target	-	+	+	+	+	+	-	+	+	+	+
pos	+	+	(+)	+	+	+	+	+	+	+	+
pos_ud	+	+	(+)	+	+	+	+	+	+	+	+

Table 4.8: Presence of columns in the table `words`

4.4.8 Table `morphemes`

Column	Content	Origin
`id`	an automatically generated numeric ID for the morpheme	postprocessing
`utterance_id_fk`	the ID of the utterance the morpheme belongs to, linking to the table `utterances`	data
`word_id_fk`	the ID of the word the morpheme belongs to, linking to the table `words`	data
`language`	the language of an individual morpheme; equals the corpus language by default	data

Table 4.9: Columns of the table `morphemes`

Column	**Content**	**Origin**
`type`	the morpheme type (actual vs. target, (see Section 4.4.10). 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.	
`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
`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 4.5.3. 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
`pos_raw`	the original part-of-speech tag (before standardization)	data
`pos`	a part-of-speech tag. Parts of speech are also standardized. The project-internal set of tags is given in Section 4.5.4	data/postprocessing
`lemma_id`	the lemma dictionary ID (only available for Chintang)	postprocessing

Table 4.9: 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	+	+	+	+	+	+	+	+	+	+	+
utterance_id_fk	+	+	+	+	+	+	+	+	+	+	+
warning	+	+	+	+	+	+	+	+	+	+	+

Table 4.10: Presence of columns in the table `morphemes`

4.4.9 Table `all_data` (view)

This table view that is generated by joining and denormalizing all of the tables mentioned above. 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, e.g. utterance_id.

When using the `all_data` view, it is critical to understand that many of the `morpheme` and `word` fields are `NULL`. This is due to a decision by developers to keep both the word and morpheme tiers **even when they do not align correctly in the input data**. This allows users to use the word, morpheme, gloss, etc., tiers independently without throwing out all data that does not align. Unfortunately, the input corpora are quite problematic, and as such, there are many misalignments.

4.4.10 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").

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 Database features three columns: `word_actual`, `word_target`, and `word`. The latter is intended for easy searches regardless of the actual/target

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 4.11: Actual and target tiers in the original subcorpora

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 Database.

4.5 Conventions

4.5.1 Transcription conventions

The transcription conventions used in the ACQDIV Corpus Database 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"). An exception is when is hyphens, e.g. he-eh, and underscores, e.g. ice_cream. These exceptions are mainly low frequencies errors in the input data.

- 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.
- 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.

4.5.2 Roles and macroroles

The ACQDIV Corpus Database 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 Database.

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

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).

4.5.3 Grammatical glosses

The ACQDIV Corpus Database 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

4.5.4 Part-of-speech tags

The ACQDIV Corpus Database 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 set contains the 17 following categories.[7]

ADJ	adjective	PART	particle
ADP	adposition	PRON	pronoun
ADV	adverb	PROPN	proper noun
AUX	auxiliary	PUNCT	punctuation
CCONJ	coordinating conjunction	SCONJ	subordinating conjunction
DET	determiner	SYM	symbol
INTJ	interjection	VERB	verb
NOUN	noun	X	other
NUM	numeral		

The 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 4.12.

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.

Table 4.12: Problematic mappings of raw to UD POS tags.

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

Corpus	Mapping	Comment
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.
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 4.12: Problematic mappings of raw to UD POS tags.

Corpus	Mapping	Comment

Table 4.12: Problematic mappings of raw to UD POS tags.

4.5.5 Additional remarks on the POS categories

When it comes to pronouns and the like, Universal Dependencies (UD) works very differently from the ACQDIV POS tags. There is DET, which can basically be anything that is used adnominally but not an adjective, and there is PRON, which is a similar dustbin category but for nominal/referential use. ACQDIV is both more and less precise because the corpora allow us to distinguish e.g. between adnominal demonstratives (PRODEM) and adnominal numerals (NUM, both DET in UD) but on the other hand we lump a lot of things together to PRODEM, where our corpora do not have a smallest common denominator, whereas UD distinguishes PRON and DET.

All article-like things are DET in UD, even adnominal numerals (cf. http://universaldependencies.org/u/pos/index.html). That is why all the Turkish categories above end up as DET. In the ACQDIV system, the elements that are clearly deictic (ART:DEM, ART:WH) end up as PRODEM. ART:IDEF, ART:INDEF, and ART/INDEF are all inconsistent labels for a single thing, viz. the numeral *bir* 'one', which is very often used like an indefinite article in Turkish. In UD this is very clearly DET, but since deicticity is a bit less clear here and PRODEM is already big and ugly we chose to classify it as NUM in ACQDIV. One may replace this by PRODEM if they wish – both decisions are compromises where the ACQDIV system does not really match the grammar of Turkish too well.

The label "ART" is another case of inconsistent labeling. Basically it is underspecification – it could be either ART:DEM or ART:INDEF. We are not sure why this label is used (perhaps it stems from an earlier phase of the KULLDD project where they did not differentiate between adnominal demonstratives and numerals like *bir* yet. Anyways, since we do not know if it is PRODEM or NUM, it is marked ???. If ons chooses to include *bir* in PRODEM, ART would become PRODEM, too, because it could no longer mean NUM.

Chapter 5

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 database. Note that the input data used for the ACQDIV Corpus Database 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 CHAT and Toolbox. Section 5.1 gives a brief introduction to the two formats. Most corpora were originally formatted as CHAT. For details on the conversion work done by the ACQDIV core team to convert CHAT and Toolbox into the ACQDIV Corpus Database, see Chapter 6. The remaining sections in this chapter deal with the particularities of the individual subcorpora.

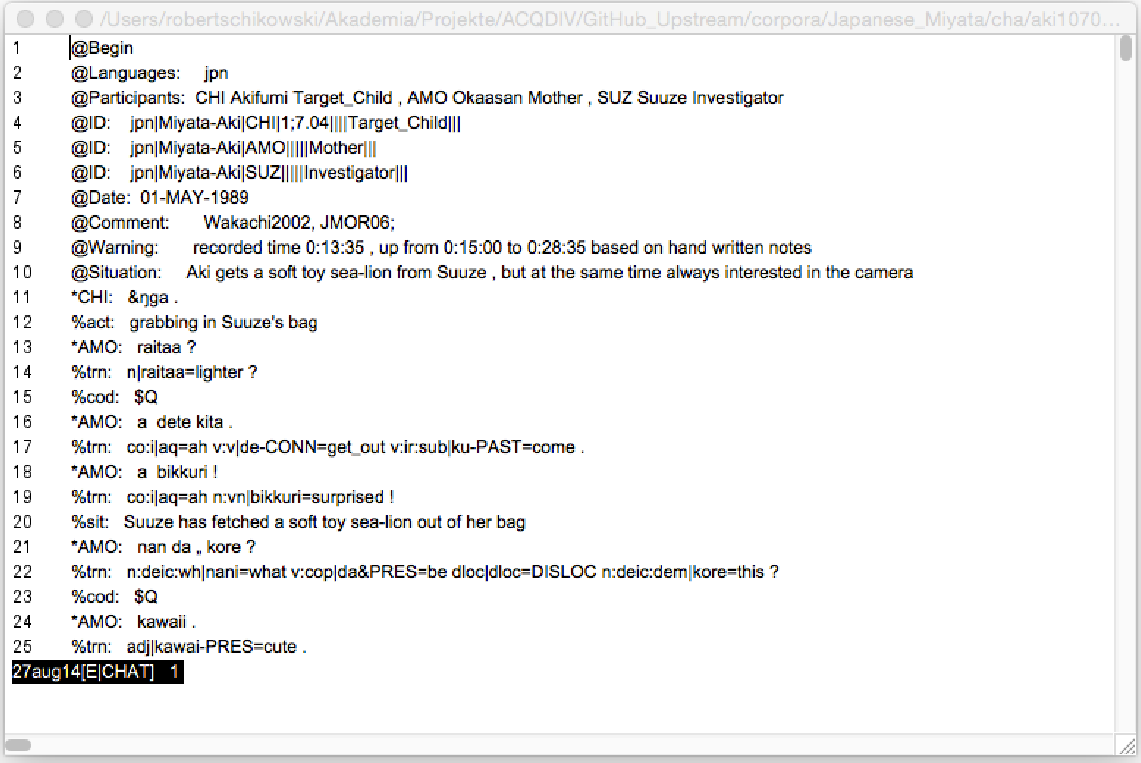
5.1 Original corpus formats

5.1.1 CHAT

CHAT is the original format of most subcorpora: Cree, Indonesian, Inuktitut, Japanese MiiPro and Miyata, Nungon, Russian, Sesotho, Turkish, and Yucatec. CHAT is the format associated with the CHILDES online archive of child language acquisition corpora (MacWhinney 2000). The full specification can be found at https://talkbank.org/manuals/CHAT.html. 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 5.1. The file was opened in CLAN, the editor associated with the format.

