Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 156–165
Marseille, 11–16 May 2020

The ACQDIV Corpus Database and Aggregation Pipeline

Anna Jancso*, Steven Moran*, Sabine Stoll
Department of Comparative Language Science & Center for the Interdisciplinary Study of Language Evolution (ISLE)
University of Zurich
Plattenstrasse 54
8032 Zurich, Switzerland
{anna.jancso, steven.moran, sabine.stoll}@uzh.ch
*equal contributors

Abstract

We present the ACQDIV corpus database and aggregation pipeline, a tool developed as part of the European Research Council (ERC) funded project ACQDIV, which aims to identify the universal cognitive processes that allow children to acquire any language. The corpus database represents 15 corpora from 14 typologically maximally diverse languages. Here we give an overview of the project, database, and our extensible software package for adding more corpora to the current language sample. Lastly, we discuss how we use the corpus database to mine for universal patterns in child language acquisition corpora and we describe avenues for future research.

Keywords: language acquisition, typological databases, corpus linguistics, TalkBank, Toolbox

1. Overview

In this paper, we present the ACQDIV corpus database and our extensible corpus aggregation pipeline that generates the database from disparate corpus input formats. We begin by describing in Section 2 the motivations behind the database's compilation for the ERC-funded project "ACQuisition processes in maximally DIVerse languages: min(d)ing the ambient language".[1] One of the most pressing questions in cognitive science is: how is a child able to acquire any of the world's 7000 or so extremely diverse languages? Thus, the goal of the five-year ACQDIV project is to identify universal cognitive processes that enable language acquisition despite the substantial cross-linguistic variation found in the world's languages. To create the language sample needed to study worldwide linguistic diversity, ten typologically very different languages were identified to simulate maximal variation in grammar (Stoll and Bickel, 2013). In Section 3, we describe the languages and corpus formats in detail, including additional language acquisition corpora that we have added through open access sources and international collaborations.

Our database aggregation pipeline is called ACQDIV and is written as a PYTHON (Python Software Foundation, 2018) package and available on PYPI.[2] The input data formats that we support are the two most-often used corpus encoding formats:

- CHILDES CHAT
- SIL TOOLBOX

In short, CHILDES CHAT is an encoding defined as part of TalkBank, an open-source system for sharing and studying conversational interactions.[3] In our work, we use several of the corpora from the Child Language Data Exchange System (CHILDES), a subset of TalkBank (MacWhinney, 2000). We also can process input corpora in TOOLBOX, an encoding standard developed by SIL International.[4]

In Section 4, we explain how we extract, transform, and load these different corpus formats and corpus-specific annotation schemes into a syntactically and semantically interoperable database (Moran et al., 2016). In Section 5, we provide instructions for how programmers can extend our PYTHON package to include new corpora from TOOLBOX files or from the more than 130 corpora encoded in CHILDES CHAT, which are openly available in TalkBank. In brief, this process consists of creating an INI configuration file to identify the metadata associated with each language and corpus, and then mapping categories, e.g. mappings between corpus-specific part-of-speech labels or morphological categories and glosses, into cross-linguistic standardized annotation sets that include Universal Dependencies (Nivre, 2016) and the Leipzig Glossing Rules (Comrie et al., 2015).

Finally, in Section 6 we briefly discuss some of the research that we have conducted with the ACQDIV corpus database and avenues for future research on one of the most pressing questions in cognitive science: what cognitive mechanisms enable children to learn any of the world's 7000 or so languages?

2. Background

The goal of the ACQDIV project is to identify universal cognitive processes that enable language acquisition despite the substantial cross-linguistic variation found in the world's languages. To create the language sample needed to study typological diversity, ten grammatically maximally different languages were identified by applying a fuzzy clustering algorithm that takes as input a set of languages and their typological feature values (e.g. grammatical case, inflectional categories, nominal synthesis)

[1] https://www.acqdiv.uzh.ch
[2] https://pypi.org/project/acqdiv
[3] https://talkbank.org
[4] https://www.sil.org

(Stoll and Bickel, 2013). The algorithm is applied to language data from thousands of languages encoded in two broad-coverage typological databases: the World Atlas of Linguistic Structures (Dryer and Haspelmath, 2013) and AUTOTYP (Bickel et al., 2017). It outputs five clusters of maximally diverse languages based on a dozen typological variables that are known to be encoded in a variety of different ways cross-linguistically. To ensure diversity, two languages from each cluster were chosen.

