The ACQDIV Database: Min(d)ing the Ambient Language

Steven Moran, Robert Schikowski, Danica Pajović, Cazim Hysi and Sabine Stoll
Department of Comparative Linguistics, University of Zurich
Plattenstrasse 54, 8032 Zurich, Switzerland
{steven.moran, robert.schikowski, danica.pajovic, cazim.hysi, sabine.stoll}@uzh.ch

Abstract

One of the most pressing questions in cognitive science remains unanswered: what cognitive mechanisms enable children to learn any of the world's 7000 or so languages? Much discovery has been made with regard to specific learning mechanisms in specific languages, however, given the remarkable diversity of language structures (Evans and Levinson, 2009; Bickel, 2014) the burning question remains: what are the underlying processes that make language acquisition possible, despite substantial cross-linguistic variation in phonology, morphology, syntax, etc.? To investigate these questions, a comprehensive cross-linguistic database of longitudinal child language acquisition corpora from maximally diverse languages has been built.

Keywords: language acquisition, corpus, cross-linguistic, database

1. Introduction

In this paper, we present the ACQDIV database compiled for the European Research Council-funded project: ***Acq**uisition processes in maximally **div**erse languages: min(d)ing the ambient language*.[1] During the project's first year, we have compiled together a set of ten electronic corpora from very linguistically diverse languages. In doing so, we have had to solve issues including:

- Compiling disparately formatted and annotated corpora and making them technologically and linguistically interoperable
- Creating workflows for extracting, transforming and loading the data into standardized formats and interfaces for analysis
- Producing a single unified database structure that allows us to mine patterns in naturally spoken language at the utterance, word and morpheme levels

With this unified cross-linguistic resource in place, our project is beginning to investigate questions involving language acquisition, including:

- Are there underlying patterns in the ambient language (i.e. child-directed speech) that children use in acquiring language?
- If patterns exist, do they verify or refute current theories of child language acquisition (e.g. universalist approaches vs rule-based learning)?
- How do children, say between the ages of 2-3, acquire language-specific features that are known to vary so dramatically across languages in form and content?[2]

We describe the language sample and how it was selected in Section 2. The sample is not only geographically, culturally and linguistically diverse, but the corpora are technologically and theoretically heterogeneous. Therefore we briefly explain our approaches toward interoperability and the resulting database, which to our knowledge, is one of the single largest spoken and unified language corpora to date in Section 3.[3] Given the unified ACQDIV database, we are now starting phase two of the project: addressing the pertinent research questions mentioned above. We show some initial data explorations in Section 4. And we present our preliminary research findings in Section 5.

2. The Language Sample

A fuzzy clustering algorithm that takes as input a set of languages and their typological feature values (e.g. grammatical case, inflection categories, nominal synthesis) and outputs clusters of maximally diverse languages is developed in Stoll and Bickel (2013a). The algorithm is applied to language data from thousands of languages in two broad coverage typological databases – the World Atlas of Linguistic Structures (WALS)[4] and AUTOTYP[5] – and to a dozen typological variables that are known to be encoded in a variety of ways cross-linguistically. Five clusters of maximally diverse languages are identified.
To address the cross-linguistic research questions proposed in the ACQDIV project, we identified two languages from each of the five clusters. For nine languages there already exists richly annotated longitudinal child language acquisition corpora.[6] General information about the languages is

[1] Primary investigator, Prof. Sabine Stoll, Department of Comparative Linguistics, Psycholinguistics Lab, University of Zurich, Switzerland. ERC award period: 2014-2019. Website: `http://www.acqdiv.uzh.ch/`.

[2] We focus on grammatical negation and aspect strategies.

[3] The Child Language Data Exchange System (CHILDES) is the child language acquisition component of TalkBank, a system for sharing and studying different data sources of human and animal communication (MacWhinney, 2000). CHILDES contains transcript and media data from conversations with children from 26 languages, but each corpus is transcribed in the CHAT format and each can only be analyzed individually with the CLAN tool thus limiting cross-linguistic analyses.

[4] `http://wals.info/`

[5] `http://www.autotyp.uzh.ch/`

[6] Led by Dr. Dagmar Jung (U. Zurich), the ACQDIV project

given in Table 1.