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


/Users/robertschikowski/Akademia/Projekte/ACQDIV/GitHub_Upstream/corpora/Japanese_Miyata/cha/aki1070...
27aug14[E|CHAT] 1

```
@Begin
@Languages:    jpn
@Participants:  CHI Akifumi Target_Child , AMO Okaasan Mother , SUZ Suuze Investigator
@ID:    jpn|Miyata-Aki|CHI|1;7.04||||Target_Child|||
@ID:    jpn|Miyata-Aki|AMO|||||Mother|||
@ID:    jpn|Miyata-Aki|SUZ|||||Investigator|||
@Date:  01-MAY-1989
@Comment:      Wakachi2002, JMOR06;
@Warning:      recorded time 0:13:35 , up from 0:15:00 to 0:28:35 based on hand written notes
@Situation:    Aki gets a soft toy sea-lion from Suuze , but at the same time always interested in the camera
*CHI:   &ŋga .
%act:   grabbing in Suuze's bag
*AMO:   raitaa ?
%trn:   n|raitaa=lighter ?
%cod:   $Q
*AMO:   a  dete kita .
%trn:   co:i|aq=ah v:v|de-CONN=get_out v:ir:sub|ku-PAST=come .
*AMO:   a  bikkuri !
%trn:   co:i|aq=ah n:vn|bikkuri=surprised !
%sit:   Suuze has fetched a soft toy sea-lion out of her bag
*AMO:   nan da „ kore ?
%trn:   n:deic:wh|nani=what v:cop|da&PRES=be dloc|dloc=DISLOC n:deic:dem|kore=this ?
%cod:   $Q
*AMO:   kawaii .
%trn:   adj|kawai-PRES=cute .
```

Figure 5.1: The first lines of a typical CHAT file, opened in CLAN

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.

To overcome this problem, we built a custom CHAT parser as part of the ACQDIV Corpus Database pipeline.[1] Our CHAT parser takes as input: a folder of CHAT files per corpus and a configuration file that defines basic metadata about the language, corpus, and mappings between the corpus-specific glosses and parts-of-speech, and with the ACQDIV parts-of-speech schemes (an in-house version based on the Leipzig Glossing Rules and Universal Dependencies tags; see Section 4.5 for details). Although CHAT has a semi-well-defined format, we have found that the morphology tier `%mor:` differs from corpus encoder-to-corpus encoder. As such, we extend our CHAT parser for corpus-specific reading, cleaning, and parsing routines. For a detailed description, see Section ??.

For users of CHAT and potential developers of the ACQDIV Corpus Database pipeline, there exists the following CHAT tools:

- 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.

[1] https://github.com/uzling/acqdiv/parsers/chat/README.md

- Chatter is a parser that can transform CHAT to TalkBank XML. It can be downloaded from http://talkbank.org/software/chatter.html. At the time of writing, there is no comprehensive documentation available.

5.1.2 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/. The subcorpora in ACQDIV that are encoded in Toolbox are: Chintang, Indonesian, Ku Waru, Qaqet, Russian, and Tuatschin.

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 line break 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 5.2 shows an example for one record in a typical Toolbox file.

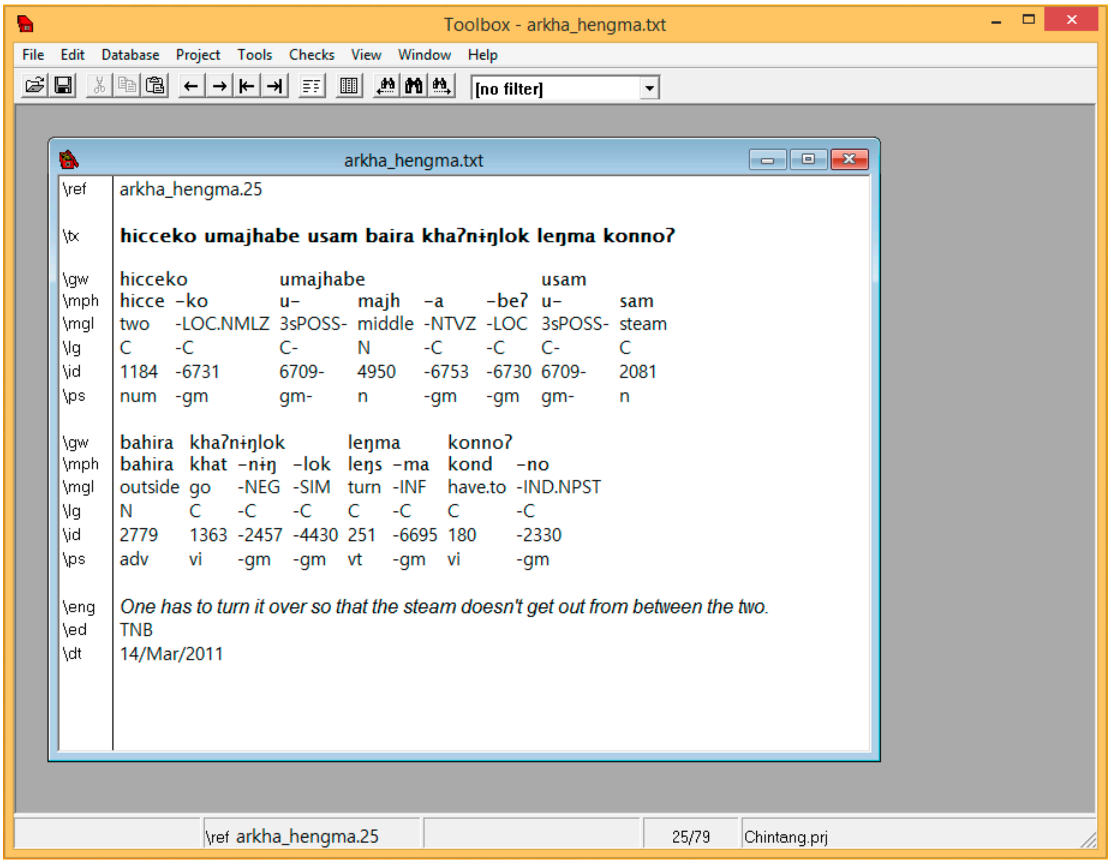
5.2 Chintang

5.2.1 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 Database, is as follows:

> 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.



```
\ref  arkha_hengma.25

\tx   hicceko umajhabe usam baira khaʔnɨŋlok leŋma konnoʔ

\gw   hicceko         umajhabe                   usam
\mph  hicce -ko       u-      majh   -a    -beʔ  u-      sam
\mgl  two   -LOC.NMLZ 3sPOSS- middle -NTVZ -LOC  3sPOSS- steam
\lg   C     -C        C-      N      -C    -C    C-      C
\id   1184  -6731     6709-   4950   -6753 -6730 6709-   2081
\ps   num   -gm       gm-     n      -gm   -gm   gm-     n

\gw   bahira  khaʔnɨŋlok        leŋma konnoʔ
\mph  bahira  khat -nɨŋ  -lok  leŋs -ma   kond    -no
\mgl  outside go   -NEG  -SIM  turn -INF  have.to -IND.NPST
\lg   N       C    -C    -C    C    -C    C       -C
\id   2779    1363 -2457 -4430 251  -6695 180     -2330
\ps   adv     vi   -gm   -gm   vt   -gm   vi      -gm

\eng  One has to turn it over so that the steam doesn't get out from between the two.
\ed   TNB
\dt   14/Mar/2011
```

Figure 5.2: The first lines of a typical Toolbox file, opened in the Toolbox program

5.2.2 Recording scheme

5.2.3 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.

5.2.4 Corpus format

The input format used for the ACQDIV Corpus Database is Toolbox. Table Table 5.2 shows how the fields in the ACQDIV Corpus Database are related to tiers in the input.

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

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 5.1: Recording scheme for the Chintang corpus

target table	**target field**	**source**
sessions	source_id	file name
utterances	source_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 5.2: Chintang tiers

5.2.5 Language Notes

Complex morphology

- Nouns have case, number, possession
- Pronouns have case, number, clusivity
- Demonstratives have case, number, “access” (to referents)
- Verbs have, person, tense, aspect, mood, polarity, agreement with 1-2 arguments
- “vector verbs/V2”: grammaticalized verb stems are chained with normal verb stems to mark aspect, aktionsart, movement, valency manipulation; this creates very long verb forms and a large number of distinct verb forms
- Morphological quirks: suffixes may be copied within a single word form, free prefix ordering, endoclitics and prefixes within complex verb forms, unusual exponence patterns (several affixes marking single function, portmanteaus, zero-marked forms with complex meaning)
- vertical deixis incorporated into several grammatical areas:
  - Distal deixis
  - Case markers
  - Verbs of movement towards speaker

Syntax simple clauses

- Default word order SOV (with a lot of flexibility in argument ordering)
- Case frequently aligns S=P=T=G vs. A, but:
- Pronouns either take ERG optionally or not at all (-> DAM)
- ERG is deleted in valency alternations
- There are valency classes with alternative case frames
- Agreement does not exhibit a uniform alignment pattern because each marker has a different pattern (may e.g. be S=A, S=P, S=A=P...)
- There is a highly frequent alternation similar to an antipassive in which A behaves similarly to S (zero case-marking, S-AGR); this is triggered by the definiteness of the argument triggering P-AGR (nonspecific -> detransitivized)
- A large proportion of transitive verbs are ambitransitive

Syntax complex clauses

- Multitude of non-finite forms (some of them without any of the usual morphology, some with quirky reduced patterns): infinitive, converbs, participles
- Central role of nominalization. clitic NMLZ are highly polyfunctional:
    - ○ Make a non-nominal nominal; non-nominal may be a verb but also any other POS or phrase, including complete clauses and case-marked nouns (genitive is -LOC-NMLZ)
    - ○ Make a non-nominal adnominal (-> relative clauses, externally or internally headed)
    - ○ At the same time mark definiteness -> may sometimes even attach to nouns
- Complex clauses where an INF or CONV is combined with a finite light verb/modal verb ('can', 'must', 'begin to' etc.) exhibit complex raising patterns including LDA. Some differential marking patterns (e.g. based on volitionality) only exist within this area.
- Conjunctions

Lexicon

- High number of Nepali loans (e.g. all numbers above 3)
- Large number of ideophones, most of them reduplicated adverbs with a variety of reduplication patterns, including triplication

5.3 Cree

5.3.1 Publication, accessibility, documentation

The Cree corpus (Brittain 2015) is associated with the Chisasibi Child Language Acquisition Study (CCLAS), which started in 2004 and will reportedly 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 Database. Some documentation for the CCLAS corpus can be found in the Cree Auto-Parser Guide (Acton 2013).