For nine of these languages, there exists open-source or privately-owned longitudinal child language acquisition corpora. A longitudinal language acquisition corpus consists of some number of sessions, i.e. a so-called "target child" is recorded in his or her environment, e.g. for an hour a month over several years, and these sessions are then transcribed and linguistically annotated. For the tenth language, our team is in the process of collecting, transcribing and annotating a longitudinal child language acquisition corpus for Dënesųłiné, an endangered language spoken by the Chipewyan people of northwestern Canada. Additionally, we have added several more open source corpora from TalkBank and from new international collaborators, who encode their data in TOOLBOX, to leverage our corpus database pipeline and because researchers are interested in having their data in a format that is interoperable with the existing corpora in ACQDIV. Such a broad and accessible corpus database allows researchers to ask novel questions about child-directed speech in culturally and linguistically diverse languages.

3. The Language Sample

Currently, we process 15 different child language acquisition corpora, which represent 14 languages. Table 1 shows general information and statistics about these corpora.

Our data sample represents vastly different morphological systems, ranging from isolating to polysynthetic languages. For example, in isolating languages like Indonesian, words and morphemes are mainly in a one-to-one relationship, as shown in Example (1). This is in stark contrast to polysynthetic languages like Cree, where an entire sentence can be analyzed as a single word made up of many morphemes, as illustrated in Example (2).

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

(2) Chi-wâp-iht-â-n â kâ-pushch-ishk-iw-â-t.
2-light-by.head-TR.INAN.NON3-2SG>0 Q 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)

Our sample represents not only vastly different morphological systems, but these languages are spoken by culturally different peoples, by vastly different community sizes, and in geographically diverse areas. Moreover, the languages represent a genealogically diverse sample of language families and they include data from different levels of UNESCO's language endangerment status index (*definitely endangered*, *vulnerable*, and *safe*) (Moseley, 2010). Computationally, the sample also represents a considerable amount of data on lesser-resourced and poorly described languages.

Due to the nature of data collection and dissemination of child language corpora from very different cultures, some of the corpora in the full ACQDIV sample are not open source. However, there is restricted access to Chintang (Stoll et al., Unpublished), Inuktitut (Allen, Unpublished), Russian (Stoll and Meyer, 2008), Tuatschin, Turkish (Küntay et al., Unpublished), and Yucatec (Pfeiler, Unpublished). Access is made available via the ACQDIV corpus database terms of agreement.[5]

In accordance with the TalkBank's code of conduct, 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 and Nisisawa, 2009; Miyata and Nisisawa, 2010; Nisisawa and Miyata, 2009; Nisisawa and Miyata, 2010), Japanese Miyata (Miyata, 2004a; Miyata, 2004b; Miyata, 2004c), Ku Waru (Rumsey et al., 2019), Nungon (Sarvasy, 2017), and Sesotho (Demuth, 2015). The ACQDIV database (public version) is available on Zenodo (Moran et al., 2019).

4. Data Extraction and Aggregation

Our source code for processing the corpora is available online in a GitHub repository (Moran and Jancso, 2019). The ACQDIV aggregation pipeline's workflow follows the fork-and-pull model. We have written our code base in PYTHON and we release the package via PYPI. It can then be run with the following commands.[6]

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
```

[5] https://www.acqdiv.uzh.ch/en/resources.html (last accessed 29-02-2020).

[6] See the ACQDIV GitHub repository for comprehensive instructions.

Corpus	ISO 639-3	# Recording sessions	# Words	# Morphemes	Status	Population	Macroarea
Chintang	ctn	477	987673	1589827	definitely endangered	3.7K	Eurasia
Cree	cre	25	44751	11686	vulnerable	87K	North America
English_Manchester1	eng	804	2016043	2098914	safe	328M	Eurasia
Indonesian	ind	997	2489329	2725605	safe	23.2M	Papunesia
Inuktitut	ike	77	71191	91685	vulnerable	34.5K	North America
Japanese_MiiPro	jpn	192	1011670	1009599	safe	128M	Eurasia
Japanese_Miyata	jpn	213	373021	372495	safe	128M	Eurasia
Ku_Waru	mux	9	65723	92438	safe	41K	Papunesia
Nungon	yuw	4	19659	19262	safe	1.7K	Papunesia
Qaqet	byx	106	56239	105165	definitely endangered	15K	Papunesia
Russian	rus	450	2029704	NA	safe	166.2M	Eurasia
Sesotho	sot	69	177963	330009	safe	5.6M	Africa
Tuatschin	roh	51	118310	NA	vulnerable	1.2K	Eurasia
Turkish	tur	373	1120077	215822	safe	71M	Eurasia
Yucatec	yua	234	262382	171633	safe	766K	North America

Table 1: ACQDIV corpora

You 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 gives us a Digital Object Identifier (DOI) for reference. This allows users to cite particular versions of the pipeline and the database for scientific replicability.

Our data extraction and aggregation pipeline accepts CHILDES CHAT and SIL TOOLBOX as corpus input formats. CHILDES is the child language acquisition component of the TalkBank system, which is an open-source system for sharing and studying conversational interactions. CHAT is the specification used to encode and analyze child-directed and child-surrounding speech from adults. Each file in CHAT represents one recording session. Files are encoded in Unicode UTF-8 plain text according to a set of (semi-loose) specifications.[7] An example of a CHAT file is given in Figure 1.

Each speaker utterance is prefixed with an asterisk. Each utterance may then have optional tiers below it that encode information such as the morphological gloss (%gls), morphological annotation (%xcod), and the translation (e.g. %eng for English). Note, however, that these tier labels may differ from corpus creator-to-corpus creator.

The second corpus format that our pipeline accepts as an input format is SIL International's TOOLBOX format. TOOLBOX is a data management tool for collecting and analyzing lexical data and interlinear text. Field linguists often use it for creating a morphologically annotated lexicon. Corpus linguists also use this format for its ease of encoding recording sessions through conversational interactions, where each utterance turn is encoded via several user-defined idiosyncratic tiers (e.g. \tx for text; \gw for gloss; \ps for part-of-speech), each of which is separated by a blank line. As in CHAT, every file represents a recording session (file names differ from corpus to corpus, but often include the recording data in the file name). TOOLBOX files are encoded in plain text Unicode UTF-8 format. An example is given in Figure 2.