	ISO 639-3	Language	Speakers	Classification
1	cre	Cree	87K	Algic
1	chp	Dene	12K	Na-Dene
2	ind	Indonesian	23.2M	Austronesian
2	yua	Yucatec	766K	Mayan
3	ctn	Chintang	3.7K	Sino-Tibetan
3	ike	Inuktitut	34.5K	Eskimo-Aleut
4	rus	Russian	166.2M	Indo-European
4	sot	Sesotho	5.6M	Niger-Congo
5	jpn	Japanese	128M	Japanese
5	tur	Turkish	71M	Altaic

Table 1: Language sample

These languages differ to a great extent linguistically in their number of contrastive sounds (cf. Moran et al. (2014)), morphology (e.g. ranging from isolating to agglutinating to polysynthetic) and syntactic structure (e.g. constituent and word orders). For example, in some languages words represent full phrases, in others the word and morpheme are nearly synonymous. Compare utterances in Indonesian and Cree, two languages in our dataset:

(1) O, Ei lagi minum susu.
oh Ei more drink milk
'Oh, Ei is drinking more milk.'

(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.'

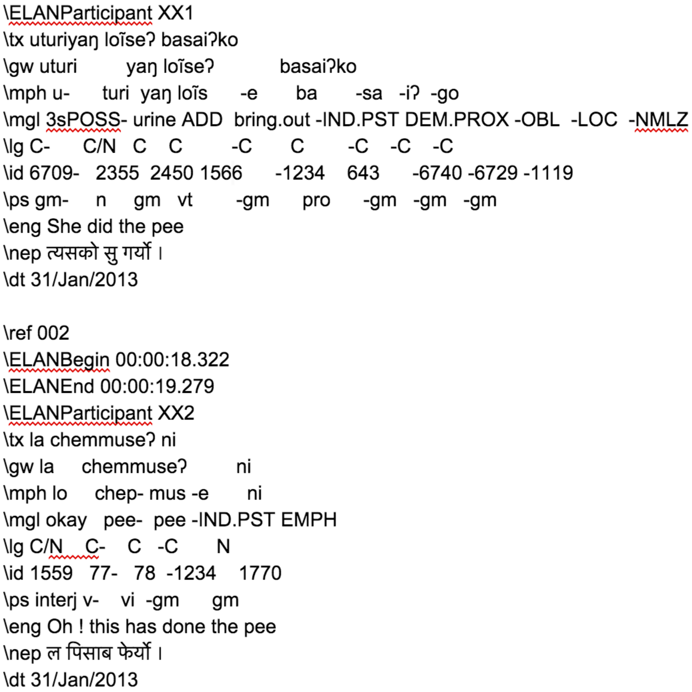
Indonesian is an example of a language with a fairly low degree of synthesis, whereas Cree belongs to one of the most genuinely polysynthetic languages of the world (and features both noun incorporation and polypartite stems). Clearly the frequency in which Indonesian or Cree children hear a particular form is a function of synthesis combinatorics.[7]

Language	Format	Kids	Sessions	Words
Chintang	Toolbox	4	419	828272
Cree	CHAT	1	10	21525
Indonesian	Toolbox	8	997	2496828
Inuktitut	CHAT-like	5	77	73302
Japanese	XML	7	341	1235364
Russian	Toolbox	5	448	2022992
Sesotho	XML	4	129	237247
Turkish	CHAT-like	8	373	1139877
Yucatec	CHAT-like	3	234	120441

Table 2: Corpora

An overview of the corpora is given in Table 2. We brought together these corpora, and importantly some of the expert linguists who compiled them, to create a single syntactically and semantically interoperable database, which we discuss below. Each corpus is a detailed multi-year longitudinal study that consists of spoken language utterances by numerous participants in culturally distinct settings. Each corpus contains target children and child-directed speech mainly from mother-child interactions. Some corpora, like Chintang, contain a variety of participants, including parents, family members, playmates, etc.

