VARIATION IN LANGUAGE DEVELOPMENT

Elena Lieven & Sabine Stoll
Max Planck Institute for Evolutionary Anthropology
Outline of the course

- Session 1: Introduction, questions, background, data (Sabine Stoll)
- Session 2: Crosslinguistic studies of infant speech perception and production (Elena Lieven)
- Session 3: Composition of early vocabulary: nouns and verbs (Sabine Stoll)
- Session 4: The communicative environment, input and uptake (Elena Lieven)
- Session 5: Crosslinguistic corpus studies (Elena Lieven)
- Session 6: Crosslinguistic experimental studies (Elena Lieven)
- Session 7: Acquisition of tense and aspect (Sabine Stoll)
- Session 8: Acquisition of ergativity (Sabine Stoll)
Outline for Session 1

- Major questions of the field
- Prerequisites for language learning (first year of life)
- Diversity in acquisition tasks:
  - Languages:
    - Grammatical features and tasks of the children
    - Sampling issues
  - Methods of data collection:
    - Diaries
    - Questionnaires
    - Experiments
    - Longitudinal studies
  - Data collection of Chintang, an endangered language of Eastern Nepal
Major questions

- Are there some universal acquisition principles? What is innate?
- What are the factors responsible for the order of acquisition (general cognitive development, language specific factors, cultural environment of child rearing etc.)?
- Are the strategies children use in learning a language more similar within a specific language than across languages?
- If there are different strategies, do they depend on the structure of the language?
- What role does the input and the cultural context play for the acquisition process?
Some observations

- Huge diversity in the languages of the world.
- Different languages pose different challenges to acquisition (Slobin).
- “One cannot study universals without studying particulars.” (Slobin, 1985: 4)
- Crosslinguistic research as a method to reveal both universals and language-specific patterns (Slobin, 1982).
Why is variation an important topic?

- What is universal in language acquisition is one of the major questions, to find out about this, we need to look at variation.

- Predictions of nativist theories: there is no qualitative difference in language acquisition across children.

- Prediction of usage-based theories: qualitative differences are possible and the input has a strong impact on how children's development looks like.
# Variation in Language Acquisition

<table>
<thead>
<tr>
<th>Within individual languages</th>
<th>Across languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• temporal variation (time of acquisition)</td>
<td>• grammatical variation</td>
</tr>
<tr>
<td>• qualitative variation</td>
<td>• cultural variation</td>
</tr>
<tr>
<td>• context-specific variation</td>
<td></td>
</tr>
</tbody>
</table>
Language Comprehension

- Prerequisites for language comprehension:
  - acoustic perception of speech
  - pattern recognition
  - pattern analysis
  - pattern memorization
Prelinguistic development: Prerequisites for language learning

Birth to 6 months

- Recognition of mother’s voice (de Casper & Fifer 1980)
- Distinguish native language from other languages (Mehler et al. 1988, Moon et al. 1993)
- Categorial perception of speech sounds (Eimas et al. 1971)
- Recognition of identity of sounds across contexts (Kuhl 1980)
- Segment speech (Saffran et al. 1996)
- Preference for infant-directed speech (Fernald et al. 1989; Cooper & Aslin 1990)
Prelinguistic development

6-12 months

- Discrimination of phonemic contrasts. Up to approx. 10 months discriminations of all contrasts. Then, only contrast of native language are distinguished. (Werker & Tees 1984)

- By 7 1/2 months children listen longer to familiarized words within longer sentences. (Jusczyk & Aslin 1995)
Prelinguistic development

Abilities around 9 months (e.g. Tomasello 2003)

- Joint attention
- Recognition of symbols
- Imitation (role reversal)
- Intention reading
- Pointing
Development of linguistic abilities: individual variation

- First words (9 months to 1 year of age)
- Strong variation in when children start speaking and how they progress (within a given language)

### Table 4.3. Age of acquisition for 10- and 50-word vocabularies in six children

<table>
<thead>
<tr>
<th>Child</th>
<th>Sex</th>
<th>10 words</th>
<th>50 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>M</td>
<td>1;0</td>
<td>1;5</td>
</tr>
<tr>
<td>S2</td>
<td>F</td>
<td>1;1</td>
<td>1;6</td>
</tr>
<tr>
<td>S3</td>
<td>M</td>
<td>1;2</td>
<td>1;7</td>
</tr>
<tr>
<td>S4</td>
<td>F</td>
<td>1;2</td>
<td>1;8</td>
</tr>
<tr>
<td>S5</td>
<td>M</td>
<td>1;4</td>
<td>1;10</td>
</tr>
<tr>
<td>S6</td>
<td>F</td>
<td>1;3</td>
<td>1;7</td>
</tr>
</tbody>
</table>

Source: Robb, Bauer, & Tyler 1994:40. Used with permission from Alpha Academic.
Development of linguistic abilities: individual variation

Fig. 2.—Number of words on the Infant form reported to be comprehended by children at each month—median values and spread of score distributions. a, Observed values. b, Fitted values. A portion of this figure is adapted from Fenson et al. (1993, p. 103), with permission of the Singular Publishing Group, Inc.