5.3.2 Recording scheme

The following information holds for the subcorpus included in the ACQDIV Corpus Database ("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 5.3: Recording scheme for the Cree corpus

5.3.3 File system and formats

Cree file names use to be 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” with some files containing an undocumented suffix “ms” behind the date, e.g. “10-A1-2005-11-21ms”. More recently, these files have been renamed with a seemingly arbitrary filename, e.g. 020114.cha.

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.

The input format used for the ACQDIV Corpus Database is TalkBank CHAT. The files are available online from TalkBank.[2]

5.3.4 Corpus format

The Cree morphology tiers are structured as follows:

- Words are separated by spaces, morphemes are separated by “~”.
- “%%%” 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.

[2] https://phonbank.talkbank.org/access/Other/Cree/CCLAS.html

5.3.5 Language Notes

The following notes come from the primary sources (Dyck et al. 2006, Junker 2000):

> Dyck, C., Brittain, J., and MacKenzie, M. 2006. ”Northern East Cree accent”. In Proceedings of the 2006 Annual Conference of the Canadian Linguistics Association. 27-30.
>
> Junker, M.-O. (ed.). 2000-2014. The Interactive East Cree Reference Grammar. Retrieved from http://www.eastcree.org/cree/en/grammar/

Morphology

- Derivation: many nouns and verbs are derived from the combination of bound morphemes with recognisable form, but indeterminate semantics, further derivations (causative, applicative, etc.) and/or compounding (also nouns with fully inflected intransitive verbs) are possible. Combination possibilities of free and bound elements almost unlimited.
- Noun stem classes: dependent on possessed stem form and coda of stem, affecting allomorphy of stem and suffixes
- Nominal inflection: number, animacy, obviation (obligatory proximate/obviative marking of overt NPs to distinguish third persons), possession, locative, vocative, diminutive, simulative
- Many different pronoun types: personal, inclusive (’me too’), alternative, indefinite, dubitative, demonstrative, absentive, hesitation, focus, identification
- Personal pronouns distinguish number, animacy, clusivity and obviation
- Reduplication:
  - Of verbs for intensity, repetition
  - Of numerals for distributivity (’each’)
- Verb classes: distinguished for transitive/intransitive and animate/inanimate S or P
- Verbal inflection:
  - Person of S, A and P marked by prefixes and/or suffixes
  - Complex tense (neutral, present, preterite), aspect (habitual/iterative), mood & evidentiality (indicative, independent, conjunct, dubitative, indirect, subjective) marking that follows the stem
  - Two imperatives (delayed and immediate)
  - Conjunct mood triggers a change in the vowel of the first syllable of the stem according to complex morphophonological rules.
  - For many TAM-subparadigm there are also ’relational’ forms which express that the object is possessed by a third person who is not co-referential with the agent of the predicate
- Participles are used for nominalisation, but take a mixture of nominal and verbal inflection
- So called “preverbs” (located between person prefix and) are often used for conjunction, relativisation, subordination, tense, modality, aspect, etc.

Syntax

- Direct/inverse system with hierarchy 2-1-3(proximate)-3(obviate)-3(nonspecific) is important for the whole of syntax
- VOS word order is most unmarked for direct, VSO for inverse (South East Cree, no information on North East Cree syntax)

Phonology and Orthography

- Few consonants /p, t, tʃ, k, k^w, s, ʃ, h, m, n/ with [w,j] as allophones of /i,u/

- Vowels follow historical spelling, but many distinctions have been lost and phonologically only /i, u, a, ə, ʊ/ with many allophones. Distinctions are tense vs. lax and for the lax vowels rounding. Historical diphthongs may be monophthongised.
- Syllable structure: (C)V(C), word-finally (C)V(C)(C) with an additional onset consonant after the coda. Onsets may be any consonant, codas only [h, s, ʃ] (and [m, n] in syncopated structures.)

5.4 English Manchester

5.4.1 Publication, accessibility, documentation

The English Manchester corpus (Theakston et al. 2001) should be cited as:

> Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, 28, 127–152.

It is made available online as part of TalkBank/CHILDES data under the Creative Commons CC BY-NC-SA 3.0 license. Detailed information about the corpus and its contents are available at: https://childes.talkbank.org/access/Eng-UK/Manchester.html.

5.4.2 Recording scheme

number of children	12
age ranges	1;8.22–2;0.25
recording rhythm	60 min in every 3 week period for a year
recording environment	indoors at home
other speakers	mothers
other languages	unknown

Table 5.4: Recording scheme for the English Manchester corpus

5.4.3 File system and formats

The input format used for the ACQDIV Corpus Database is TalkBank CHAT. The files are available online from TalkBank.[3] All files are encoded as UTF-8 text.

5.5 Indonesian

5.5.1 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.

[3] https://childes.talkbank.org/access/Eng-UK/Manchester.html

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.

5.5.2 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 5.5: Recording scheme for the Indonesian corpus

5.5.3 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.

5.5.4 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 5.6 shows how the tiers in the ACQDIV Corpus Database 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 ??).

- 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 5.6 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 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 Database uses the extended format throughout.

target table	target field	source tier
sessions	source_id	file name
utterances	source_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 5.6: 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 5.7: Indonesian tiers with differing contents in the first two Toolbox records

- 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 5.1.2).

5.5.5 Language Notes

The following language notes come from the primary sources (Sneddon et al. 2012, Macdonald 1976):

> Sneddon, James Neil. 1996. *Indonesian: A Comprehensive Grammar.* London: Routledge.
>
> Macdonald, R. Ross. 1976. *Indonesian Reference Grammar.* Washington D.C.: Georgetown University Press.

Morphology

- Noun phrases are not marked for case, but possession can be marked with enclitic pronouns
- Plural formed either with para preceding N or by fully reduplicating N. Plural is not expressed if clear from context or reference is generic.
- Numeral classifier system, but reduced from original variety and obligatoriness and now usually only used with numeral 1.
- Nominalisation may be marked (by demonstrative, nominalising prefix) or unmarked (i.e. only by syntactic position)

- Derivational morphology: concatenative
  - Productive prefixes and suffixes and circumfixes (usually valency changing)
  - Some prefixes/suffixes and all infixes are no longer productive
  - Various forms of reduplication are extremely productive and can take further derivational morphology
  - Compounds may simply be juxtaposed or morphologically bound

Simple clause syntax

- Default word order SVO
- Monomorphemic TAM markers precede predicate
- Negation: Nominal negation vs. other negation, prohibitative (=special negator)
- Tripartate voice system: simple verbs are usually intransitive, transitives (agent focus) and passives (patient focus) are derived by prefixes *meN*- and *di*-
  - Enclitic pronouns used with active transitive verbs agree with the patient.
  - Passives may also be formed by putting a (proclitic) 1/2 pronoun before simple verb (P A=V) or by attaching an enclitic 3 pronoun to a simple verb (P V=A)
  - Passives formed with *di*- can only be used with 3 person
- Ditransitives are derived by suffixes -*kan*/-*i* attached to a transitive verb.

Complex clause syntax

- If subject of independent and dependent clause is the same, it cannot be repeated.
- Variety of subordinators, relativisers, sentence linkers
- yang has many functions (used as relativiser, nominaliser, adjectiviser)

Lexicon

- Heavily influenced by Arabic, Sanskrit, Dutch and other indigenous languages of the Indonesian archipelago.
- Pronouns distinguish person, number and clusivity (only in 1PL) and also vary with social deixis. No pronoun for general reference (animals or things), although increasingly 3 is used for this. Some pronouns may be cliticised, but in all situations full pronouns are also possible. Pronouns are often avoided for reasons of politeness.
- Prepositions and conjunctions are mainly borrowed (apart from three original locational prepositions) and seem to be recent innovations in the language, having rather fluid semantics and many semantic overlaps.

5.6 Inuktitut

5.6.1 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 5.8: Recording scheme for the Inuktitut corpus

5.6.2 Recording scheme

5.6.3 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).

5.6.4 Corpus format

The Inuktitut morphology tier 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.
  - "&amp;" (sic) connects a stem gloss with a grammatical gloss (e.g. "here&amp;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.
  - "&lt;", "&gt;" (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.

5.6.5 Language Notes

The following language notes come from the primary sources (Allen 1996, Swift 2008):

> Allen, Shanley E. M. 1996. *Aspects of Argument Structure Acquisition in Inuktitut.* Amsterdam: John Benjamins
>
> Swift, Mary. 2004. Time in Child Inuktitut. A Developmental Study of an Eskimo-Aleut Language. Berlin/New York: Mouton de Gruyter.

Morphology

- Highly polysynthetic, over 400 productive affixes and clitics which attach to verbal, nominal or uninflected particle roots, allowing PoS to change several times within one word.
- Complex, partly opaque morphophonology.
- Nouns inflected for case (8 cases), number (SG, DU, PL) and possession (person and number of possessor), with portmanteau forms (e.g. ABS.SG.1SG.POSS)
- Affixal verbs, usually cannot be roots, but lexically function in much the same way as root verbs.
- Verbs inflection maps grammatical relations in portmanteaus with mood (e.g. IND.2SG>1SG), 6-7 persons and 9 moods. Also see Documentation/other_grammar/Inuktitut_moods.
- Antipassivisation can be used with all transitive verbs, but is often not overtly marked. Therefore, it makes more sense to distinguish between two-argument and one-argument inflection, rather than transitive and intransitive inflection.
- Tense distinction between NFUT (Ø) and FUT (many markers), FUT is possibly not obligatory. Many optional metric tense markers (future and past). Interpretation of NFUT depends on Aktionsart (telicity).
- Tense/aspect markers used when using a verb in a different way to its default interpretation.