As is standard practice in child language acquisition corpus development, recordings are captured at regular intervals (e.g. every two weeks) for a period of a year or more, centered around some number of target children. For example, the Chintang corpus is nearly 1 million words in its entirety, contains hundreds of participants, was compiled between 2004 and 2015, and is morphologically annotated, part-of-speech tagged, and translated into English and Nepali. An example of a conversational exchange encoded in the CLRP Toolbox format is given in Figure 1.



```
\ELANParticipant XX1
\tx uturiyaŋ loĩse? basai?ko
\gw uturi        yaŋ loĩse?          basai?ko
\mph u-      turi  yaŋ loĩs      -e      ba       -sa   -i?  -go
\mgl 3sPOSS- urine ADD  bring.out -IND.PST DEM.PROX -OBL  -LOC  -NMLZ
\lg C-      C/N   C    C         -C       C         -C    -C    -C
\id 6709-   2355  2450 1566      -1234    643       -6740 -6729 -1119
\ps gm-     n     gm   vt        -gm      pro       -gm   -gm   -gm
\eng She did the pee
\nep त्यसको सु गर्यो ।
\dt 31/Jan/2013

\ref 002
\ELANBegin 00:00:18.322
\ELANEnd 00:00:19.279
\ELANParticipant XX2
\tx la chemmuse? ni
\gw la     chemmuse?        ni
\mph lo     chep- mus -e       ni
\mgl okay   pee-  pee -IND.PST EMPH
\lg C/N    C-    C   -C        N
\id 1559   77-   78  -1234    1770
\ps interj v-    vi  -gm      gm
\eng Oh ! this has done the pee
\nep ल पिसाब फेर्यो ।
\dt 31/Jan/2013
```

Figure 1: Toolbox format

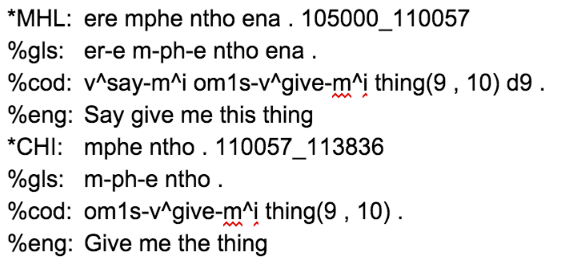
The format differs greatly, from say, conversational exchanges in the Sesotho corpus (Demuth, 1992)[8] and the

is investing significant financial and human resources into completing the fifth cluster by creating, from scratch, a substantial acquisition corpus for Dene (ISO 639-3: chp), a Athabaskan language spoken in Canada by fewer than 12k people. Dene, along with Cree (cre), form a fifth cluster which contains polysynthetic languages, i.e. languages in which "sentences" are composed typically of many morphemes. So far, the Dene corpus contains audio-visual recordings of eight children (ages 2-4) and their families. It currently consists of 200+ sessions (190 hrs+) with transcriptions and translations in ELAN, with glossing in Toolbox.

[7] Another example is verbal inflection: English typically has three forms, e.g. kick, kicks, kicked. But another language in our sample, Chintang, has over 1800 inflectional forms **per verb**.

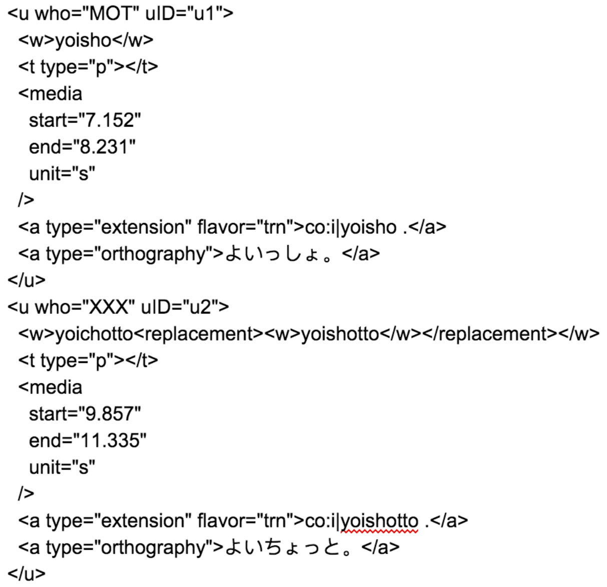
[8] This example is encoded in the CHILDES CHAT standard,