Fig. 3.—Number of words on the Infant form reported to be produced by children at each month—median values and spread of score distributions. a, Observed values. b, Fitted values. A portion of this figure is adapted from Fenson et al. (1993, p. 104), with permission of the Singular Publishing Group, Inc.
Table 1. Individual Differences in Language Development: Summary of Claims in the Literature

<table>
<thead>
<tr>
<th>Strand 1</th>
<th>Strand 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Semantics</strong></td>
<td><strong>Semantics</strong></td>
</tr>
<tr>
<td>High proportion of nouns in first 50 words</td>
<td>Low proportion of nouns in first 50 words</td>
</tr>
<tr>
<td>Single words in early speech</td>
<td>Formulate in early speech</td>
</tr>
<tr>
<td>Imitates object names</td>
<td>Unselective imitation</td>
</tr>
<tr>
<td>Greater variety within lexical categories</td>
<td>Less variety within lexical categories</td>
</tr>
<tr>
<td>Meaningful elements only</td>
<td>Use of &quot;dummy&quot; words</td>
</tr>
<tr>
<td>High adjective use</td>
<td>Low adjective use</td>
</tr>
<tr>
<td>Context-flexible use of names</td>
<td>Context-bound use of names</td>
</tr>
<tr>
<td>Rapid vocabulary growth</td>
<td>Slower vocabulary growth</td>
</tr>
<tr>
<td><strong>Grammar</strong></td>
<td><strong>Grammar</strong></td>
</tr>
<tr>
<td>Telegraphic in Stage 1</td>
<td>Inflexions and function words in Stage 1</td>
</tr>
<tr>
<td>Refers to self and others</td>
<td>Refers to self and others</td>
</tr>
<tr>
<td>by name in Stage 1</td>
<td>by pronoun in Stage 1</td>
</tr>
<tr>
<td>Noun-phrase expansion</td>
<td>Verb phrase expansion</td>
</tr>
<tr>
<td>Morphological overgeneralization</td>
<td>Morphological undergeneralization</td>
</tr>
<tr>
<td>Consistent application of rules</td>
<td>Inconsistent application of rules</td>
</tr>
<tr>
<td>Novel combinations</td>
<td>Frozen forms</td>
</tr>
<tr>
<td>Imitation is behind spontaneous speech</td>
<td>Imitation is ahead of spontaneous speech</td>
</tr>
<tr>
<td>Fast learner</td>
<td>Slow learner</td>
</tr>
<tr>
<td><strong>Pragmatics</strong></td>
<td><strong>Pragmatics</strong></td>
</tr>
<tr>
<td>Object-oriented</td>
<td>Person-oriented</td>
</tr>
<tr>
<td>Declarative</td>
<td>Imperative</td>
</tr>
<tr>
<td>Low variety in speech acts</td>
<td>High variety in speech acts</td>
</tr>
<tr>
<td><strong>Phonology</strong></td>
<td><strong>Phonology</strong></td>
</tr>
<tr>
<td>Word-oriented</td>
<td>Intonation-oriented</td>
</tr>
<tr>
<td>High intelligibility</td>
<td>Low intelligibility</td>
</tr>
<tr>
<td>Segmental emphasis</td>
<td>Suprasegmental emphasis</td>
</tr>
<tr>
<td>Consistent pronunciation across word tokens</td>
<td>Variable pronunciation across word tokens</td>
</tr>
<tr>
<td><strong>Demographic Variables</strong></td>
<td><strong>Demographic Variables</strong></td>
</tr>
<tr>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Firstborn</td>
<td>Later-born</td>
</tr>
<tr>
<td>Higher SES</td>
<td>Lower SES</td>
</tr>
</tbody>
</table>
Potential learning tasks: Languages today (6000-7000)
Language families a subsample of 2560 languages (WALS)

Dryer, 2005
Some examples of variation

Phoneme Inventories (Maddieson):

- Consonant inventories
  - 6 (Rotokas, Papua New Guinea) - 122 (!Xóõ, Southern Khoisan) out of a sample of 562 languages

- Vowel inventories
  - 2 (Yimas (Papua New Guinea) -14 (German)
Some examples of variation

- Inflectional synthesis of the verb (Bickel & Nichols 2005, WALS)
- Degree of synthesis as defined by the number of elements that make up a synthetic verb form
- Large variation form 0 categories per verb form (Vietnamese) to 13 (Koasati)
Some examples of variation