Parts of Speech

- Similarity between certain verbal inflection and nominal possessor inflection and the possibility of expressing the same meaning (e.g. *anguti-up nanuq kapi-janga* [man-ERG.SG bear.ABS.SG stab-IND2.3SG>3SG] 'The man stabbed the bear' and *anguti-up nanuq kapi-jaq-nga* [man-ERG.SG bear-ABS.SG stab-PP-ABS.3SG] 'The bear is the man's stabbed one' are almost identical) in verbal or nominal constructions shows that historically verbal inflection was probably derived from nominal inflection for possession.

Simple Clause Syntax

- Default word order SOV, ergative morphosyntactic alignment, but unmarked antipassive constructions are becoming more frequent.

Complex clause syntax

- Subordinate clauses take special mood marking (e.g. causal, conditional, dubitative, etc.) and do not need any other syntactic marking.

5.7 Japanese MiiPro

5.7.1 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).

5.7.2 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 5.9: Recording scheme for the Japanese MiiPro corpus

5.7.3 File system and formats

MiiPro files use to be composed of the code of the target child and the recording date as “YYYY MM DD” but without any separators, e.g. “aprm19990515”. More recently, the filenames on CHILDES have been changed to the target child’s name followed by a numerical identifier, e.g. Asato_11026.cha.

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

The input format used for the ACQDIV Corpus Database is TalkBank CHAT. The files are available online from TalkBank.[4]

5.7.4 Corpus format

The MiiPro morphology tier 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.

5.7.5 Language Notes

The following language notes come from the primary source (Kaiser et al. 2013):

> Kaiser, Stefan. 2013. *Japanese: a comprehensive grammar.* Abingdon: Routledge 2013.

Morphology

- Overall features: concatenative, monoexponential case and TAM, medium synthesis, low flexivity
- No person marking on verb and the dropping of arguments when clear from context means that participants are pragmatically inferred. Verbs are marked for: Polarity, politeness, tense, aspect, mood, causative, voice
- Two existential verbs for animate and inanimate nouns
- NPs are marked for case and topicality (core case markers are optional)
- Reduplication of some nouns and adverbs can indicate plurality
- Aspect and modality expressed through complex predicates (clausal chaining like structures of two or more verbs)
- Indirect, hearsay, inferential evidentiality encoded by a variety of grammaticalised strategies
- Complex verbal predicates (verbal compounds, serial verbs, clause chaining)

[4] https://childes.talkbank.org/access/Japanese/MiiPro.html

Simple clause syntax

- Topic-comment (TOP)SOV unmarked word order
- Adversative passive and causative may be combined (passivised causatives and causitivised passives)
- Many sentence final particles marking patterns of dominance with respect to information exchange

Complex clause syntax

- Relative clause marked by WO only (preposed)
- Many constructions involve relative clauses on highly abstract head nouns. Nominalisations can also be viewed as such a strategy and are used for many unexpected functions, e.g. pragmatic
- Double relativisation possible (relative clause embedded in a relative clause)
- Interrogative words can enter rather freely into coordinate structures, complex noun phrases, adverbial clauses, and sentential subjects
- No tense agreement between subordinate and main clause
- Various types of coordination and subordination using non-finite or conditional verb forms, particles, relative clauses, nominalisation

PoS

- No pronouns; a large variety of social-deictically marked personal nouns can be used as terms of address, but are usually avoided when addressing superiors where names/kinship-terms/social-roles are preferred
- Two classes of adjectives, one predicative (verb-like), the other needs the copula (noun-like)
- Plethora of numeral classifiers conditioned by lexical features (size, shape, animacy, type of animal, unit of time, etc.)

Politeness

- Complex honorification and humility encoded grammatically and dependent on social deixis and formality of speech-situation.
- Honorific and humble forms of nouns are conditioned by ingroup-outgroup distinction, both for address and for reference: e.g. When addressing an outgroup hearer or referring to someone/something belonging to an outgroup addressee's group, an honorific form is used, but when referring to an ingroup referent (person or thing), a humble form is used, while older ingroup members are often addressed with an honorific or familiar form in an informal context.
- Verbs have complex system of honorific and humble forms, sometimes with suppletive polite verb stems. The honorifics also used with imperatives, allowing for a fine-grained scale of polite commands/requests in the imperative.
- Politeness suffix is attached to verbs in formal situations or when speaking to outgroup members
- Speaker-centred speech (marking the speaker's point of view) is the norm, 'neutral' or 'pragmatically unmarked' sentences are often unidiomatic/strange

Lexicon

- Heavily influenced by Chinese, especially nouns, which has a similar role to words of Latin origin in English. There are often native Japanese and Chinese-Japanese words with the same denotation, but different connotations.

5.8 Japanese Miyata

5.8.1 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.

Content-wise 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).

5.8.2 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 5.10: Recording scheme for the Japanese Miyata corpus

5.8.3 File system and formats

The input format used for the ACQDIV Corpus Database is TalkBank CHAT. The files are available online from TalkBank.[5]

The Miyata corpus as published on CHILDES use to contain 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").

[5] https://childes.talkbank.org/access/Japanese/Miyata.html

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

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 Database 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.

5.9 Ku Waru

5.9.1 Publication, accessibility, documentation

The Ku Waru corpus (Rumsey et al. 2019) that has been deposited in ACQDIV is part of a much larger Ku Waru child language corpus that was recorded between 2004 and 2017. The larger corpus includes audio and video files. In time, it will be archived in PARADISEC. The portion included within ACQDIV was recorded during 2013 and 2014, by Ku Waru field assistant Andrew Noma. Noma transcribed the child's utterances as he heard them, and added a translation into adult Ku Waru where he thought it appropriate. He also transcribed the adults' utterances. On another line he translated the children's and adults' utterances into his own variety of non-standard English. The Ku Waru group corrected regular English misspellings in that translation but beyond that they have not tried to correct it except where it would have otherwise been unintelligible. They have phonemicized Noma's orthography, for example by converting his digraph mb to b, which represents a single prenasalised stop phoneme; and by distinguishing orthographically between phonemically distinct palatal consonants and other ones as described below. The project was overseen by Alan Rumsey. Text wrangling was done by Charlotte van Tongeren; glossing by Lauren Reed and Naomi Peck; and post-processing by Stephanie Yam.

The corpus should be cited as:

> Rumsey, Alan, Andrew Noma, Lauren Reed, Naomi Peck, Charlotte van Tongeren & Stephanie Yam. 2019. ACQDIV portion of the Ku Waru Child Language Socialization Study (KWCLSS).

It accessible via Creative Commons CC BY-NC-SA 3.0 license.

5.9.2 Recording scheme

number of children	1
age ranges	2;2.12 - 3;0.5
recording rhythm	160-65 minutes per session recorded monthly, in one continuous session per month
recording environment	inside child's family home
other speakers	mainly the child's father, less often his mother and sometimes other children
other languages	Tok Pisin (a national lingua franca)

Table 5.11: Recording scheme for the Ku Waru corpus

5.9.3 File system and formats

The input format used for the ACQDIV Corpus Database is Toolbox. Table Table 5.12 shows how the fields in the ACQDIV Corpus Database are related to tiers in the input. The metadata are provided in IMDI format.

5.9.4 Corpus format

target table	target field	source
sessions	session_id_fk	file name
utterances	utterance_id	\ref
words	word	\gw
morphemes	morpheme	\mph
morphemes	gloss_raw	\mgl
morphemes	pos_raw	\ps
morphemes	morpheme_language	\lg

Table 5.12: Ku Waru tiers

5.9.5 Language notes

Ku Waru syntax is strictly verb-final. There are systems of verb serialization and clause chaining, which make use of formally distinct series of final and non-final verbs. Final verbs have portmanteau suffixes encoding tense, mode and aspect, and the person and number of their subject. Non-final verbs do not inflect independently for TAM, but share their TAM value with the following final verb. Non-final verbs do inflect for person and number – albeit less fully than final verbs – and for whether their subject is the same or different from that of the following verb (switch-reference). As discussed and exemplified in described in Section 2.1, the two series of switch-reference verb forms are morphologically identical with corresponding optative and subjunctive forms, but are syntactically distinct from them, in that the latter occur only in final position, and the switch reference forms only in non-final position. There is a parallel difference between future verbs, which occur only in final position, and 'imminent' verbs, which occur only non-finally and express a relation of intentionality or 'imminence' between the immanent verb and the following one. In bivalent clauses the word order is usually SOV but sometimes (about 10% of the time) it is OSV. Case is marked by postpositions. These include an “optional” ergative postposition that is used on most but not all subject NPs of bivalent clauses. For further details regarding Ku Waru grammar see (Merlan & Rumsey 1991:322–343).

In the phonemic orthography the letters <b>, <d>, <j> and <g> are used for prenasalized stops: [mb], [nd], [ɲd͡ʒ] and [ŋg] respectively. The digraphs <ny> and <yn> are used for palatal nasal [ɲ], and the digraphs <ly> and <yl> for palatal lateral [ʎ] / [ʎ̥]. The trigraph <rlt> is used for retroflex flapped lateral [ɺ̢]. The letter <l> is used for the most common lateral in Ku Waru: the prestopped velarized lateral [ɡ͡ʟ] / [k͡ʟ̥]. In some utterances, the Ku Waru team has phonemicized the orthography only partially, in order to ensure that they did not override the child’s own idiosyncratic pronunciations of the words.