Japanese MiiPro corpus,[9] as shown in Figures 2 and 3.



```
*MHL:  ere mphe ntho ena . 105000_110057
%gls:  er-e m-ph-e ntho ena .
%cod:  v^say-m^i om1s-v^give-m^i thing(9 , 10) d9 .
%eng:  Say give me this thing
*CHI:  mphe ntho . 110057_113836
%gls:  m-ph-e ntho .
%cod:  om1s-v^give-m^i thing(9 , 10) .
%eng:  Give me the thing
```

Figure 2: CHAT format



```
<u who="MOT" uID="u1">
  <w>yoisho</w>
  <t type="p"></t>
  <media
    start="7.152"
    end="8.231"
    unit="s"
  />
  <a type="extension" flavor="trn">co:i|yoisho .</a>
  <a type="orthography">よいっしょ。</a>
</u>
<u who="XXX" uID="u2">
  <w>yoichotto<replacement><w>yoishotto</w></replacement></w>
  <t type="p"></t>
  <media
    start="9.857"
    end="11.335"
    unit="s"
  />
  <a type="extension" flavor="trn">co:i|yoishotto .</a>
  <a type="orthography">よいちょっと。</a>
</u>
```

Figure 3: XML

Although each corpus is in a different encoding format, each corpus is transcribed at the utterance, word and morpheme levels and additional annotation tiers include utterance timestamps, morphological analysis, part-of-speech labels, etc.[10] To bring the data together, we developed our own ETL pipeline.

3. Extract, Transform, Load

The ETL pipeline (Inmon, 1992; Kimball, 1996) developed for this project is written in Python and transforms original corpora data formats[11] into a single standards-abiding digital format encoded in relational tables using SQLAlchemy for its database ORM, SQLite for data storage, and plain CSV files and R data frames for output formats. The data formats we encountered are:

- Corpus-compiler specific specifications encoded in SIL's Toolbox,[12] e.g. Russian (Stoll and Roland, 2008), Chintang and Indonesian[13]; see Example 1 above
- Various versions of the CHILDES standard called CHAT[14]
- Talkbank XML

The data are transformed by:

- Converting the files into Unicode (UTF-8 NO-BOM NFD) plain text
- Correcting legacy character codes that were lost in translation and identified with unigram character code and grapheme models
- Extracting annotator comments from utterance tiers and removing punctuation
- Associating all annotations explicitly with the three main levels utterance–word–morpheme, possibly reassembling conflicting structures in the original data aligning word and morpheme level annotations within an utterance explicitly
- Unifying metadata standards and data types (e.g. unifying speaker role labels, age formats)
- Unifying linguistic terminology from the different annotations; creating terminological interoperability by mapping linguist expert opinion of grammatical categories to a unified set[15]
- Inferring additional information from annotations, e.g. determining the sentence type from punctuation in the translation

The data are then loaded into a simple relational database, with the tables: Sessions, Speakers, Utterances, Words and Morphemes. Each session has multiple speakers and multiple utterances. Each utterance is in a one-to-many relationship with words and each word is in a one-to-many relationship with morphemes. Sessions contain additional metadata (e.g. when, where, whom). Utterances contain information like timestamps and addressee. Words and morphemes contain linguistic analysis, e.g. part-of-speech tags, morphological glosses.

see: http://childes.psy.cmu.edu/manuals/CHAT.pdf.

[9] http://childes.psy.cmu.edu/

[10] The compiled data provide not only a platform for addressing questions of language acquisition, but also a resource for developing data-rich NLP tasks for and with under-resourced languages data.

[11] Collected between 1984-2005, hence the void of technological incompatibility of data formats.

[12] http://www-01.sil.org/computing/toolbox/

[13] http://childes.psy.cmu.edu/manuals/10eastasian.pdf

[14] Although the CHAT standard is fairly well defined, it allows for a great deal of encoding flexibility on part of the corpus compiler, which proved challenging in our work.

[15] Based on the Leipzig Glossing Rules as far as possible, however we acknowledge that lossless translation between categories is an issue of intense debate (cf. Haspelmath (2010a), Haspelmath (2010b), Newmeyer (2010)).

From the relational tables, we output the data into different formats, e.g. CSV files and R data frames, to provide broad data accessibility and to allow our team to leverage its expertise in data science, visualizations and statistics.[16]

4. Exploratory Analysis

In a nutshell, our ELT process takes disparate legacy data formats and creates a unified relational database ripe for qualitative and quantitative analysis. An initial step towards our research goals involves developing techniques for visual exploration of the data.

