Practical introduction in creating and analyzing child language transcripts and using the CHILDES databases

Filip Smolík (smolik@praha.psu.cas.cz)
Anna Chromá

August 11, 2019

Child language transcripts

- One of the traditional methods in child language research
  - Probably first after diaries
- Material for creating corpora of child language
- Source of dependent variables in experimental work
- Source with good ecological validity

Roger Brown

- A First Language: The Early Stages (1973)
- Analyzed transcripts of Adam, Eve and Sarah
- Introduced the use of MLU
- MLU
  - Mean length of utterance
  - Brown measured in morphemes
- Analyzed the order of emergence of 14 grammatical morphemes
  - Pointed out that frequency in adult lanugage

CHILDES

- Child Language Data Exchange System
  https://childes.talkbank.org/
- Conceived by Brian MacWhinney at Carnegie Mellon University (Pittsburgh)
  - https://www.cmu.edu/dietrich/psychology/people/core-training-faculty/macwhinney-brian.html
- Originally, a repository of child language transripts using a shared format
- CHAT – format used in Childes transripts
- CLAN – a set of analytic software tools

CHILDES content

- A number of transcripts from various languages
- Important to read the documentation

- … walktrough though the web…

- Possible to download or browse the transcripts

Other formats of transcripts

- SALT – used in speech-language pathology, esp. in US
- Language may be transcribed as a part of broader observational data
  - Interact (Mangold), Observer (Noldus)
  - ELAN – free alternative
- You can make up your own
  - The improtant thing is to have the format
    - Clear, unequivocal, consistent

LENA

- Not really language trancript analysis, but it is increasingly use to study communication
  - Device worn by the child
  - Records communication over the whole day, automatically calculates the number of words and syllables produced and heard
  - Can be used to examine the general amounts of communication and its time distribution

Basic format of CHAT transripts

- Three types of lines (“tiers”)
  - Distnguished by the first character
  - Only three characters allowed: @, *, %
  - @ - headers, context, end of file
    - Are relevant for the whole transcript or for a larger section.
    - Minimum content of headers and file intro/ending is described in the minCHAT specification (manual p. 22)
  - * - main tiers, contain the transcript material
    - Often “orthographized”
  - % - dependent tiers
    - Modify, explain, code details of the content of the main tier
    - Always belong to the closest main tier above

Format requirements

- Minimum file structure
  - minCHAT specification (manual p. 22)
  - There is CLAN utility for checking whether the structure
- @Begin, @End
- Identification of participants in the transcript
  - Identification of the target child, should include age
- Languages used in the transcript

General principles

- The manual shows many types of codes
  - You don’t have to use all
  - It gives you a “grammar” of the coding system
    - You can choose which devices you will be using
    - Or adapt and expand the system in line with these “grammatical principles”

@ “regulatory” tiers

- @Begin, @End
- Identifying info in the header
  - @ID:
    - Also defines speaker label, always uppercase 3-letter (CHI)
- Description of the context
  - @Location
- Comments about the context and situation
  - @Comment, @Situation
  - May occur in the middle, refer to a longer stretch of the transcript

* main tier

- Always followed by speaker ID, colon and tab:
  - Must end with . or ! or ?
  - *CHI:	mommy like [*] tea.
- Contain the main transcript material
  - Usually in “adult” orthography
  - Type and level of standardization depend on the project
  - The standardization should facilitate search for various types of content
- You must have a criterion for what is an utterance

Other main tier notes

- Special form markers @
  - samál@f (family specific form)
- Unintelligible material
  - xxx, yyy
- Part words, omissions
  - He 0is sit(ting) here;
- Errors [*]
- Scope codes
  - Eg. <want that one> [!= cries] – paralinguistic material across multiple words

Main tier terminators

- . ! ?
- Can be combined for special meanings
  - +… trailing off
  - +/. Interruption
  - etc…

What is an utterance

- Important decision
- Children will often chain sentences with “and”
  - or without any connective but without pauses
- You need to set criteria