5.10 Nungon

5.10.1 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.

5.10.2 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 5.13: Recording scheme for the Nungon corpus

5.10.3 File system and formats

The Nungon files are transcribed in CHAT and encoded in UTF-8 text. Filenames include both the code and age referring to the target child, e.g. “TowetOe-020310”. There is a single but frequent non-ASCII character ⟨ö⟩.

5.11 Qaqet

5.11.1 Publication, accessibility, documentation

The Qaqet corpus was recorded between 2014 and 2019, and the processing of the corpus data (transcription, translation, annotation) is an on-going enterprise. Corpus construction was generously funded by the Volkswagen Foundation’s Lichtenberg program (Az 87 100) (2014-2022). The entire corpus is archived with the Language Archive Cologne, and the subset of annotated recordings is deposited with ACQDIV.

The corpus (Hellwig et al. 2014) should be cited as:

> Hellwig, Birgit, Carmen Dawuda, Henrike Frye & Steffen Reetz. 2014. The Qaqet Corpus at the Language Archive Cologne. Online: http://hdl.handle.net/11341/00-0000-0000-0000-202A-0@view.

5.11.2 Recording scheme

number of children	7
age range	0;7 - 7;10 (focus: 2;0-3;11)
recording rhythm	60 minutes per session recorded weekly
recording environment	natural environment inside/outside, as determined by the parents (no researchers present
other speakers	siblings and other children, parents and other adults
other languages	Tok Pisin (national lingua franca)

Table 5.14: Recording scheme for the Qaqet corpus

5.11.3 File system and formats

The Qaqet data are provided in UTF-8 text files in Toolbox format. Each file has an accompanying IMDI metadata file.

5.11.4 Corpus format

5.11.5 Language Notes

The following language notes come from the primary source (Hellwig 2019):

> Hellwig, Birgit. 2019. *A grammar of Qaqet*. Berlin: De Gruyter Mouton. (Mouton Grammar Library, 79.)

It is not easy to establish genetic links between the East Papuan languages. There are, however, a number of typological features that are widespread amongst them and that are not widely shared with neigh-boring Oceanic languages. As Lindström et al. (2007:128) note, "[t]he puzzle presented by these languages is their recurrent structural similarities together with the absence of formal correspondences (...)." This section summarizes the salient typological features of Qaqet, and places them in an East Papuan and areal perspective. For detailed accounts of their distribution in East Papuan and neighboring Oceanic, see Donohue & Musgrave (2007), Dunn, Reesink & Terrill (2002), Dunn et al. (2008), Lindström et al. (2007), Reesink (2005), Ross (1996), Stebbins et al. (2009) and Terrill (2002).

Qaqet has a smallish phoneme inventory (16 consonant and 4 vowel phonemes) that includes a phonemic contrast between voiceless and voiced plosives, a phonemic contrast between /r/ and /l/, intervocalic lenition of voiceless plosives and the recent development of fricative phonemes from this process of lenition. The first two features are considered characteristic features of Oceanic, not of East Papuan (Dunn et al. 2008:743). All four features are shared by the Baining languages and Kuot, but apparently not by the other East Papuan languages of the Bismarck Archipelago (Stebbins et al. 2009). Like many East Papuan languages (Lindström et al. 2007:126), Qaqet has complex syllable structures, allowing for consonant clusters and word-final consonants. The prosodic structures of East Papuan languages are not well-understood, but there exists an excellent description of Kuot prosody (Lindström & Remijsen 2005). Qaqet differs from Kuot in that it does not have lexical stress, but it is otherwise remarkably similar to Kuot in its inventory of pitch movements that mark (mostly) the right edge of intonation units.

Overall, there is good evidence for the existence of different word classes (including the open classes of nouns, adjectives, verbs and adverbs). And although adjectives share morphosyntactic similarities with nouns, there are formal differences that justify setting up a distinct adjective class. The word classes are distinguished on the basis of their morphosyntactic properties, but there is one qualification to be added: many roots can occur underived in different word classes. This phenomenon is also attested in Mali Baining, and like Stebbins et al. (2011:58-59, 95-97) , I consider it a case of conversion (not of acategoriality). Furthermore, there is also good evidence for the existence of phrases (with fixed and contiguous structures), including noun phrases. Nevertheless, there exists a small number of examples that deviate from the typical noun phrase structure, and some examples even show discontinuous structures. It is very likely that information structural constraints account for this variation, but the number of attested examples is too small to attempt larger generalizations.

In the nominal domain, Qaqet exhibits a number of characteristic East Papuan features (Dunn et al. 2002 33-36; Dunn et al. 2008 743; Stebbins et al. 2009a; Terrill 2002). Most notably, nouns are marked for noun class (distinguishing two sex-based and six shape-based classes) and number (distinguishing singular, dual and plural), and the classes overtly appear in a large number of grammatical environments: as free pronouns and as arguments that developed from the free pronouns (object suffixes on verbs and prepositions); and on noun phrase elements agreeing with the noun (adjectives, the numerals 'one' and 'two', demonstratives, some indefinite pronouns, and the interrogative pronouns 'which' and 'who'). Furthermore, the noun classes are mapped onto a smaller system (labeled 'gender' throughout this grammar), which surfaces in the form of possessor indexes on possessed nouns, of subject indexes on verbs, and as associative pronouns. Both nominal classification and the dual/plural distinction are widespread among the East Papuan languages, and are largely absent from Oceanic. But although East Papuan languages tend to have some form of nominal classification, the formal and semantic properties of these systems are very different. The Qaqet system shares similarities with those of other Baining languages, but it remains to be seen whether similarities are also found with the other East Papuan languages of the Bismarcks.

In terms of noun phrase structure, Qaqet has pre-head determiners (demonstratives, indefinites and articles) and post-head adjectival and nominal (including numeral) modifiers. Post-head adjectives and numerals are shared by many East Papuan languages, while pre-head demonstratives are less common (albeit attested, too). The main evidence for Oceanic influence in this domain is the existence of definite and indefinite articles preceding the noun: articles are generally considered an Oceanic borrowing, which are now found in most East Papuan languages of New Britain.

In possessive noun phrases, Qaqet displays the typical Papuan order of the possessor preceding the possessed – despite not (or no longer) having the concomitant Papuan verb-final constituent order. Qaqet also does not have any possessive classifiers, which are typical of Oceanic languages (although they are largely absent in neighboring Oceanic languages, too). There is evidence for Qaqet having the category of inalienable possession (which is typical of Oceanic languages), but it only concerns a handful of items (a few kinship nouns, relational nouns and irregular proforms). The category surfaces in the form of prefixes (and not suffixes, as would be typical for Oceanic).

Similarly, the pronominal categories are characteristic of East Papuan (Dunn et al. 2002: 40-41; Stebbins et al. 2009a) : Qaqet has dedicated dual pronouns, and it shows no evidence for a distinction between inclusive and exclusive first person. Like in many other East Papuan languages, participants are indexed on the verb, but they are indexed by means of subject proclitics and object suffixes (and not by means of subject suffixes, as is otherwise the preference in East Papuan) (Dunn et al. 2002: 52-57; Dunn et al. 2008: 743).

In the verbal domain, Qaqet exhibits the typical lexicalization patterns reported for many Papuan languages (see, e.g., Foley & Foley (1986)). As such, verbs are highly compositional, consisting of simple verbs with general meanings that combine with other elements (usually prepositions, in the case of Qaqet) to form complex verbs with idiomatic, non-compositional, meanings. Similarly, Qaqet pays attention to many of the categories reported for other Papuan languages: it distinguishes continuous and non-continuous aspect (through distinct aspectual verb stems), and controlled and non-controlled events (through argument structure alternations). The only conspicuous absence is stem alternations that depend on the person and number of arguments – a pattern that is found in many East Papuan languages (Dunn et al. 2002:38-57). While the semantic patterns seem to be of Papuan origin, their formal means of expression differ. In particular, verb serialization is not attested synchronically, and it cannot be identified as a source of lexicalization in the verb lexicon either. There is some evidence, though, that at least some particles and modifiers to the predicate originated in multiverb constructions (possibly in serial verb constructions). Instead, Qaqet makes pervasive use of prepositions. It is possible to trace a diachronic development from prepositional phrases functioning as adjuncts via prepositions introducing arguments entailed by the verb semantics to particles and suffixes that have become lexicalized as parts of complex verbs. Verbs obligatorily index their subject (S/A) argument as a proclitic on the verb. There is no case marking.

More generally, prepositions are an areal pattern of the Bismarck Archipelago: they are not typical of (East) Papuan languages (which tend to have postpositions), and they are thus sometimes attributed to contact with Oceanic (Dunn et al. 2002: 33; Dunn et al. 2008: 743). Qaqet does not have any inherited postpositions, but there is a recent development from possessed nouns to relational nouns, which can arguably be analyzed as developing into postpositions.

Prepositions, relational nouns and adverbs predominantly convey spatial meanings. In addition, Qaqet has a complex and prevalent directional system that is based on the mountainous landscape, distinguishing 'down', 'up' and 'across'. This system is shared by the other Baining languages (Stebbins et al. 2011:192-206). Furthermore, Qaqet has a syntactically-defined word class of particles that expresses information on modality, time, aspect, discourse structure and speech act, and that plays an important role in structuring discourse. Again, Qaqet shares this word class with other Baining languages (Stebbins et al. 2011:93-94, 218-229).