```
@UTF8
@PID:   11312/c-00027872-1
@Begin
@Languages: sot
@Participants:  MTS Mantso Sister ,
    KAT Katherine_Demuth Investigator , JUL
    Julia Grandmother , MAR Maria Mother , NEU Neuoe_Cousin
    Relative , CHI Tsebo Target_Child , HLE Hleso Brother
@Options:   bullets
@ID:    sot|Demuth|MTS|10;00.||||Sister|||
@ID:    sot|Demuth|KAT|||||Investigator|||
@ID:    sot|Demuth|JUL|||||Grandmother|||
@ID:    sot|Demuth|MAR|||||Mother|||
@ID:    sot|Demuth|NEU|2;04.||||Relative|||
@ID:    sot|Demuth|CHI|3;08.00||||Target_Child|||
@ID:    sot|Demuth|HLE|6;00.||||Brother|||
@Media: 030800cd, audio
*CHI:   ke disa kae moekesaize le moroho ? ¿0_9911¿
%gls:   ke-di-is-a kae eksaese le mo-roho ?
%xcod:  sm1s-t^p_om10-v^go/c-m^in wh exercise_book(9 , 10)
cj n^3-greens(3, 4) ?
%eng:   Where do I take the exercise and the vegetables ?
*MAR:   e ? ¿9200_10376¿
%gls:   e ?
%xcod:  wh ?
%eng:   What ?
```

Figure 1: Example of CHAT format (Demuth, 2015)

Both CHAT and TOOLBOX files have accompanying metadata information that summarizes when, where and how the recordings were made, and also provides information about the speakers, such as who is present in the recording, what role do they play (e.g. target child, parent, caregiver) and how old they are. For CHAT, the metadata is provided at the top of each file in terms of key-value pairs prefixed with "@", as shown in Figure 1. For TOOLBOX, metadata is kept in separate files, typically in the XML formats called

[7] The full CHAT specification is given in the TalkBank manual which we discuss below.

\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

Figure 2: Example of TOOLBOX format (Stoll et al., Unpublished)

IMDI (Broeder and Wittenburg, 2006) and more recently its newer version CMDI (Broeder et al., 2012).

With our corpus database aggregation pipeline, illustrated in Figure 3, we bring together the CHAT and TOOLBOX corpora and their metadata into a single unified relational database format. We have chosen a relational database because it allows us to easily analyze parent and child utterances for statistical patterns across morphologically different languages. Also, our aggregation pipeline creates syntactically interoperable data from the different corpus input formats by unifying them into the relational database. During this transformation, we also make the idiosyncratic annotation schemes in the different input corpora semantically interoperable by re-encoding all morphological labels into unified and standardized vocabularies.[8]

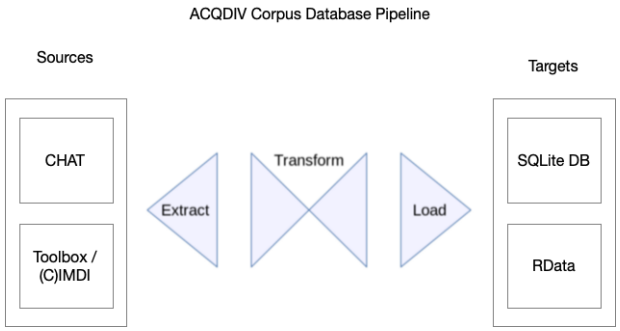

Figure 3: ACQDIV Corpus Database Pipeline

Our current output formats are a relational SQLITE database and an R data object that includes data frames for each table and view in the database. We chose SQLITE because it requires no special setup or configuration. Since our database is not very large (∼1.5GB), and in most cases will be used by individual users, we do not expect scalability issues. However, as we use object-relational mapping (ORM), users can also easily load the data into other database management systems. We also provide an R data object as an output format because many linguists prefer to use R for data analysis (R Core Team, 2018).