% dependent tiers

- Provide additional information or coding for material contained on the above
  - Many types, new types can be created specifically for a project
    - %err: – describes and explains an error
    - %com – general commentary
    - %pho – phonetic transcript
    - %act – describe accompanying action
    - %cod – general-purose coding line
    - %mor – morphological coding
    - See manual around page 85
  - There can be mutiple dependent tieres for any main tier

Tips for recording transcripts

- Digital recorders common and inexpensive today, do good job
  - You want to test them, test recording modes, test how well they pick up
  - Today, not much reason to use external mics
    - Perhaps if you want the child to wear the recorder
    - But if needed, common clip-on mics are sufficient
  - LED lights on the recorder may attract the child's attention, if the recorder has them, consider covering

Tools

- A good text editor
  - Text editor = word processor (like Word)
  - An editor that shows all contents of a text file
    - Saves no hidden formatting information
  - A lot of them available, typically used by programmers
    - Or people using (La)TeX
  - Notepad++, https://notepad-plus-plus.org/
    - Free to use, easy to install
  - Good for writing and reading transcripts, and also for searching in them

Analyzing transripts

- Usually involves some amount of manual data extraction
  - Reading the transripts and searching for varous things
    - Words, morphemes, structures
  - Completely automatic analysis usually not possible
    - And not desirable, if you want to get a good idea about the data
  - But you can facilitate the analysis with some automatic steps
- The best way: learn Perl or Python
  - But before you do that, you can get pretty far using simpler tools

CLAN

- The basic analytic utility for childes transcripts
- Can be quite complicated and awkward
- But probably the easiest way to calculate basic quantitative indices line MLU
- Analyses have to be done in a command-line style interface

Basic quantitative indices

- MLU
  - Mean length of utterance
    - May be measured in morphemes, words, syllables
    - Words is simple, others require more detailed coding (or syllable boundary estimation)
- Number of tokens
- Number of different words, number of types
- Type-token ratio (TTR)
  - Used to be popular, but depends a lot on the transcript length

MLU in CLAN

- CLAN MLU utility
  - If you don't have the %mor tier, you have to use this
  - mlu @ +t*CHI –t%mor
  - mlu @ +t*CHI –t%mor +d
    - Creates Excel file (excel may complain while opening but it will do it eventually)

Searching for specific material

- Oftentimes you will be interested in specific words, sequences, or codes
  - Find all errors
  - Find all instances of the auxiliary
  - Find all prepositions “on”, “in”, “at”
- Some automation can help you with these tasks
  - Regular expressions in Notepad++

Searching for things

- Find all nouns
  - Impossible unless you have explicit coding
  - Or unless you have a list of words that will be coded as nouns
- Find all instances of “be”
  - Possible if you can list all possible forms of “be” (be, is, was, were, am, are)
- Often we have to manually review the search results
  - But it is still much faster than fully manual search

Search considerations

- It is generally easy to search for specific forms or a small set of forms
  - Harder/effortful if you have a longer list of forms, but possible
    - Probalby some CLAN utility could help you
    - Or you can program a search utility in Perl
- Much can be achieved using text editor search
- All searching is facilitated by regular expressions

Regular expressions

- Allow you to search for patterns of characters
  - Using various “wildcards” and other specifications of what you are searching for
  - E. g. find the string “is” when it is preceded by “John” anywhere on the line
- Especially useful to limit the search to children’s material in the transript
  - E. g. find the string “is” on lines that begin with *CHI:

Examples in Notepad++

- ^*CHI.*(je|jsou)
  - All occurrences of „je“ or „jsou“ on *CHI lines
  - Better put spaces around (je|jsou):
  - ^*CHI.* (je|jsou) ; or even better
  - ^*CHI.*\b(je|jsou)\b – here, \b is a zero-width word-boundary marker
- ^*CHI.*\[*\]
  - Find all children’s lines with error markers
- ^*CHI.*jsou ?[\.\!\?]$
  - Find all children’s lines with “jsou“ at the end
    - Lines ending with ., ! or ?
    - A space may but may not precede the sentence-ending punctuation
    - Does not match sentences with complex ending punctuation (+… etc.)