One technique that we have developed so far uses an interactive sunburst visualization (Stasko and Zhang, 2000) that we populate with language- and speaker-specific corpus data from the database.[17] For example, Figures 4 and 5 show the distribution of verbal categories produced by a Russian child between the ages of 1;08.10 and 1;09.09.[18]

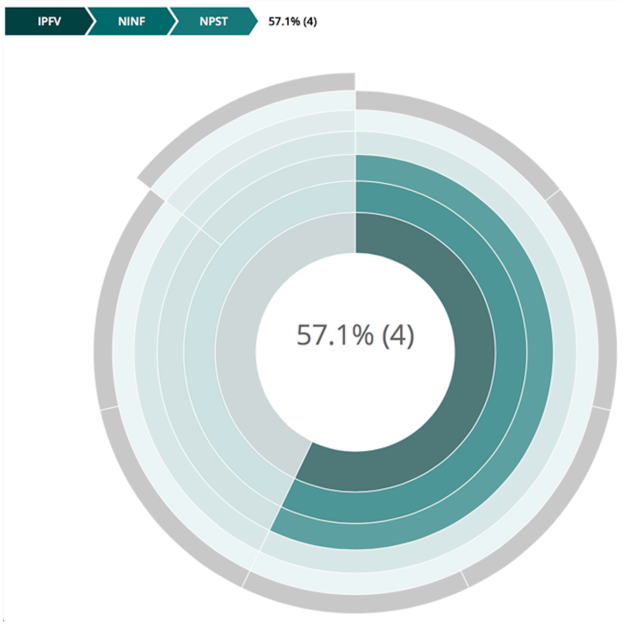


Figure 4: Distribution of imperfective verbs

These figures illustrate a Russian child's preference for using the imperfective aspect with non-past verb forms and the perfective aspect with past tense and imperative verb forms. During the first month of the recordings, the child uses the imperfective aspect exclusively with non-past verb forms. As expected, the child uses the perfective aspect with past tense verb forms, but the child uses mainly the perfective aspect with imperative forms (the segment in the third layer) in this age range.

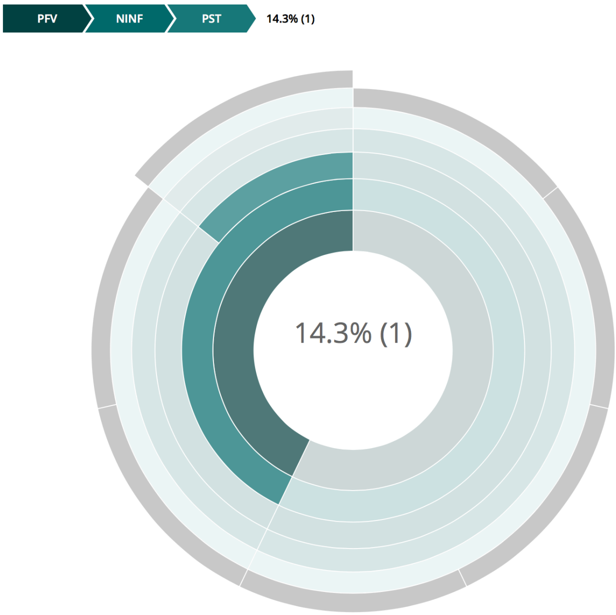


Figure 5: Distribution of perfective verbs

Similar patterns were detected for the other Russian children and this observation supports the findings of Gagarina (2000) and Filiouchkina (2005), both of whom investigate the distribution of the usage of imperfective and perfective Russian verbs with present and past forms. Gagarina (2000) shows that children start using both aspectual forms relatively early, but the occurrences of which are not evenly distributed: verbs that are uttered in imperfective forms are mostly used in their non-past form, and perfective verbs mostly denote telicity and are used in their past form. A strong correlation between the usage of the perfective aspect with telic verb forms is also found in other languages, such as English, French, Italian and Greek (Stoll, 2001), which suggests a cross-linguistic preference by children. A first question for exploratory analysis is whether this pattern is due to the frequency distribution of verb forms in caregiver's speech.