Another shared areal pattern is clausal constituent order (Dunn et al. 2002: 32-33, 36-37; Dunn et al. 2008: 743) : Qaqet has AVO   SV constituent order, i.e., it differs from the verb-final order otherwise widespread in (East) Papuan languages. This order is found in all clause types (affirmative and negative statements, questions and commands), and it is also found in one non-verbal construction (the locative construction). In addition, there are two non-verbal constructions (the equative and attributive constructions) that exhibit the reversed order of the predicate being followed by the subject.

As a consequence of its predominant verb-medial order, Qaqet does not have any clause chaining and/or switch reference, which is common in verb-final Papuan languages. Qaqet also does not employ multi-verb constructions synchronically, and, interestingly, there is also no clear evidence for subordination. Qaqet essentially combines main clauses with each other, indicating the differ-ent semantic relationships through a large number of conjunctions and particles.

5.12 Russian

5.12.1 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.

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 5.15: Recording scheme for the Russian corpus

5.12.2 Recording scheme

5.12.3 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 Database the original data were all reencoded to UTF-8. The Russian data do not contain any non-ASCII characters.

5.12.4 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 5.16 shows how the tiers in the ACQDIV Corpus Database are related to tiers in the input.

target table	target field	source tier
sessions	source_id	file name
utterances	source_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 5.16: 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.

5.12.5 Language Notes

Phonology

- Average vowel inventory but many consonants, most of them in two phonemically distinct variants (palatalized vs. non-palatalized/velarized).
- Complex phonotactics including some single-consonant words (prepositions).

Morphology

- Conservative Indo-European morphology with corresponding SAE parts-of-speech; high degree of fusion and inflection.
- Nouns inflect for 2 numbers and 6 cases (+ 3 more in certain residues) and have three genders (M, F, N). Pronouns and numerals are clearly nominal, too, but display a number of inflectional peculiarities.
- Modifier POS such as adjectives or demonstrative and possessive pronouns index number, case, and gender of their head referent. Adjectives have default long forms and alternative short forms, which are only used predicatively and express temporary qualities in this case. Adjectives also feature a predicative comparative form (often expressed by stem alternation; alternating with a synthetic form for attributive use) and an adverbial form.
- Verbs variably index person/number (default) or gender/number (past tense) of one argument (usually S/A). They have 3 basic TM forms (past, non-past, imperative) and a few additional synthetic forms. Non-finite forms are the infinitive, various participles, and a converb ("adverbial participle"). Reflexive verbs are marked by a clitic that is the same for all persons.
- All verb forms are marked for one of 2 aspects. Both aspects have a large number of exponents, including prefixes, suffixes, stem alternations, combinations of all these, and suppletivism. The use of the exponent solely depends on the lexical identity of the verb, so that the alternating forms are often viewed as distinct lexemes linked by the grammatical system. Most often the IPFV form is morphologically basic and the PFV is derived from it. PFV is also used to express immediate future.

- Imperfective motion verbs express an additional distinction (unidirectional vs. multidirectional movement) in a similar fashion.

Simple Sentences

- Word order is free, i.e. based on information structure; old > new yields the default WO SVO. Pro-drop is possible.
- The default alignment is accusative. Many nominal paradigms (details depend on POS, inflection class, and number) have a DOM pattern where animate P are marked by GEN and inanimate P by NOM. Other paradigms feature a dedicated ACC case. The S/A of negated sentences can be marked by NOM (default) or GEN (less definite and related referential factors).
- Verbs agree with S/A.
- The copula is most often zero, making noun-only clauses a frequent phenomenon.
- Counting is complicated. '1' simply agrees with its head referent as an adjective would. When the unmarked case would be expected for the complete NP, the numeral exposes it but the noun is marked by SG.GEN (numerals '2', '3', '4') or SG.PL ('5' or higher). With other cases, both the numeral and the plural noun are marked for the expected case. '1', '2', '3', '4' feature NOM/GEN DOM, hence inanimate P ⟶ NOM numeral + SG.GEN noun, animate P ⟶ GEN numeral + PL.GEN noun. Compounds (twenty-three etc.) behave like their last digit but do not have DOM (hence NOM numeral + SG.GEN (21, 22, 23, 24) or PL.GEN (25, 26, 27, 28, 29, 30). Adjectives follow the same case rules as nouns but are always PL, even when the noun is SG.

Complex Sentences

- SAE inventory of conjunctions, relative pronouns, complementizers
- One (not so common) converb.

5.13 Sesotho

5.13.1 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.

5.13.2 Recording scheme

5.13.3 File system and formats

Examples for file names in the originally 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.

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 5.17: Recording scheme for the Sesotho corpus

More recently these filenames have changed in CHILDES and look like: TseboNeuoe_040500ab.cha. The content however is the same.

All files are encoded as UTF-8 text and only contain ASCII characters. The transcripts are available online.[6]

5.13.4 Corpus format

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.

- 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.

[6] https://childes.talkbank.org/access/Other/Sesotho/Demuth.html

5.13.5 Language Notes

Morphology

- Overall features: primarily concatenative, medium synthesis
- Productive and syncretic (prefixal) noun classes (which may also mark number) marked with different sets of concordance prefixes on nouns, adjectives, demonstratives, possessives, pronouns, numerals; noun class additionally conditions verbal agreement (subject and object) in the third person. Historically the noun class system was semantically based, but now more restricted.
- Case: locative
- Diminutive used for both nouns and verbs
- Closed class of nominal adjectives which have adjectival and relative concord markers, some (verbal) adjectives only take relative concords
- Extensive and productive verbal derivation: applicative, causative, passive, neuter, reciprocal, intensive, reversive, neuter reversive and extensive
- Verbal inflection:
    - Subject and object agreement with person (noun class for third persons); object agreement optional in present tense: conditioned by WO/information structure (no agreement with SVO (so-called indefinite present), optional agreement with SOV (so-called definite present)).
    - Tense: present, immediate past, immediate future, future
    - Aspect: continuous, perfect, consecutive
    - Mood: infinitive, indicative, dependent, subjunctive, hortative, imperative, potential (indicative and dependent) and their negations
    - Modality: deontic
    - Information structure: temporal focus
- Verbal negation is extremely complex and asymmetrical; simultaneously affecting multiple levels (tone, agreement, word structure, syntax; additionally suppletive stems, TAM markers and derivational affixes); often multiple negative morphemes.
- A so-called ”definite/indefinite” distinction in the present tense:

Simple clause syntax

- Pro-drop
- Unmarked WO is SVO
- Many light verbs at various stages of grammaticalisation (lexical > auxiliary) are used in (up to triverbal) coverbal constructions to express additional tenses/aspects. These constructions may additionally take inflectional TAM markers as well. Most light verbs require complements in a particular mood or aspect. The most grammaticalised light verbs may sometimes lose agreement, be contracted with the lexical verb and become prefixes
- Ditransitives requires most animate object to be closest to the verb. If both objects are animate, then order is interchangeable.
- Impersonal constructions in active and passive (like in German) are possible

Complex clause syntax

- Subordination by marking dependent mood on subordinated predicate
- Relative clauses are marked by relative clause agreement on relative clause’s predicate

Phonology

- Vowel harmony, tone sandhi, clicks, ejectives
- Tone as secondary TAM and phrase structure marker

Lexicon

- Some influence from Afrikaans, English and Zulu

5.14 Turkish

5.14.1 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.

5.14.2 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 5.18: Recording scheme for the Turkish corpus

5.14.3 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⟩.

5.14.4 Corpus format

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).

- 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 Database, 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).

5.14.5 Language Notes

The following language notes come from this primary source: (Jaklin 1997)

> Kornfilt, Jaklin. 1997. *Turkish.* London: Routledge.

Morphology

- Concatenative, monoexponential, exclusively suffixing
- Nominal morphology:
  - Marked for case (6) and number (SG,PL)
  - Case and postpositions interact, sometimes the same meaning can be expressed by either an inflected noun or a postpositional phrase.
  - Case assignment of object argument may vary lexically and direct objects receive accusative case only when they are specific (DOM).
  - Reduplication and various forms of compounding are frequent nominal derivational strategies.
  - Human/non-human distinction for nouns: non-human 3 person PL subjects have singular agreement.
- Verbal morphology:
  - Conjugation classes
  - Agreement with person and number of subject
  - Person agreement paradigms depend on the **TAM** of verb, but are not portmanteau forms
  - Various valency changing derivations (passive, causative, reciprocal, etc.)
  - TAM markers often a combination of T, A and M (e.g. evidentiality and past); additionally more than one TAM marker can be attached to a verb
  - Various imperative suffixes, ranging from colloquial to formal, from less polite to more polite
  - Reflexive suffix (-in)
  - Suffix marking reciprocity

  - voice
- Copula:
  - Various copular suffixes, clitics and verbs can all attach verbal morphology. Choice depends on whether it is an nominal, adjectival, adverbial predicate
- Negation:
  - Verbal negation by suffix, nominal negation by negative copula
  - The present ('aorist') tense marker has allomorphs depending on negation
  - No special prohibitative

Simple clause syntax

- Default word order is SOV, accusative morphosyntactic alignment
- Pro-drop

Complex clause syntax

- Non-finite and subordinate clauses are also pro-drop.
- Infinitives, participials (gerunds) and nominalised verbs (most common strategy) behave in similar ways and are the main strategies of subordination. Subordinators with finite subordinate clauses are possible, but uncommon.
- Non-finite verbs cannot take the full range of TAM suffixes.

Lexicon

- Numerical classifiers (marginal)

5.15 Yucatec

5.15.1 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.

5.15.2 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 5.19: Recording scheme for the Yucatec corpus

5.15.3 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 re-encodings. The only characters that naturally appear in Spanish or Yucatec orthography are vowels with acute accents, ⟨ñ⟩, and ⟨ʼ⟩ (= modifier letter apostrophe, U+02BC).