Our relational database schema is comprised of seven tables, as shown in Figure 4. The main table is the CORPORA table, which is related in a cascading one-to-many relationship with the SESSIONS (recordings) table, each of which has multiple UTTERANCES, WORDS, and MORPHEMES.[9]

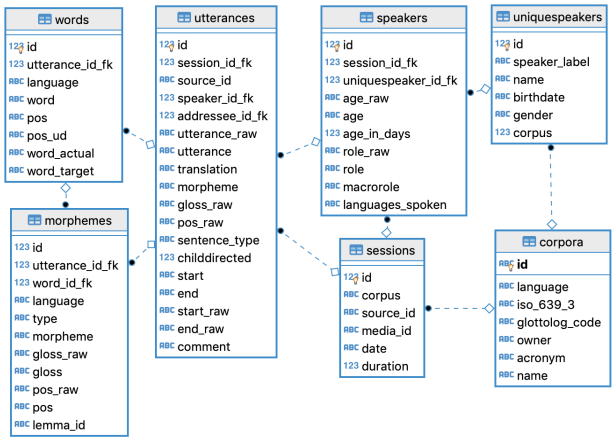

Figure 4: ACQDIV Corpus Database Schema

For each corpus, regardless of format, we extend a general corpus parser, written in PYTHON, from which we parse a set of the input data (recording session and speakers metadata, information about the utterances spoken, including attributes such as the utterance duration, its words, morphemes, morphological annotation, parts-of-speech, etc.). However, given the current limitation of tools for accessing multiple CHAT (and TOOLBOX) corpora at the same time, and given the wide range of idiosyncratic morphological annotation schemes encoded in them, our aggregation pipeline provides users the ability to transform CHAT and TOOLBOX files into a single accessible relational database format.

Given the nature of our typologically diverse language sample, it should not be surprising that some language-specific categories must be broadened or narrowed to attain some level of semantic interoperability. Consider for example the rich noun class system in Sesotho (a Bantu language of South Africa), which has more than 20 different plural strategies, each labeled differently by the corpus collectors (in Figure 1 note the morphological code "n^3" that denotes noun class three). We chose to not only keep this rich set of annotations in our database (e.g. MORPHEMES.GLOSS_RAW), but to also collapse these different annotations into a single label, NOUN, so that the Sesotho data can be searched (along with all

[8] A next step towards semantic interoperability would be to encode the language data, and the relationships within the data, with an ontology. We leave this for future research.

[9] A very detailed description of the database, its design, problems encountered and solutions implemented for the corpus-specific input formats, is provided in our comprehensive ACQDIV Corpus Database Manual (Moran et al., 2019b).

other languages in the database) with the tag set from the Universal Dependencies for nouns, verbs, etc.

This example highlights the most problematic linguistic challenge that we encountered in the development of the ACQDIV database and corpus aggregation pipeline. That is, the same morphological phenomena may be encoded differently per corpus. However, it is not only the labels that differ from corpus to corpus, the actual format for encoding the morphological tier is highly heterogeneous. This requires that we build corpus-specific parsers for the morphology tiers. Particularly problematic are the CHAT corpora because of the flexibility that this corpus encoding format allows its users – basically anything goes. In our TOOLBOX corpora, 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 (see Figures 1 and 2 above). TOOLBOX corpora, and to some extent the Cree and Sesotho 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 CHAT corpora encode the morpheme, gloss, and part-of-speech labels on the same tier. For example, in the Japanese MiiPro corpus (Miyata and Nisisawa, 2009; Miyata and Nisisawa, 2010; Nisisawa and Miyata, 2009; Nisisawa and Miyata, 2010):