<table>
<thead>
<tr>
<th>1s</th>
<th>1p</th>
<th>2s</th>
<th>3s</th>
<th>1s</th>
<th>1p</th>
<th>2s</th>
<th>3s</th>
<th>PT</th>
</tr>
</thead>
<tbody>
<tr>
<td>treffen</td>
<td>treffen</td>
<td>treffen</td>
<td>treffen</td>
<td>treffen</td>
<td>treffen</td>
<td>treffen</td>
<td>treffen</td>
<td>treffen</td>
</tr>
<tr>
<td>‘to meet’</td>
<td>‘to meet’</td>
<td>‘to meet’</td>
<td>‘to meet’</td>
<td>‘to meet’</td>
<td>‘to meet’</td>
<td>‘to meet’</td>
<td>‘to meet’</td>
<td>‘to meet’</td>
</tr>
</tbody>
</table>

| Lieven & Stoll, DGS Summer School, 2010 | 19 |
Some examples of variation

<table>
<thead>
<tr>
<th>German</th>
<th>Chintang</th>
</tr>
</thead>
<tbody>
<tr>
<td>ge-troff-en</td>
<td>u-ma-tup-yokt-e-hê</td>
</tr>
<tr>
<td>an-ge-troff-en</td>
<td>3-NEG-meet-NEG-PAST-1</td>
</tr>
<tr>
<td>aber:</td>
<td>‘He did not meet me’</td>
</tr>
<tr>
<td>*ge-an-troff-en</td>
<td>or:</td>
</tr>
<tr>
<td>*ge-troff-an-en</td>
<td>ma-u-top-yokt-e-hê</td>
</tr>
<tr>
<td></td>
<td>ma-top-u-yokt-e-hê</td>
</tr>
</tbody>
</table>

Assumption so far: Free ordering of affixes is impossible
Some examples of variation

- Word order (Subject, Verb, Object)
- All 6 logically possible orders are attested
  - SOV (Japanese)
  - SVO (Mandarin)
  - VSO (Irish)
  - VOS (Nias, Austronesian)
  - OVS (Hixkaryana, Carib, Brazil)
  - OSV (Nadëb, Brazil)
- No predominant order
Some examples of variation

- Structural diversity: semantic mapping
- Inclusive vs. exclusive pronouns

German:  
ich  
wir

Chintang:  
akka  
ani  
ananga

inclusive  exclusive
Some examples of variation

- Structural diversity: world knowledge
- Kinship and suffixes (Martuthunira, Australia)

Suffix used to mark kinship. Same generation (brother, sister, grandfather) vs. alternating generation (father, mother, great-grandfather). Suffix is used to mark same generation set.

Nyinta wiya-nmayi-nha marrkara-ngarli- ku -rla?
2sgNOM see -COLL-PAST brother PLURAL-ACC-PST
Did you see your younger brother’s?

Dench, 1987
Languages today

- Approx. 6000-7000 living languages in the world
- Approx. 300 language families
- Approx. 500-700 languages are described or we know at least something about them
- Approx. 300 variables that typologists have extensive knowledge about
Our knowledge about languages

- Estimation of the minimum number of languages under the following assumptions (Bickel, 2008):
  - on average at every point in time approx. 5000 languages (more realistically: 4000-12000)
  - on average maximally 1000 years per language (criterium of interintelligibility)
  - probable age of modern languages minimum 100’000 years
    \[
    \frac{100,000}{1000} \cdot 5000 = 500,000
    \]
  - todays population < 1% of total population (amount of languages)
  - of those we know only about 1% (500).
  - But soon, we will not be able to learn anything.
Language death

Data from Ethnologue 13, map Bibiko 2006

Lieven & Stoll, DGfS Summer School, 2010
Language development in the future (estimated death rate of 50% every 50 years):
Survey of languages with acquisition studies
Euroversals (Haspelmath 1998, 2001)

- Definite vs. indefinite article
- Relative pronouns
- ‘HAVE’- perfect
- Participle based passive
- External possessor
- ‘nobody-came’- constructions
- ‘as-big-as’-constructions
- agreement with obligatory NP
Euroversals: Relative pronouns

Data from Comrie & Kuteva 2005, N=166, EU vs. Rest, p<.001; EU vs. Asia, p=.05

Lieven & Stoll, DGfS Summer School, 2010
Data in language acquisition research

- Diaries (very early in development)
- Questionnaires (MCDI)
- Experiments
- Corpora
  - Longitudinal corpora
  - Cross-sectional corpora
- Modelling
Diary studies

- German: Preyer (1882), Stern & Stern (1928), Lindner (1885), Leopold (1948)
- French: Gregoire (1937)
- Russian: Gvozdev (1949)
- Polish: Zarebina (1965)
- many modern studies: e.g. Tomasello (1992)
Diary studies

Advantages:

- Every utterance of the child can be noted down in the early phases of acquisition

Disadvantages:

- Only feasible in the very early stages of acquisition
- Difficult to exactly note down the linguistic and extralinguistic context
Questionnaires

Advantages:

- large number of mothers can be tested
- standardized ‘test’
- results can be compared easily

Disadvantages:

- cultural issues might apply
- relies on the memory and interpretation of the mother
Advantages:
- controlled setting and variables which are the same for all children
- cause and effect can be tested

Disadvantages:
- only one context is tested
- not possible to administer in a lot of cultures
Longitudinal studies

- Audio- or video recording of the child and her caretakers in natural interaction
- Transcription of their speech
- Translation of the data
- Tagging of the corpus (translation, morphological glossing, parts-of-speech glossing)
- Further annotations
- Linking to video- or audio data
- Analysis of the data
Longitudinal studies

- Advantages:
  - spontaneous behaviour of the children is recorded
  - development of individual children can be characterized
  - input of the surrounding adults available for analysis

- Disadvantages:
  - sampling issues
  - potential contextual influences (recording situation)
  - small number of children (what is the norm?)
  - only correlations can be attested
Child Language Data Exchange System (CHILDES)


- Includes corpora and tools for studying conversational interactions.
  - 130 corpora (30 languages)

- Tools include:
  - programs for computer analysis of transcripts
  - methods for linguistic coding
  - systems for linking transcripts to digitized audio and video
CHILDES is the child language component of the TalkBank system. TalkBank is a system for sharing and studying conversational interactions.