5.15.4 Corpus format

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 Database, `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.

5.15.5 Language Notes

The following language notes come from the primary sources (Bohnemeyer 2002, Maimonides et al. 1989)

> Bohnemeyer, Jürgen. 2002. *The grammar of time reference in Yukatek Maya.* Munich: LINCOM EUROPA.
>
> Bolles, David and Alejandra Bolles. 2001. *A Grammar of the Yucatecan Mayan Language.* Revised Edition. Lancaster, California: Labyrinthos.

Morphology

- Mildly polysynthetic
- Inflection may go between root and derivation, complex words are not tightly bound and allow intrusive morphology.

- Incorporation of nouns and adverbs possible
- Verbs:
  - Inflected for person, number, transitivity, mood, verb class, voice
  - There are a group of portmanteau suffixes called 'status' in the Mayanist literature, the choice of this suffix depends on mood/aspect (incompletive, completive, subjunctive, imperative, extrafocal) and predicate class (transitive, inactive, positional, inchoative, active).
  - Two sets of bound pronouns ("A" and "B") cross-reference person and number on verb. Their alignment is variable, but set A usually indexes A and certain S and set B indexes P and other S.
  - Many verbal derivations: valency changing, aspectual (passive, causative, applicative, perfect, gerundive, etc., etc.)
- Nouns:
  - Inflected for number (irregular, allomorphs lexically determined)
  - Set A bound pronouns are used to cross-index possessor
  - Set B bound pronouns are suffixed to prepositions to express the theme/goal of stative predicates. Portmanteau forms of one preposition and set B suffixes comprises independent pronouns, used for non-core arguments or for topicalised core-arguments
- Numerals and *hay* 'how many?' are followed by suffixal numeral classifiers (animate/inanimate or semantically motivated).
- Adjectives: inflect for plural if they do not directly precede the noun.
- Demonstratives formed by *le* N-DEM.suffix, inflecting for distance and number

Simple Clause Syntax

- V-initial word order
- Both ergative and accusative alignment are possible because of split S; the agreement set ("A" or "B") used for S depends on predicate type (incompletive vs. other types).
- Independent verbal clauses must be marked with one of the 15 pre-verbal aspect-mood markers, these markers can be seen as predicates with verbal clause arguments
- Focus constructions, relative clause constructions and content questions affect aspect and mood marking.

Complex Clause Syntax

- Verbal cores, but not whole clauses, can be complements of matrix clauses. Embedded verbal cores must have subjunctive or incompletive.
- Finite clauses can only be linked paratactically or subordinated via conjunctions.

Lexicon

- Mayan numerals (originally vigesimal) are now only used for 1-4, otherwise Spanish numerals are used.
- Much influence from Spanish, Yucatec morphology may attach to Spanish roots.

Chapter 6

Creating the ACQDIV Corpus Database

The ACQDIV Corpus Database is dynamically generated from the original data described in Chapter 5 using a Python aggregation tool that we developed and describe in ??. A new version of the database can be generated each time the original data change, as long as the input formats (i.e. CHILDES CHAT and Toolbox) are not changed substantially, e.g. restructuring of the morpheme tiers in CHAT, renaming of the Toolbox labels.

The original corpus data are extremely heterogeneous and often have greater or lesser 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, timestamps, normalization of labels)

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 can be repeated each time the ACQDIV Corpus Database is generated. This 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 ?? and the documentation that comes with the individual scripts involved in each step.

6.1 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.

6.1.1 Non-textual formats

Most corpora were already formatted as structured text (e.g. CHILDES CHAT 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.

6.1.2 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.

6.1.3 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.

6.1.4 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 Database 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 Database.

- 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 Database 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).

6.2 Cleaning of corpus formats

The two corpus formats that were accepted as input for the ACQDIV Corpus Database are CHILDES CHAT and Toolbox. However, three corpora – Inuktitut, Turkish and Yucatec – were delivered in broken CHAT, so we had to fix them. 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[1] 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 given a new metadata header for every file in each corpus. All

[1] http://talkbank.org/software/chatter.html

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 our cleaning scripts, 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.

6.3 Parsing the corpus data

Parsing in a narrow sense concerns the process that transforms one of the accepted input formats (CHILDES CHAT and 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.

6.3.1 CHILDES CHAT

Initially we parsed the TalkBank XML format, but if proved difficult to maintain. Now we have a suite of parsers in our Python package,[2] which parse CHILDES CHAT. CHILDES CHAT contains a metadata header and text body (described above in Section 5.1.1 and online).[3] For details of our parsing routines, see the documentation in our GitHub repository.[4]

A particularly problematic aspect of parsing the CHILDES CHAT corpora is the flexibility that this encoding format allows its users in terms of morphological annotation. In our TOOLBOX corpora (see next section), morphology is much more consistently encoded because the corpus collectors were quite consistent in using the Leipzig Glossing Rules, which describe a standard for segmenting and labeling morphological annotations that is often used by field linguists.

Morpheme, gloss, and part-of-speech information can be stored either on the same or on separate tiers. The TOOLBOX corpora, and to some extent the Cree and Sesotho CHILDES CHAT corpora, code morphology on separate tiers, which is programmatically easier to parse. However, multiple tiers for morphological analysis introduce a higher probability for misalignments due to user-introduced errors (and thus a higher percentage of the aggregated data being misaligned) between morphemes and their part-of-speech tags and/or glosses.

In comparison to TOOLBOX, most CHILDES CHAT corpora encode the morpheme, gloss, and part-of-speech labels on the same tier. For example, in the Japanese MiiPro corpus (Miyata & Nisisawa 2009, 2010, Nisisawa & Miyata 2009, 2010):

[2] https://github.com/acqdiv/acqdiv

[3] https://talkbank.org/manuals/CHAT.pdf

[4] https://github.com/acqdiv/acqdiv/tree/master/src/acqdiv/parsers/chat

```
*MOT: osanai to . 33893_34834
%xtrn: v:c|os-NEG-PRES ptl:conj|to .
```

The V:C is the part-of-speech label, the OS is the morpheme and NEG and PRES are glosses. As the example shows, there is no gloss for the stem OS. And vice versa, there are no morphemes for the suffixes NEG and PRES, only glosses. Unfortunately, we have found that incomplete morpheme data is common in the CHILDES CHAT corpora. For example, in many corpora affixes are the only annotated glosses and for stems there may be no gloss at all.

Morphemes can also be segmented in different ways, including dashes (−), equal sign (=), hash (#), colon (:), and plus (+). These delimiters have different meanings depending on where they occur, e.g. ':' often codes gloss subcategories, as shown above, but in other corpora it may have more than one use.

6.3.2 Toolbox

We also wrote our own parsers for Toolbox files, which involves the following conceptual steps:

- Files are split into records based on line break 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.).

More details are available online: https://github.com/acqdiv/acqdiv/tree/master/src/acqdiv/parsers/toolbox.

6.4 Parsing the metadata

Metadata are stored in separate files and/or structured differently from corpus data in the case of the Toolbox corpora. They are therefore parsed separately from the latter. For details, see: https://github.com/acqdiv/acqdiv/tree/master/src/acqdiv/parsers/metadata.

Metadata are either read from CHILDES CHAT files directly (from the header of recording session files) or from IMDI (or CMIDI) XML files, which are 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`).

6.5 Parsing session durations

To the extent that we have access to the original corpora media files (e.g. audio or video files), we have extracted recording session durations. The basic format is to use `wget` to download media files. And then to use the `exiftool` command line tool to extract the media durations.[5] The media files resides on various volumes on the UZH server, which can be passed as a parameter to the durations extraction script. For most corpora, the filenames of the media files are not the same as the filenames of the transcripts (Chintang, and to some extent Russian, are exceptions). For these corpora, we extract the media filenames from the metadata (when available) and map the filenames from the transcripts to the media filenames. Session durations are inserted during database postprocessing (Section 6.6) into the session table in `sessions.duration` in total number of seconds. Current coverage for session duration timestamps is given in Table 6.1.

Corpus	Durations	Total sessions
Chintang	473	477
Cree	0	25
English_Manchester1	0	804
Indonesian	931	997
Inuktitut	45	77
Japanese MiiPro	148	148
Japanese Miyata	75	213
Ku_Waru	0	9
Nungon	0	4
Qaqet	0	106
Russian	437	450
Sesotho	67	69
Turkish	367	373
Yucatec	11	234

Table 6.1: Session duration coverage

Japanese Miyata contains only media files for Tai.[6] Sesotho contains only 70 media files. Yucatec contains several hundred and needs to be revisited for potential filename mismatches. Getting the remaining session durations is a work in progress.

[5] https://www.sno.phy.queensu.ca/~phil/exiftool/

[6] https://media.talkbank.org/CHILDES/Japanese/Miyata/

6.6 Data Extraction and Aggregation

Our source code for processing the corpora is available online in our GitHub repository (Moran et al. 2019b).[7] The ACQDIV aggregation pipeline's workflow follows a fork-and-pull request model.[8] Our code base in written in PYTHON and we release the software package via PYPI. The command line tool can then be run with the following commands.[9]

Install the ACQDIV package with PIP (note the optional PYTHON virtual environment):

```
python3 -m venv venv
source venv/bin/activate

pip install acqdiv
```

Contributors should install the package from source:

```
git clone git@github.com:acqdiv/acqdiv.git
cd acqdiv
pip install -r requirements.txt
```

Run the pipeline:

```
acqdiv load -c /absolute/path/to/config.ini
```

Users need to pass an INI configuration file (parameter `-c`) to the `load` command. The repository already contains a sample configuration file (see `src/acqdiv/config.ini`) in which you can adapt the paths to the corpora and for the database file.