```
*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 CHAT corpora. For example, in many CHAT corpora affixes are the only annotated glosses and for stems there may be no gloss at all.

Furthermore, morphemes can 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 (see the ACQDIV database corpus manual for a full explication for each corpus (Moran et al., 2019b)).

Another important aspect of our corpus aggregation pipeline involves cleaning the input data. This process includes:

- Removing any elements not of interest to our project, e.g. pause annotations, paralinguistic elements like coughing and sneezing, that are only coded in a few of the corpora, so we cannot compare them across the languages in the sample
- Reducing misalignments: certain elements only occur on certain tiers but not on others (e.g. punctuation might be encoded on the word tier, but not on the morphology tiers; word repetitions may occur on the word tier, but not morphology tiers)
- Unifying glosses, part-of-speech tags, speaker metadata, utterance timestamps, date formats, corpus tier names (e.g. "xtrn" vs. "xmor" in Japanese MiiPro) as to enable a consistent cross-linguistic search
- Removing punctuation (and inferring from it what type of utterance it is, e.g. "!" imperatives, "?" interrogatives)

During the cleaning procedure, we also undertake a routine for speaker unification. That is, we identify the same speakers across different recording sessions, so that we can populate a UNIQUESPEAKERS table in our database. This involves a certain amount of metadata correction (e.g. fixing typos in the input) and inference:

- Speaker name typos, e.g. "Khetheng"' vs. "Khethang" in Sesotho
- Speaker labels across corpora are generic, e.g. "CHI" for target children in CHAT corpora, so we must identify each individual
- Speaker names may be different from session-to-session, e.g. "Asato" is the same person as "Asatokun" in Japanese MiiPro

We must also undertake speaker role unification, which involves:

- Target child identification because some recording sessions have assigned the role "Child" to all children including the target child which should receive the role "Target_Child"
- Text normalization for different annotator labels for the same person, e.g. "Mother" vs. "mother", "Mother's brother" vs. "Uncle"

Part-of-speech and gloss unification is also necessary, i.e.:

- Many corpora do not use the standardized Leipzig Glossing Rules or Universal Dependency labels, but instead their own idiosyncratic conventions, e.g. "PRS" vs. "PRES", "PRSP", "PRESP", etc., so we identify and unify these different labeling schemes
- A gloss or part-of-speech can be mapped to several glosses or part-of-speech tags, so we describe and implement our decisions on what is mapped to what (see the ACQDIV corpus manual for examples and documentation)

For missing data, we also try to infer as much as information as possible, including but not limited to:

- The macrorole of the speaker (i.e. target child, child, adult) as based on age, e.g. speakers over 12 years old are labeled adults
- The speaker's gender, which we infer from speaker roles as labeled in the corpora, e.g. mother is mapped to female
- A speaker's age is inferred from their birth date and recording session date

Finally, we have found that misalignment in the various corpora between the word, morpheme, and annotation tiers poses a serious problem. This is why we link morphemes not only to the word, but also to the complete utterance in our database schema. Again, we document our decisions in the ACQDIV corpus manual, but to sum up, we provide access to the word and morpheme levels, in both cases of when they align and when they do not align.

Running the pipeline on all corpora, which includes 4081 recording sessions and 3580 metadata files, takes 35 minutes and consumes less than 100MB memory on an Intel Core i5-7200U processor with 8GB RAM. Thus, our system processes on average around 2 recording sessions with 1925 utterances and 5218 words per second.

5. Adding New Corpora

Our extensible approach to coding our corpus aggregation pipeline has been motivated by the desire to make our code base available to other developers and to make our decisions regarding the ACQDIV corpus database clear for users. For example, this includes unifying different annotation schemes, within a database design that unifies syntactically the different corpus input formats. These procedures were an extremely time-consuming endeavor, so we have tried to develop a workflow for easily adding new and diverse corpora. In this section, we describe how to add new corpora to the ACQDIV database by extending the aggregation pipeline. We give users and developers a brief overview of our current corpus input parsers and the components needed to integrate new data sources.

TalkBank contains a range of resources that are transcribed, richly annotated, and aligned with audio and video recordings. Currently, there are over 130 different corpora representing 26 languages. These materials are publicly available online. As such, we have developed the ACQDIV corpus database aggregation pipeline so that developers can add these, as well as corpora encoded in TOOLBOX, to the resulting output database. Detailed instructions on how to add new corpora are described online in the GitHub repository. Here we provide 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 CHAT and TOOLBOX is available in the repository. Second, create a new Python package under `parsers/corpora/main/<corpus_name>` with the following classes:

- Reader
- Cleaner
- SessionParser
- CorpusParser

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 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).

The cost for adding a new corpus depends on many factors, such as whether the developer is already familiar with the pipeline, but also how much cleaning and inference mechanisms have to be implemented for that corpus. Generally speaking, TOOLBOX corpora are faster to integrate because they usually follow the Leipzig Glossing Rules, while CHAT corpora often use their own encoding rules for their morphology. We found that adding a new corpus takes between a couple of hours to several days. We estimate that a developer unfamiliar with the pipeline might need up to one week to add a new corpus.