<table>
<thead>
<tr>
<th>System</th>
<th>Programs and Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guidelines</td>
<td>Downloadable Database</td>
</tr>
<tr>
<td>Contributions</td>
<td>Browseable Database</td>
</tr>
<tr>
<td>Overviews</td>
<td>The CLAN Program</td>
</tr>
<tr>
<td>Introductions</td>
<td>WebCLAN</td>
</tr>
<tr>
<td>Membership List</td>
<td>Training Videos</td>
</tr>
<tr>
<td>How to subscribe</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Links</th>
<th>Manuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>TalkBank Database</td>
<td>CHAT Transcription</td>
</tr>
<tr>
<td>IASCL information</td>
<td>CLAN Programs</td>
</tr>
<tr>
<td>Other Child Language sites</td>
<td>Database Manuals</td>
</tr>
<tr>
<td>Research based on CHILDES</td>
<td>BTS sign transcription system</td>
</tr>
<tr>
<td>Related Software</td>
<td>Teacher's manual</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phonology and Fonts</th>
<th>Teaching with CHILDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phon &amp; PhonBank</td>
<td>Topics in language acquisition</td>
</tr>
<tr>
<td>Unicode and IPA for Mac</td>
<td>Teaching Tips and Resources</td>
</tr>
<tr>
<td>Unicode and IPA for Window</td>
<td>Child Language Bibliographies</td>
</tr>
<tr>
<td>Special Procedures</td>
<td></td>
</tr>
<tr>
<td>Procedures and tools for CA analysis</td>
<td>Part of Speech Analysis by MGR</td>
</tr>
<tr>
<td>Working with digitized video</td>
<td>Parental frequency count: view, download, documentation</td>
</tr>
<tr>
<td>Working with digitized audio</td>
<td>MRC lexical dictionary</td>
</tr>
<tr>
<td>The Computerized Comprehension Task</td>
<td>Syntactic analysis by GRASP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mirrors</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antwerp and Chukyo</td>
<td>Brian MacWhinney: homepage</td>
</tr>
</tbody>
</table>
Using local transcripts and local media

1. You need to download the transcripts from "zipped transcripts" below and unzip them. If you want to work with XML versions of the data, rather than CHAT versions, please go to the CHAT files.
2. If the corpus is linked to audio or video media, you need to download those files and place the media into the transcript folders.
3. You need to download and install the CLAN program.
4. To open a transcript, you double-click on it. If there is associated media, you can play the media using escape-8 for continuous playback or command-click for playing single utterances.

### Zipped Transcripts

- English - USA
- English - UK
- Celtic
- East Asian
- Germanic
- Romance
- Slavic
- Other Languages
- Bilinguals
- Clinical
- Frog Story Narratives
- Other Narratives
- PhonBank CHAT Data
- PhonBank Phon Data

### Using local transcripts with web-served media

This is like the first method, but, if you are on a broadband Internet connection, you can skip step #2, since CLAN will use links in the folder called /media to play the media over the web.
### Childes: Guilfoyle corpus (Irish)