In order to visualize and investigate a potential influence in the usage of perfective and imperfective verb forms by the adult caregivers, we add their child-directed speech utterances during the same age ranges of the children.[19] Figure 6 shows the comparison between the verbal morphology visualization for one Russian child and his caretaker.

The *distributional bias* hypothesis by Shirai and Andersen (1995) states that children use imperfective and perfective verb forms mostly as they hear them from their caregivers, that is, imperfective (or progressive) verb forms are mostly with verbs falling into Vendler's (1957) verb class of activ-

[16] There are obviously privacy issues involved in the collection and dissemination of data on child language acquisition. The Chintang, Inuktitut, Turkish, Yucatec and Russian corpora are unpublished but included in the sample and are available under the ACQDIV Terms of Use; see: `http://www.acqdiv.uzh.ch/`. The Cree, Sesotho, Japanese MiiPro and Miyata, and Indonesian corpora are available under the creative commons license CC BY-NC-SA 3.0.

[17] We build on source code by Rodden, `http://bl.ocks.org/kerryrodden/7090426`, by adding a zoom function, drop-down menus for selecting different speakers, and a slider button for navigating through time (not shown).

[18] From 1 year, 8 months and 10 days to 1 year, 9 months, 9 days.

[19] 1;08.10–1;09.09 and 2;00.26–2;01.12.