6. Research outcomes

Once the ACQDIV corpus database aggregation pipeline and resulting interoperable database were completed, the main goal of the ACQDIV project was to investigate whether there are universal patterns in the input to children cross-linguistically. One area to investigate is child-directed speech because it has been shown to facilitate language learning through various structural features and frequently occurring patterns in the input to children. These include statistical regularities of isolated words (Lew-Williams et al., 2011), adjacent dependencies (Redington et al., 1998), non-adjacent dependencies (Mintz, 2003), and sentence frames (Fernald and Hurtado, 2006).

In fact, using the ACQDIV database we have found striking similarities across morphologically very different languages in child-directed and child-surrounded speech. Furthermore, the cultural settings of languages in our sample vary from WEIRD (western, educated, industrialized, rich and democratic) societies (Henrich et al., 2010), where the norm is directed interaction between caregivers and children (often using so-called *motherese*), to cultures where children often get more input from their peers than their parents. We highlight in particular three studies that we have published using the ACQDIV corpus database as input for our analyses.

First, we mined the ACQDIV corpus database for discontinuous, but frequently occurring repetitive patterns, and we found language-independent anchor points that predict with high accuracy parts-of-speech. These patterns, known in the literature as *frequent frames* (Mintz et al., 2002; Mintz, 2003), occur frequently in all languages in our sample (Moran et al., 2018a). An example of a frequent frame is:

- I like you
- I hate you
- I hear you
- I love you
- I and you
- I not you

The frame is a nonadjacent dependency between the sequence of three linguistic elements, e.g. the form A_B_C, in which A and C (here "I" and "you") predict information about B. As shown above, the intervening word is most likely to be a verb in a large corpus of English data. Frames can be calculated at either the word or morpheme levels. For example, consider the morphological frame:

- She **is** sleep**ing**

This frame exemplifies morphosyntactic agreement in English and only verbs can appear between the auxiliary verb *is* and the progressive suffix *-ing*. This particular dependency signals the grammatical class of the intervening element, i.e. **verb**. In this case, the intervening morpheme, in English, is always a verb.

We mined the ACQDIV corpus database for discontinuous frames and found that in our diverse language sample, frequently occurring frames are anchor points for accurately identifying the part-of-speech of the intermediate element at the morphology level across all languages in the sample. This shows that there are statistically frequent and reoccurring patterns in the input to children, regardless of the typological characteristics of the language, which may help the learner identify parts-of-speech (Moran et al., 2018a).

In another recent study (Moran et al., 2019a), we identified a cross-linguistic universal pattern in the input to children known in the psycholinguistics literature as *variation sets*. An example of a variation set in English is (Küntay and Slobin, 2002):

- Who did we **see** when we went out shopping today?
- Who did we **see**?
- Who did we **see** in the store?
- Who did we **see** today?
- When we went out shopping, who did we **see**?

In this example, the utterances, in close proximity, are anchored around the word *see* and the question "Who did you see?". To get the message across, the caregiver repeats and rephrases the same question several times, presumably trying to keep the child's attention. Hence, variation sets are repetitions of words, specifically nouns and verbs, in short sequences of interactions. They are thought to provide the child not only with repetitive information, but variation sets also allow children to learn about the different forms that a lexical item can appear in. This is particularly interesting for morphologically-rich languages, e.g. Turkish (Küntay and Slobin, 2002):

(3) **Ver** el-ler-in-i.
give hand-PL-POSS.2SG-ACC
'Give (me) your hands.'

(4) El-ler-in-i **ver**-ir-mi-sin.
hand-PL-POSS.2SG-ACC **give**-AOR-Q-2SG
'Will you give (me) your hands?'

(5) El-ler-in-i **ver**.
hand-PL-POSS.2SG-ACC **give**
'Give (me) your hands.'

To understand the actual distribution of these variation sets, we again mined the ACQDIV database and found that all child directed speech (for all languages in our sample) contains variation sets (Moran et al., 2019a). These repetitions of words in short sequences of interaction provide children not only with lexical repetition, but they are also believed to assist them in learning different lexical forms in morphologically-rich languages.

Lastly, in recent work we model language specific child-directed speech as lexical adjacency network graphs. We then identified that in child-directed speech, regardless of the language in our sample, each corpus shows the so-called *small world* property associated with networks, i.e. low average path length and high clustering coefficient in the network. We postulate that these global structural characteristics reflect principles of self-organization of the lexicon and they may facilitate cognitive processing and learning (Moran et al., 2018b). Our findings support the emergent view of language development, i.e. language acquisition is a piecemeal process that takes several stages and many years to master, on which the scaffolding of learning words and constructions and their interrelations is paramount.