<table>
<thead>
<tr>
<th>@Font: CAfont:13:0</th>
<th>@Begin</th>
</tr>
</thead>
<tbody>
<tr>
<td>@Languages: ga</td>
<td>@Id: ga</td>
</tr>
<tr>
<td>@Participants: GEA Investigator, CHI Cai Target Child, EIT Investigator, MOT Mother</td>
<td>@Id: ga</td>
</tr>
<tr>
<td>@Date: 05-JUL-1992</td>
<td>@Id: ga</td>
</tr>
<tr>
<td>@Birth of CHI: 01-DEC-1990</td>
<td>@Id: ga</td>
</tr>
<tr>
<td>*MOT: cead e sin?</td>
<td>*MOT: och ceard atfaighe agat?</td>
</tr>
<tr>
<td>%eng: that is that?</td>
<td>%eng: oh what do you have?</td>
</tr>
<tr>
<td>*GEA: xx anois babog</td>
<td>*GEA: xx anois babog</td>
</tr>
<tr>
<td>%eng: xx now a doll</td>
<td>%eng: xx now a doll</td>
</tr>
<tr>
<td>*CHI: babog</td>
<td>%xmr: doll</td>
</tr>
<tr>
<td>%eng: doll</td>
<td>*xmr: Na doll</td>
</tr>
<tr>
<td>*MOT: nach bh-fuil si go h-ainn?</td>
<td>*MOT: nach bh-fuil si go h-ainn?</td>
</tr>
<tr>
<td>%eng: isn't she beautiful?</td>
<td>%eng: isn't she beautiful?</td>
</tr>
<tr>
<td>*GEA: ceard eile e ata thios ann?</td>
<td>*GEA: ceard eile e ata thios ann?</td>
</tr>
<tr>
<td>%eng: what else is down there?</td>
<td>%eng: what else is down there?</td>
</tr>
<tr>
<td>*CHI: xx</td>
<td>*CHI: xx</td>
</tr>
<tr>
<td>%eng: xx</td>
<td>%eng: xx</td>
</tr>
<tr>
<td>*MOT: ceard e sin?</td>
<td>*MOT: ceard e sin?</td>
</tr>
<tr>
<td>%eng: what is that?</td>
<td>%eng: what is that?</td>
</tr>
<tr>
<td>*CHI: lora</td>
<td>*CHI: lora</td>
</tr>
<tr>
<td>%eng: truck</td>
<td>%eng: truck</td>
</tr>
<tr>
<td>%xmr: N truck</td>
<td>%xmr: N truck</td>
</tr>
<tr>
<td>*MOT: ab e?</td>
<td>*MOT: ab e?</td>
</tr>
</tbody>
</table>

**U1303[FI][CHAT] 1**
Bloom corpus: English

@Begin
@Languages: en
@Participants: CHI Peter Target, Child, LOI Lois investigator, PAT Patsy investigator, MOT Mother
@D: en/bloom/70/CHI/2.5.03/normal/Target,Child
@D: en/bloom/70/LOI/normal/investigator
@D: en/bloom/70/PAT/normal/investigator
@D: en/bloom/70/MOT/normal/Mother
@Comment: Mother had invited Patsy and LOI to stay for lunch and lunch
was on Peter's mind from the time they arrived. He had just
had breakfast but asked for food throughout the session.
Mother had tried to put him off in order not to "spoil his
appetite" for lunch, but he persisted and became rather
cranky. He was unable to concentrate on anything for very
long. Patsy and LOI brought new toys but Peter had just
acquired several new toys from a family about to move out of
the building. He was most excited about a large Hess oil
truck which he proudly showed and enjoyed. He also got a
magnetic alphabet and board and a cash register with wooden
coins inside. Mother was present for most of this session.
Jenny was there for part of it, but was asleep before lunch.
In the middle of his lunch, Peter asked to sit in Jenny's
chair. Mother put him there and Peter played baby for several
minutes, babbling and pretending to drink from a bottle. By
noon, he was very tired and protested when Patsy and LOI went
home.

@Situation: Lunch
*PAT: Hi Pete.
%mor: co[n]m:com:Pete.
%xgra: 1[0]:ROOT 2[1]:[VOC 3]:[PUNCT
%act: <be> Peter opens apartment door to let Lois and Patsy in
*LOI: <good morning> [>].
%mor: ad[t]good admission.
%xgra: 1[0]:ROOT 2[1]:LCT 3[1]:PUNCT
*CHI: <<[i ]> I lxx door [<].
%mor: pro[l] unk ox h|door.
%xgra: 1[2]:SUBJ 2[0]:ROOT 3[2]:OBJ 4[2]:PUNCT
*PAT: thank +you
J [1303][E][CMAT]}
Demuth corpus: Sesotho (some metadata)
Demuth corpus: Sesotho

```
Begin
Languages: se
Participants: BOY unidentified_child Playmate, NOI Noi Adult, ELS Elisha Playmate, MEM Mme_Manyilli Grandmother, CHI Hobohang Target_Child, KAT Katherine_Demuth
```

```
ID: se|Sesotho|BOY|Playmate
ID: se|Sesotho|NOI|Adult
ID: se|Sesotho|ELS|Playmate
ID: se|Sesotho|MEM|Grandmother
ID: se|Sesotho|CHI|28.|Target_Child
ID: se|Sesotho|KAT|Investigator
ID: se|Sesotho|KHT|Playmate
ID: se|Sesotho|MHL|Mother
ID: se|Sesotho|TSN|Playmate
ID: se|Sesotho|MOL|10.|Cousin
```

*MOL: ya Hao e ka Hobohang?
%gis: ya Hao e ka Hobohang?
%cod: ps9 j cp9 wh n\^name?
%eng: Where is yours Hobohang?
%sit: Rolling stones.