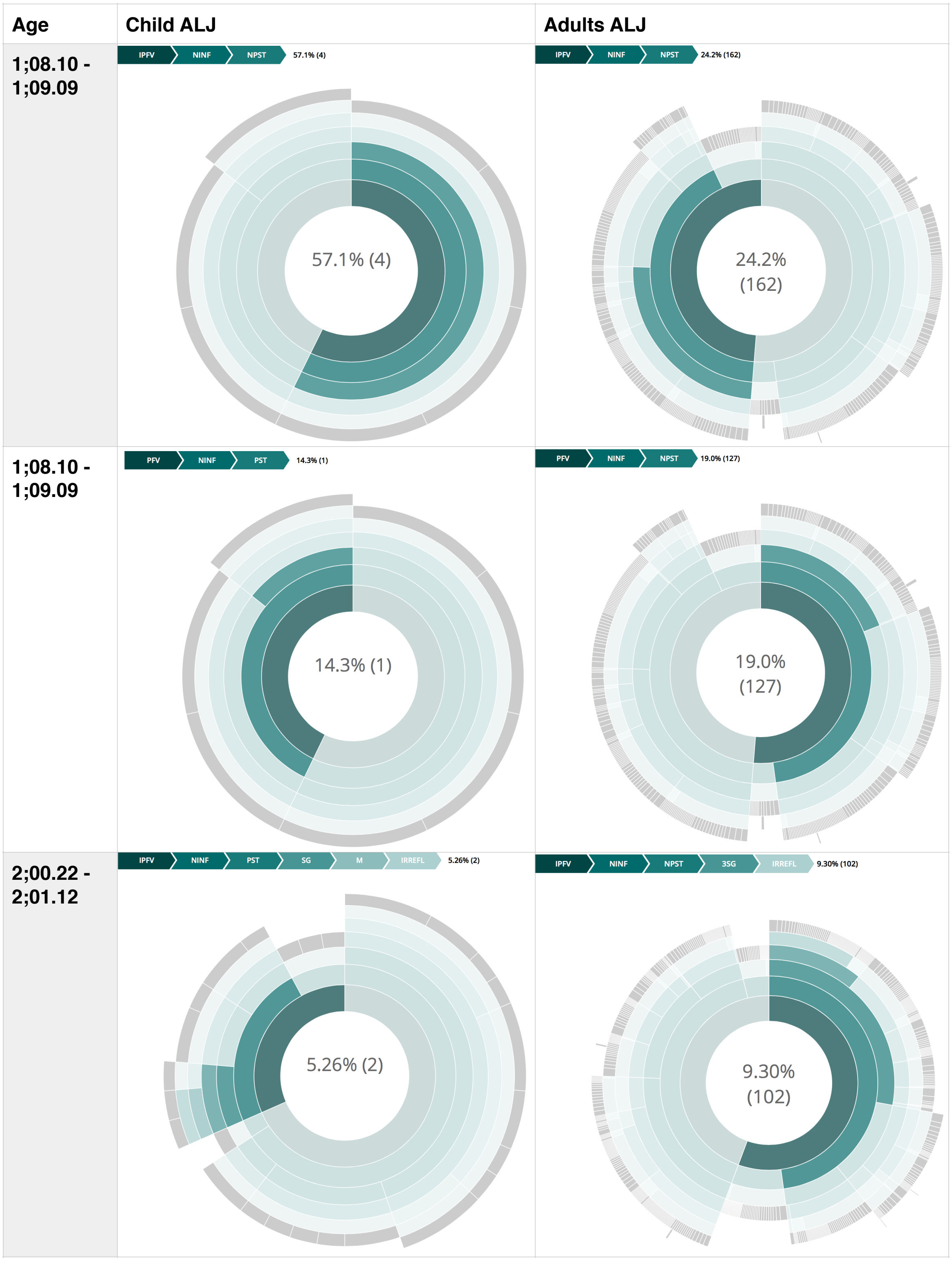


Figure 6: Comparison of verb forms used by ALJ and ALJ's caregiver

ities (in the present tense), whereas perfective verb forms are mostly used with verbs denoting achievements or accomplishments (in the past tense).[20]

In Figure 6, the first row seems to support the distributional hypothesis because the adults use imperfective verb forms mostly with the present tense (the child uses perfective verbs exclusively in their present tense). However, the pattern looks different for the perfective aspect. In the second row, the sunburst visualization for the child shows that he uses the perfective aspect with either past tense or imperative (not highlighted) verb forms, even though he also hears the perfective aspect with non-past tense forms being

[20] Examples include, activity: walk; achievement: understand; accomplishment: build.

uttered by the caregivers. Although the imperfective aspect is more frequently used with present tense verb forms by the adults, the child still hears present tense verbs mostly being associated with the imperfective aspect, hence the preference for using non-past verb forms with the imperfective aspect.

In the third row, the child (ages 2;00.26–2;01.12) starts to use equally frequently the imperfective aspect with past tense verb forms (most frequently in the singular masculine form) and the imperative form. The adults, however, retain the usage of the imperfective aspect with mostly present tense verb forms. Figure 7 illustrates the zoom function within our visualization, which allows us to also concentrate on deeper levels by zooming in to see only singular perfective past forms in their masculine form (hovering over the black tiles will show the lemma of the verb form).

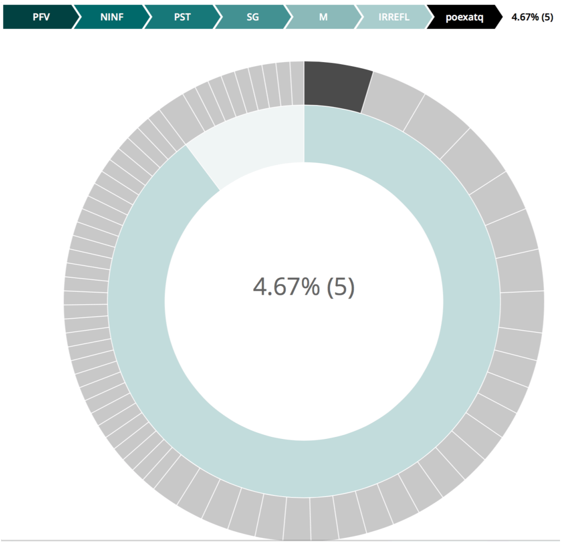


Figure 7: Zoomed in sunburst visualization

5. Preliminary Research Results

Our first research aim is to characterize patterns in child-directed speech and to relate them to first language acquisition theory from a cross-linguistic perspective. In other words, we want to identify how children can acquire any language effortlessly, despite the remarkable diversity of linguistic structures in the world's 7000 or so languages. Therefore, one area of research is to use quantitative methods to investigate patterns in child-directed speech.

In previous work, we introduce quantitative methods to characterize language acquisition development (Stoll et al., 2011; Stoll and Bickel, 2013b). Our current research analyzes the challenges children encounter when learning languages with different degrees of complexity. For example, when comparing the acquisition of the verbal system in English and Chintang (a polysynthetic Sino-Tibetan language) we find that children learning both languages become proficient at approximately the same time, despite the differences in morphological complexity that they encounter (Stoll et al., Forthcoming).