7. Future directions

For future research, we plan to increase the coverage of the ACQDIV corpus database by extending it with more corpora from TalkBank and from private corpus owners. For example, we have started to extend the aggregation pipeline with several of the phonetically transcribed corpora in PhonBank (Durand et al., 2013). PhonBank includes child language acquisition corpora with phonetic transcriptions. We are converting the individual transcription practices

into a single unified International Phonetic Alphabet encoded in the Unicode Standard via orthography profiles (Moran and Cysouw, 2018). We are now investigating whether there are universal patterns in the input to children at the phonological level (Moran and Stoll, 2018).

As we described above, the ACQDIV corpus database aggregation pipeline can be extended to include any CHAT or TOOLBOX formatted corpus. Additionally, linguistic annotations – to the extent that they can be, cf. Haspelmath (2010a; 2010b) and Newmeyer (2010) – are made semantically comparable by using our mappings tables with annotation labels from the Universal Dependencies and the Leipzig Glossing Rules. As such, TalkBank also includes language data from aphasics, second language acquisition, conversation analysis, and classroom language learning. Since these resources are encoded in CHAT, they can be readily parsed into the ACQDIV unified database format for analysis and research on topics other than child language acquisition. Additionally, given that the Universal Dependencies framework provides NLP tools for parsing and analyzing treebanks, we imagine such tools being applied to (at least some of) the corpora in our database and in CHILDES, so that these datasets can be lemmatized and extended with dependency parses – for example with SPACY (Honnibal and Montani, 2017). This information can then be integrated back into the ACQDIV corpus database, so that new research questions can be asked.

Lastly we note that there are several other tools and software packages available for interacting with the CHILDES data, including TalkBank's CLAN program (allows regular expression search and does basic statistics like mean length of utterance); the NLTK (Loper and Bird, 2002) which has an API for querying the CHILDES XML files; and recently there is CHILDES-DB which has a browsable web application and an API in R (Sanchez et al., 2018).

These tools provide exciting opportunities for cross-linguistic research into language acquisition. Our aim here has also been to provide a rich set of accessible data for research on child language acquisition, both qualitative and quantitative in nature. We believe the data formats that we produce provide the basis for integration with existing projects and the opportunity for creating new and exciting tools for future research.

8. Summary

In this paper, we have given an overview of the ACQDIV corpus database and aggregation pipeline, which is available as a PYTHON package via PYPI. Our ACQDIV software package integrates 15 corpora from 14 typologically maximally diverse languages into a single syntactically and semantically interoperable database. Our aggregation pipeline is designed to be extensible, and as such, can be extended to include corpora from both CHILDES CHAT and SIL TOOLBOX input formats. Here, we described the technological challenges and architectural decisions we have made so that we have been able to create the tools needed to be able to work with corpora encoded in different input formats and with different annotation schemes and encodings.

Given the vast typologically diversity in the language sample that we currently work with, we have developed a workflow that allows users to map morphological labels and annotations into a unified cross-linguistic format, using Universal Dependencies and the Leipzig Glossing Rules. Given this rich resource of longitudinal child language acquisition corpora from a diverse set of the world's languages, we discussed briefly the kinds of research on universal patterns in the input to children that we have identified. Using the ACQDIV corpus database, our research shows that there is a repertoire of universal distributional patterns in the input to children, that it is a cross-linguistic phenomenon, and that it needs further investigation to help identify the cognitive processes that underlie child language acquisition.

9. Acknowledgements

The research leading to this paper was part of the project 'Acquisition processes in maximally diverse languages: Min(ding) the ambient languages (ACQDIV)' that has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7-2007-2013) (Grant agreement No. 615988; PI Sabine Stoll). We gratefully acknowledge Robert Forkel and Sebastian Bank for comments and feedback on the ACQDIV software package and Janis Goldzycher for contributing to the code base. Also many thanks to three anonymous reviewers.

10. Author Contributions

SM & AJ designed the database, developed the software package, and wrote the paper. SST designed the ACQDIV project and supervised the project. All authors read and approved the final manuscript.

11. Bibliographical References

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., Bierkandt, L., Zúñiga, F., and Lowe, J. B. (2017). The AUTOTYP typological databases. version 0.1.0. Online: https://github.com/autotyp/autotyp-data/tree/0.1.0.

Broeder, D. and Wittenburg, P. (2006). The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2):119–132.

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. (2012). CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.

Comrie, B., Haspelmath, M., and Bickel, B. (2015). The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Online: `https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf`.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Durand, J., Gut, U., Kristoffersen, G., Rose, Y., and MacWhinney, B. (2013). *The PhonBank ProjectData and Software-Assisted Methods for the Study of Phonology and Phonological Development*. Oxford University Press.

Fernald, A. and Hurtado, N. (2006). Names in frames: Infants interpret words in sentence frames faster than words in isolation. *Developmental science*, 9(3).