*MOL: e ne le moo.
%gis: e ne le moo.
%cod: sm9-fhe-sm9-cp loc.
%eng: It was here
%sit: H, Mol e

*MOL: jwale a fihla a e nda.
%gis: jwale u-a-fhi-a u-a-e-nk-a.
%cod: av sm2s-t\^p-v\^take-m\^in sm2s-t\^p-om9-v\^take-m\^in.
%eng: Now you arrive you take it
*MOL: mae ke e tswara.
%gis: mae ke-e-tswara.
%cod: pnts sm1s-t\^p_om9-v\^grab-m\^in.
%eng: I grab
%sit: x2
```
Chang corpus: Chinese
Jiwon corpus: Korean

Lieven & Stoll, DGfS Summer School, 2010
Languages with accessible corpora (CHILDES)
Language families with accessible corpora

- Basque (1)
- Sino-Tibetan (2)
- Indo-European (16)
- Uralic (2)
- Afro-Asiatic (1)
- Japanese (1)
- Korean (1)
- Niger-Congo (1)
- Dravidian (1)
- Tai-Kadai (1)
- Altaic (1)
Sampling Problems

We have only a very limited amount of data available of a very restricted set of languages with a very limited number of the variables that are found in the languages of the world. The sample is skewed towards European languages with a strong bias towards Germanic and Romance.

==> Difficult to make generalizations; mostly our claims can be only language specific or at best probabilistic
Solution: do field work

- Chintang child language corpus (Stoll, Lieven) financed by Volkswagen foundation programs DOBES (PI: Bickel) and Dilthey (Stoll) and by the MPI for Evolutionary Anthropology.

- Fieldwork since 2004.

- Collaboration between linguists, anthropologists and language acquisition researchers
Chintang: a polysynthetic language spoken in eastern Nepal

**Linguistics:**
Grammar
Discourse
Lexicon

How is the language structured?

**Ethnography:**
Rituals and religion
Social structure
Everyday life

How is the language used?

**Psycholinguistics:**
Acquisition strategies
Multilingualism

How is the language learned?

Balthasar Bickel
Novel Kishore Rai
Vishnu Rai

Martin Gaenzsle
Judith Pettigrew

Sabine Stoll
Elena Lieven

Lieven & Stoll, DGfS Summer School, 2010
# Chintang-project members

<table>
<thead>
<tr>
<th>Linguistics</th>
<th>Anthropology</th>
<th>Language Acquisition</th>
</tr>
</thead>
</table>
# Chintang corpus

<table>
<thead>
<tr>
<th>Category</th>
<th>Words (rounded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child language</td>
<td>499.728</td>
</tr>
<tr>
<td>Conversations</td>
<td>89.900</td>
</tr>
<tr>
<td>Narratives</td>
<td>37.500</td>
</tr>
<tr>
<td>Poems and songs</td>
<td>1.500</td>
</tr>
<tr>
<td>Myths and ritual texts</td>
<td>37.000</td>
</tr>
<tr>
<td>Descriptions</td>
<td>12.300</td>
</tr>
<tr>
<td>Sum</td>
<td>677.928</td>
</tr>
</tbody>
</table>
Chintang

- Sino-Tibetan language spoken in eastern Nepal
- approx. 6000 speaker
- trilingual society
- endangered language
- children still learn the language
# Chintang corpus

<table>
<thead>
<tr>
<th>N (children)</th>
<th>Recording</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2-3 h per month</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3-4 h per month</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3-4 h per month</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

- **monthly**
- **bimonthly**
- every 3 month
Chintang Child language project workflow: Managing a large scale project
# Building up the Chintang corpus

<table>
<thead>
<tr>
<th>1. Recording</th>
<th>2. Metadata</th>
<th>3. Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chintang natural speech (different ages &amp; genres)</td>
<td>Information about location, speakers, etc.</td>
<td>Text production</td>
</tr>
<tr>
<td><strong>Location:</strong> Nepal</td>
<td><strong>Location:</strong> Nepal</td>
<td><strong>Location:</strong> Nepal</td>
</tr>
<tr>
<td><strong>Staff:</strong> student assistants</td>
<td><strong>Staff:</strong> student assistants</td>
<td><strong>Staff:</strong> student assistants (Chintang native speakers)</td>
</tr>
<tr>
<td><strong>Output:</strong> session.mpg</td>
<td><strong>Software:</strong> imdi-editor</td>
<td><strong>Software:</strong> Transcriber</td>
</tr>
<tr>
<td></td>
<td><strong>Output:</strong> session.wav</td>
<td><strong>Input:</strong> session.wav</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Output:</strong> session.imdi</td>
</tr>
<tr>
<td>*add session to session-monitor</td>
<td>*check completeness of imdifuile</td>
<td>*assign the session to a transcriber</td>
</tr>
<tr>
<td>*add DVD to media-monitor</td>
<td>*upload to server</td>
<td>*update the monitor (person, status)</td>
</tr>
</tbody>
</table>

*the transcriber also does the Nepali-translation*
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Translation into the local lingua franca</strong></td>
<td><strong>Translation into an international language</strong></td>
<td><strong>Linguistic analysis</strong></td>
</tr>
<tr>
<td><strong>Location:</strong> Nepal</td>
<td><strong>Location:</strong> Germany</td>
<td><strong>Location:</strong> Germany</td>
</tr>
<tr>
<td><strong>Staff:</strong> student assistants (Chintang native speakers)</td>
<td><strong>Staff:</strong> student assistants</td>
<td><strong>Staff:</strong> student assistants</td>
</tr>
<tr>
<td><strong>Software:</strong> Elan</td>
<td><strong>Software:</strong> any texteditor</td>
<td><strong>Software:</strong> Toolbox</td>
</tr>
<tr>
<td><strong>Input:</strong> session.trs, session.wav, session.mpg</td>
<td><strong>Input:</strong> session.txt</td>
<td><strong>Input:</strong> session.txt</td>
</tr>
<tr>
<td><strong>Output:</strong> session.eaf</td>
<td><strong>Output:</strong> session.txt</td>
<td><strong>Output:</strong> session.txt</td>
</tr>
</tbody>
</table>

**4. Translation Nepali**
- *receive .trs and .eaf*
- *check completeness*
- *check speakercodes (use impdifile and speaker-DB)*
- *export as toolboxfile and convert (UTF-8, linebreaks)*
- *upload to server*
- *update the monitor (person, status, pwords)*
- [*calculate workspeed]*

**5. Translation English**
- *assign the toolboxfile to a translator*
- *update monitor*
- *receive toolboxfile*
- *check completeness*
- *upload to server*
- *update monitor*
- [*calculate workspeed]*