In other recent work, we build on existing studies that model language acquisition data using network theory by comparing the child-directed speech from the nine typologically diverse languages in the ACQDIV database. We show that networks created from child-directed speech all exhibit small-world structural characteristics even though the networks from these morphologically very different languages reflect the frequency effects of what linguists consider a word. Our finding is in line with network construction in child language acquisition models that have defined links in terms of semantic or grammatical relationships, both of which exhibit convergent features in their global structures (Ke, 2007). It has been suggested that small-world (and free-scale) structural characteristics reflect self-organization in the lexicon – a feature that may account for universal properties like fast retrieval from the mental lexicon, but which may also help to account for the fact that children can learn any language's patterns, despite the remarkable diversity in morphological structures.

6. Summary

In sum, we have created a syntactically and semantically interoperable database compiled from technologically disparate and typologically maximally diverse child language acquisition corpora. This single unified database allows us to investigate qualitatively and quantitatively questions in cross-linguistic child language acquisition. Although we have just recently started the research phase of our project, we show here briefly some visualization techniques that have been developed so far and we discuss some of our initial research findings regarding identifying patterns in cross-linguistic child-directed speech.[21]

7. Acknowledgements

The research leading to these results has received funding from the European Research Council (FP7, Grant Agreement no. 615988). We also thank three anonymous reviewers for their feedback.

8. Bibliographical References

Bickel, B. (2014). Linguistic diversity and universals. In Nicholas J. Enfield, et al., editors, *The Cambridge Handbook of Linguistic Anthropology*. Cambridge University Press, Cambridge.

Demuth, K. (1992). Acquisition of Sesotho. In D. Slobin, editor, *The Cross-Linguistic Study of Language Acquisition*, volume 3, pages 557–638. Lawrence Erlbaum Associates.

Evans, N. and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(05):429–448.

Filiouchkina, M. (2005). How tense and aspect are acquired: a cross-linguistic analysis of child Russian and English. *Nordlyd*, 32(1).

Gagarina, N. (2000). The acquisition of aspectuality by Russian children: the early stages. *ZAS Papers in Linguistics*, 15:232–246.

[21] SM & RS devised the conceptual architecture of the database. SM was responsible for the technological realization. SM, RS, DP & CH wrote the code. SM & DP wrote the paper. SS acquired the funding and developed the sampling algorithm.

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.

Inmon, W. H. (1992). *Building the data warehouse*. John Wiley & Sons, Inc.

Ke, J. (2007). Complex networks and human language. *arXiv preprint cs/0701135*.

Kimball, R. (1996). *The data warehouse toolkit*. John Wiley & Sons, Inc.

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

Steven Moran, et al., editors. (2014). *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

Shirai, Y. and Andersen, R. W. (1995). The acquisition of tense-aspect morphology: a prototype account. *Language*, pages 743–762.

Stasko, J. and Zhang, E. (2000). Focus + context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization*, pages 57–65. IEEE.

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

Stoll, S. and Bickel, B. (2013b). The acquisition of ergative case in Chintang. In Edith L. Bavin et al., editors, *The acquisition of ergativity*, pages 183–208. Benjamins, Amsterdam.

Stoll, S. and Roland, M. (2008). Audio-visual longitudinal corpus on the acquisition of russian by 5 children.

Stoll, S., Bickel, B., Lieven, E., Banjade, G., Bhatta, T. N., Gaenszle, M., Paudyal, N. P., Pettigrew, J., Rai, I. P., Rai, M., and Rai, N. K. (2011). Nouns and verbs in Chintang: children's usage and surrounding adult speech. *Journal of Child Language*, 39:284–321.

Stoll, S., Mazara, J., and Bickel, B. (Forthcoming). The acquisition of polysynthetic verb forms in Chintang. To appear in Fortescue, M., Mithun, M., and Evans, N. (Eds.) Handbook of Polysynthesis. Oxford: Oxford University Press.

Stoll, S. E. (2001). *The acquisition of Russian aspect*. Ph.D. thesis, University of California, Berkeley.

Vendler, Z. (1957). Verbs and times. *The philosophical review*, 66(2):143–160.