Haspelmath, M. (2010a). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Haspelmath, M. (2010b). The interplay between comparative concepts and descriptive categories (reply to Newmeyer). *Language*, 86(3):696–699.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Küntay, A. and Slobin, D. I. (2002). Putting interaction back into child language: Examples from Turkish. *Psychology of Language and Communication*, 6(1):5–14.

Lew-Williams, C., Pelucchi, B., and Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, 14(6):1323–1329.

Loper, E. and Bird, S. (2002). NLTK: the natural language toolkit. *arXiv preprint cs/0205028*.

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

Mintz, T. H., Newport, E. L., and Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26:393–424.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Moran, S. and Cysouw, M. (2018). *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press.

Moran, S. and Jancso, A. (2019). ACQDIV database aggregation pipeline (version 1.0.0). Zenodo. Online: `http://doi.org/10.5281/zenodo.3558643`.

Moran, S. and Stoll, S. (2018). Worldwide frequency of phonemes predicts their age of acquisition. In *Poster presented at the 42nd Boston University Conference on Language Development (BUCLD), November 3–5, Boston, USA*.

Moran, S., Schikowski, R., Pajović, D., Hysi, C., and Stoll, S. (2016). The ACQDIV database: Min(d)ing the ambient language. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Moran, S., Blasi, D. E., Schikowski, R., Küntay, A. C., Pfeiler, B., Allen, S., and Stoll, S. (2018a). A universal cue for grammatical categories in the input to children: frequent frames. *Cognition*, 175:131–140.

Moran, S., Pajović, D., and Stoll, S. (2018b). Cross-linguistically small world networks are ubiquitous in child-directed speech. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4100–4105, Paris, France, May 7–12. European Language Resources Association (ELRA).

Moran, S., Lester, N. A., Gordon, H., Küntay, A., Pfeiler, B., Allen, S., and Stoll, S. (2019a). Variation sets in maximally diverse languages. In *In Proceedings 43rd annual Boston University Conference on Language Development (BUCLD). November 2–4, Boston, USA. Cascadilla Press.*

Moran, S., Schikowski, R., and Stoll, S. (2019b). ACQDIV corpus database user manual. Online: `https://github.com/acqdiv/corpus_manual`.

Moseley, C. (2010). Atlas of the world's languages in danger (3rd edn. ed.). Online: `http://www.unesco.org/languages-atlas/`.

Newmeyer, F. J. (2010). On comparative concepts and descriptive categories: a reply to Haspelmath. *Language*, 86(3):688–695.

Python Software Foundation, (2018). *The Python Language Reference, version 3.7*.

R Core Team, (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redington, M., Chater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K. E., Yurovsky, D., and Frank, M. C. (2018). *childes-db: a flexible and reproducible interface to the Child Language Data Exchange System*. PsyArXiv, Apr.

Stoll, S. and Bickel, B. (2013). Capturing diversity in language acquisition research. *Language Typology and Historical Contingency: In Honor of Johanna Nichols. Amsterdam: John Benjamins*, pages 195–216.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., and 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.

12. Language Resource References

Allen, Shanley. (Unpublished). *Allen Inuktitut child language corpus*.

Brittain, Julie. (2015). *Corpus of the Chisasibi Child Language Acquisition Study (CCLAS)*.

Demuth, Katherine. (2015). *Demuth Sesotho corpus*.

Küntay, Aylin C. and Kocbas, Dilara and Sabri Tasci, Süleyman. (Unpublished). *Koc University Longitudinal Language Development Database (KULLD) on lan-*

guage acquisition of 8 children from 8 to 36 months of age.

Miyata, Susanne and Nisisawa, Hiro Yuki. (2009). *MiiPro - Asato Corpus*. Talkbank.

Miyata, Susanne and Nisisawa, Hiro Yuki. (2010). *MiiPro - Tomito Corpus*. Talkbank.

Miyata, Susanne. (2004a). *Aki Corpus*. Talkbank.

Miyata, Susanne. (2004b). *Ryo Corpus*. Talkbank.

Miyata, Susanne. (2004c). *Tai Corpus*. Talkbank.

Moran, Steven and Jancso, Anna and Stoll, Sabine. (2019). *ACQDIV database (public) (Version 1.0.0) [Data set].*

Nisisawa, Hiro Yuki and Miyata, Susanne. (2009). *MiiPro - Nanami Corpus*. Talkbank.

Nisisawa, Hiro Yuki and Miyata, Susanne. (2010). *MiiPro - ArikaM Corpus*. Talkbank.

Nivre, Joakim et al. (2016). *Universal dependencies v1: A multilingual treebank collection*.

Pfeiler, Barbara. (Unpublished). *Pfeiler Yucatec child language corpus*.

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

Sarvasy, Hannah. (2017). *Sarvasy Nungon corpus*.

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

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