**6. Glossing**
- *assign toolboxfile to a glosser*
- *update monitor*
- *receive toolboxfile*
- *check completeness*
- *update monitor*
- *add information using scripts (e.g. POS, agegroup of speakers)*
- *upload to server*
Training of assistants in Nepal
Project

Name: Chintang and Puma Documentation Project
Title: Documentation of Chintang and Puma, two Kiranti languages of Eastern Nepal
ID: CPDP

Contact Information

Name: Prof. Dr. Balthasar Bickel
Address: Institut für Linguistik, Beethovenstr. 15, 04107 Leipzig, Germany
E-mail: bickel@uni-leipzig.de
Organisation: University of Leipzig

Descriptions

Language: English
Text: The aim of the project is to provide a rich linguistic and ethnographic documentation of two highly endangered but almost totally undocumented languages in eastern Nepal, Chintang and Puma. These languages belong to the Kiranti family of Tibeto-Burman. Chintang is spoken by the Chintang Rai in Chintang Village Development Committee (VDC) of Dhankuta district. Puma is spoken by the Puma Rai in Diplung, Mauvabote, Devistan and...
Transcription...
... into the target language Chintang (in Transcriber)

Transcriber 1.5.1+ (06/25/2005)

(no speaker)

LDCK2

kalpe mo dhara ta

LDCh3

kamala kham tisakina dhara thano
Translation into Nepali and English (in ELAN)
audura
ai
dhara catthã?

आ बारी
एडी!
घारानाई हामे।

tap
Hey!
I hit the tap.

va
ba
dhara
dha

ही
घुरा

this
tap!
Glossing (in Toolbox)

Lieven & Stoll, DGfS Summer School, 2010

70
Toolbox file with POS-tags
Toolbox file with agecodes
Meta-data editor (IMDI)

Project
- Name: Chintang and Puma Documentation Project
- Title: Documentation of Chintang and Puma, two Kiranti languages of Eastern Nepal
- ID: CPDP

Contact Information
- Name: Prof. Dr. Balthasar Bickel
- Address: Institut für Linguistik, Beethovenstr. 15, 04107 Leipzig, Germany
- E-mail: bickel@uni-leipzig.de
- Organisation: University of Leipzig

Descriptions
- Language: English
- Text: The aim of the project is to provide a rich linguistic and ethnographic documentation of two highly endangered but almost totally undocumented languages in eastern Nepal, Chintang and Puma. These languages belong to the Kiranti family of Tibeto-Burman. Chintang is spoken by the Chintang Rai in Chintang Village Development Committee (VDC) of Dhankuta district. Puma is spoken by the Puma Rai in Diplung, Mauwabor, Devshan and...
## Monitor of workflow

<table>
<thead>
<tr>
<th>Session Name</th>
<th>Transcription Status</th>
<th>Translation Status</th>
<th>Interlinear Status</th>
<th>Length</th>
<th>Child Utterances</th>
<th>Adult Utterances</th>
<th>Other Child Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 30</td>
<td>done (GAR)</td>
<td>Nepali done (GAR)</td>
<td></td>
<td>1422</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 51</td>
<td>done (AnR)</td>
<td>all done (AnR, DR)</td>
<td>in progress (KK)</td>
<td>244</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 53</td>
<td>done</td>
<td>done</td>
<td></td>
<td>142</td>
<td>27, 70</td>
<td>17, 55</td>
<td>98, 142</td>
</tr>
<tr>
<td>Session 55</td>
<td>done (AnR)</td>
<td>all done (AnR, DR)</td>
<td>in progress (KK)</td>
<td>169</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 56</td>
<td>done (JK)</td>
<td>Nepali done (JK)</td>
<td></td>
<td>478</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 57</td>
<td>transcriptions are</td>
<td>done</td>
<td></td>
<td>447</td>
<td>177, 167</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>Session 58</td>
<td>done</td>
<td>done</td>
<td></td>
<td>535</td>
<td>195, 280</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>Session 59</td>
<td>done</td>
<td>done</td>
<td></td>
<td>135</td>
<td>51, 81</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Session 60</td>
<td>done</td>
<td>done</td>
<td></td>
<td>1153</td>
<td>293, 346</td>
<td>514</td>
<td></td>
</tr>
<tr>
<td>Session 61</td>
<td>progress (AnR)</td>
<td>progress (AnR)</td>
<td></td>
<td>723</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 62</td>
<td>done (DW)</td>
<td>all done (DW, DR)</td>
<td>progress</td>
<td>894</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 63</td>
<td>done (LK)</td>
<td>all done (LK, DS)</td>
<td>done (SES)</td>
<td>106</td>
<td>42, 35</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>Session 64</td>
<td>done</td>
<td>done</td>
<td></td>
<td>221</td>
<td>85, 106</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Session 65</td>
<td>transcriptions are</td>
<td>done</td>
<td></td>
<td>264</td>
<td>97, 112</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Session 66</td>
<td>done</td>
<td>done</td>
<td></td>
<td>130</td>
<td>44, 38</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Session 67</td>
<td>done</td>
<td>done</td>
<td></td>
<td>513</td>
<td>164, 290</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Session 68</td>
<td>done</td>
<td>done</td>
<td></td>
<td>494</td>
<td>11, 356</td>
<td>127</td>
<td></td>
</tr>
<tr>
<td>Session 69</td>
<td>done</td>
<td>done</td>
<td></td>
<td>359</td>
<td>2, 367</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Session 70</td>
<td>done</td>
<td>done</td>
<td></td>
<td>284</td>
<td>0, 282</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Session 71</td>
<td>done</td>
<td>done</td>
<td></td>
<td>1586</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 72</td>
<td>done &amp; being</td>
<td>Nepali done (DKR)</td>
<td></td>
<td>873</td>
<td>125, 274</td>
<td>474</td>
<td></td>
</tr>
<tr>
<td>Session 73</td>
<td>done</td>
<td>Vt done</td>
<td>done</td>
<td>832</td>
<td>227, 69</td>
<td>336</td>
<td></td>
</tr>
<tr>
<td>Session 74</td>
<td>done (DKR)</td>
<td>all done (DKR, DS)</td>
<td></td>
<td>936</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 75</td>
<td>done</td>
<td>done</td>
<td></td>
<td>815</td>
<td>150, 237</td>
<td>427</td>
<td></td>
</tr>
<tr>
<td>Session 76</td>
<td>done</td>
<td>done</td>
<td></td>
<td>723</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The session monitor
Adult vs. child language