We also have a test suite in place to verify that no regression is introduced in the source code and to check the integrity of the database:

```
pytest tests/unittests
pytest tests/systemtests
```

We release versions of the ACQDIV Corpus Database pipeline on PYPI and we archive them in Zenodo, which provides a Digital Object Identifier (DOI) for reference. This allows users to cite particular versions of the pipeline and the database for scientific replicability.

Due to the nature of the data and the dissemination of child language corpora from very different cultures, some of the corpora in the full ACQDIV sample are not open source (as discussed in Section 4.1. Note that there is restricted access to Chintang (Stoll et al. Unpublished), Inuktitut (Allen Unpublished), Russian (Stoll & Meyer 2008), Tuatschin, Turkish (Küntay et al. Unpublished), and Yucatec (Pfeiler Unpublished). Access is made available via the ACQDIV corpus database terms of agreement.[10]

In accordance with the TalkBank's code of conduct,[11] corpora published in CHILDES must be released under the CC BY-NC-SA 3.0 license. In the ACQDIV corpus database, these corpora include: Cree (Brittain 2015), English Manchester (Theakston et al. 2001), Japanese MiiPro (Miyata & Nisisawa 2009, 2010, Nisisawa & Miyata 2009, 2010), Japanese Miyata (Miyata 2004a,b,c, 2012), Ku Waru (Rumsey et al. 2019),[12] Nungon (Sarvasy 2017b), and Sesotho (Demuth 2015). The ACQDIV Corpus Database (public version) is available on Zenodo (Moran et al. 2019a).

[7] https://github.com/acqdiv/acqdiv

[8] https://gist.github.com/Chaser324/ce0505fbed06b947d962

[9] For detailed instructions, see also: https://github.com/acqdiv/acqdiv.

[10] https://www.acqdiv.uzh.ch/en/resources.html

[11] https://talkbank.org/share/rules.html

[12] Forthcoming.

6.7 Adding new corpora

We have designed the ACQDIV Corpus Database aggregation pipeline with the aim of it being extensible. For example, TalkBank contains a range of rich resources that are transcribed, richly annotated, and aligned with audio and video recordings (currently, there are over 130 different corpora representing 26 languages). These data are publicly available online, and as such, we have developed the ACQDIV corpus database aggregation pipeline so that developers can add these, as well as corpora encoded in Toolbox format, to the resulting output database. Detailed instructions on how to add new corpora are described online in the GitHub repository.[13] Here is a very brief description of the workflow.

First, create a new section in the configuration file for the corpus being added. A template section for CHILDES CHAT[14] and Toolbox[15] is available online. Second, create a new Python package under `parsers/corpora/main/<corpus_name>` with the following classes:

- Reader
- Cleaner
- SessionParser
- CorpusParser

Specifically:

- Reader class:
  - Create a class called '<corpus_name>Reader' in a file named 'reader.py' in the package.
  - Make it inherit from the class 'acqdiv.parsers.chat.readers.reader.CHATReader'.
  - Make sure every method of 'CHATReader' has a correct implementation, otherwise override the method.
- Cleaner class:
  - Create a class called '<corpus_name>Cleaner' in a file named 'cleaner.py' in the package.
  - Make it inherit from the class 'acqdiv.parsers.chat.cleaners.cleaner.CHATCleaner'.
  - Make sure every method of 'CHATCleaner' has a correct implementation, otherwise override the method.
- Session parser class:
  - Create a class called '<corpus_name>SessionParser' in a file named 'session_parser.py' in the package.
  - Make it inherit from the class 'acqdiv.parsers.chat.parser.CHATParser'.
  - Override the methods 'get_reader()' and 'get_cleaner()' to return an instance of the newly implemented reader and cleaner class, respectively.
- Corpus parser class:
  - Create a class called '<corpus_name>CorpusParser' in a file named 'corpus_parser.py' in the package.

[13] https://github.com/acqdiv/acqdiv/blob/master/CONTRIBUTING.md
[14] https://github.com/acqdiv/acqdiv/tree/master/src/acqdiv/parsers/chat
[15] https://github.com/acqdiv/acqdiv/tree/master/src/acqdiv/parsers/toolbox

- Make it inherit from the class ‘acqdiv.parsers.corpus_parser.CorpusParser’.
- Implement the method ‘get_session_parser()’ to return an instance of the newly implemented session parser class.

These classes should inherit from already implemented base classes and override any methods that need adaption. Creation of the appropriate objects follows the abstract factory pattern. The main issues to address include are the corpus-specific morphological parsing and the mapping of parts-of-speech and morphological glosses to the standardized label schemes (i.e. Universal Dependencies and/or the Leipzig Glossing Rules).

One needs also to add a Mapper class. To do so, add a mapping of the corpus name to the newly implemented corpus parser in the ‘mappings‘ dictionary of the class ‘CorpusParserMapper’ of the module ‘acqdiv.parsers.corpus_parser_mapper.CorpusParserMapper’.

Lastly, for the loader procedure, add the ini file to the ‘configs’ list in the method ‘load()’ of the module ‘acqdiv.loader’. Then follow the instructions above in Section 6.6. Voilà!

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.

Bohnemeyer, Jürgen. 2002. *The grammar of time reference in Yukatek Maya.* Lincom.

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.

Donohue, Mark & Simon Musgrave. 2007. Typology and the linguistic macrohistory of Island Melanesia. *Oceanic Linguistics* 348–387.

Dunn, Michael, Stephen C Levinson, Eva Lindström, Ger Reesink & Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 710–759.

Dunn, Michael, Ger P Reesink & Angela Terrill. 2002. The East Papuan languages: A preliminary typological appraisal. *Oceanic Linguistics* 41(1). 28–62.

Dyck, Carrie, Julie Brittain & Marguerite MacKenzie. 2006. Northern East Cree accent. In *Proceedings of the 2006 Annual Conference of the Canadian Linguistics Association*, 27–30.

Foley, William A & William A Foley. 1986. *The Papuan languages of New Guinea.* Cambridge University Press.

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.

Hellwig, Birgit. 2019. *A Grammar of Qaqet*, vol. 79 Mouton Grammar Library. De Gruyter Mouton.

Hellwig, Birgit, Carmen Dawuda, Henrike Frye & Steffen Reetz. 2014. The Qaqet Corpus at the Language Archive Cologne. Online: http://hdl.handle.net/11341/00-0000-0000-0000-202A-0@view.

Jaklin, Kornfilt. 1997. Turkish. *London/NY: Routledge* .

Junker, M. 2000. The interactive East Cree reference grammar. http://www.eastcree.org/cree/en/grammar/.

Kaiser, Stefan, Yasuko Ichikawa, Noriko Kobayashi & Hilofumi Yamamoto. 2013. *Japanese: A comprehensive grammar.* Routledge.

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.

Lindström, Eva & Bert Remijsen. 2005. Aspects of the prosody of Kuot, a language where intonation ignores stress. *Linguistics* 43(4). 839–870.

Lindström, Eva, Angela Terrill, Ger Reesink & Michael Dunn. 2007. The languages of island Melanesia. *Genes, Language, and Culture History in the Southwest Pacific: A Synthesis. Oxford University Press, Oxford* 118–140.

Macdonald, Roderick Ross. 1976. *Indonesian reference grammar.* Georgetown University Press.

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

Maimonides, Moses, David Bolles & Alejandra Bolles. 1989. *A grammar of the Yucatecan Mayan language* 2. Labyrinthos.

Merlan, Francesca & Alan Rumsey. 1991. *Ku Waru: Language and Segmentary Politics in the Western Nebilyer Valley.* Cambridge University Press.

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, Anna Jancso & Sabine Stoll. 2019a. ACQDIV database aggregation pipeline (Version 1.0.0). Zenodo. Online: http://doi.org/10.5281/zenodo.3558643.

Moran, Steven, Anna Jancso & Sabine Stoll. 2019b. ACQDIV database (public) (Version 1.0.0) [Data set]. Zenodo. Online: http://doi.org/10.5281/zenodo.3558641.

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.

Reesink, Ger. 2005. Sulka of East New Britain: a mixture of Oceanic and Papuan traits. *Oceanic Linguistics* 145–193.

Ross, Malcolm D. 1996. *8 Contact-induced Change and the Comparative.* Oxford University Press on Demand.

Rumsey, Alan, Andrew Noma, Lauren Reed, Naomi Peck, Charlotte van Tongeren & Stephanie Yam. 2019. ACQDIV portion of the Ku Waru Child Language Socialization Study (KWCLSS). Forthcoming in PARADISEC.

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.

Sneddon, James Neil, K Alexander Adelaar, Dwi N Djenar & Michael Ewing. 2012. *Indonesian: A comprehensive grammar.* Routledge.

Stebbins, Tonya et al. 2009. The Papuan languages of the Eastern Bismarcks: migration, origins and connections. In Bethwyn Evans (ed.), *Discovering history through language. Papers in honour of Malcolm Ross*, 223–243. Canberra: Pacific Linguistics,.

Stebbins, Tonya N et al. 2011. *Mali (Baining) grammar.* Pacific linguistics, Research School of Pacific and Asian Studies, The ….

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, 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. Unpublished. 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.

Swift, Mary D. 2008. *Time in child Inuktitut: A developmental study of an Eskimo-Aleut language*, vol. 24. Walter de Gruyter.

Terrill, Angela. 2002. Systems of nominal classification in East Papuan languages. *Oceanic linguistics* 63–88.

Theakston, A. L., E. V. M. Lieven, J. M. Pine & C. F. Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language* 28. 127–152.