```
UTTERANCE COUNTING

length (#utterances) 782

217 by target child 437 by adults 115 by other children

217 of these glossed
```
Status of target child 1

### ALL CHILDREN

<table>
<thead>
<tr>
<th>Category</th>
<th>Translated Length</th>
<th>Translated Duration</th>
<th>Translated Pwords</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phonological</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>406277</td>
<td>203:18:16</td>
<td>152:34:01</td>
</tr>
<tr>
<td>Gossed</td>
<td>313856</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Grammatical</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>207229</td>
<td></td>
</tr>
<tr>
<td>Translated</td>
<td>406277</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gossed</td>
<td></td>
<td>313856</td>
<td></td>
</tr>
<tr>
<td><strong>Duration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>371:14:24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TRANSCRIPTION & TRANSLATION

<table>
<thead>
<tr>
<th>Category</th>
<th>Translated Length</th>
<th>Translated Duration</th>
<th>Translated Pwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>47233</td>
<td>55:40:39</td>
<td>112769</td>
</tr>
</tbody>
</table>

### GLOSSING

<table>
<thead>
<tr>
<th>Category</th>
<th>Glossed Length</th>
<th>Glossed Duration</th>
<th>Glossed Pwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>43142</td>
<td>51:33:18</td>
<td>103512</td>
</tr>
</tbody>
</table>

---

Lieven & Stoll, DGfS Summer School, 2010
Lieven & Stoll, DGfS Summer School, 2010

Media

- tape
  - DIGITIZATION
    - session1
      - .mpg
      - .wav
    - session2
      - .mpg
      - .wav

Text

- TRANSCRIPTION
  - Transcriber
    - session.trs

- TRANSLATION
  - (Nepali & English)
    - ELAN
      - session.eaf

- GLOSSING
  - Toolbox
    - session.txt

- ARCHIVING
  - LAMUS

Metadata

- METADATA-COLLECTION
  - IMDI-Editor
    - session.imdi

- UPDATE
  - METADATA-COLLECTION
    - IMDI-Editor
      - session.imdi

---

**clean1**: (convert .eaf -> .txt)
- check completeness of records
- count utterances
- convert to UTF-8 and Unix linebreaks

**clean2**: count phonol. & gramm. words
- add POS-tags
- clean \\mph-\\mgl mismatches
- add agecodes

---

Lieven & Stoll, DGfS Summer School, 2010

78
Thank you for your attention!