Reproducible Identification of Pragmatic Universalia in CHILDES Transcripts

GNU meets OpenScience

Daniel Devatman Hromada[123]
daniel@wizzion.com

[1]Université Paris 8 / Lumières
École Doctorale Cognition, Langage, Interaction
Laboratoire Cognition Humaine et Artificielle

[2]Slovak University of Technology
Faculty of Electronic Engineering and Informatics
Department of Robotics and Cybernetics

[3]Universität der Künste
Fakultät der Gestaltung, Berlin

Table of Contents

1. Introduction
   - Psycholinguistics
   - Reproducibility
   - Universalia
2. Corpus, Tools and Method
3. Three analyses
4. Conclusion

Developmental Psycholinguistics

DP

Is a science which uses experimental methods of developmental psychology in order to study acquisition, learning and development of linguistic structures and processes in human children.

Multiple epistemological and methodological problems include:

1. child's behaviour is often very instable
2. the very fact of being subjected to experiment impact child's responses
3. the invasivity problem

These problems do not exist when researcher decides to **observe instead of experiment**!

The Hallmark Principle

Reproducibility

"*Non-reproducible single occurrences are of no significance to science*" (Popper, 1992)

Experimentator-independent reproducibility can be attained *iff*:

1. all experimentators use the same dataset
2. use the same (or least very similiar) set of tools
3. the first experimentator faithfully protocols the usage of such tools
4. other experimentators follow the protocol
5. analysis is deterministic

Pragmatic and Ontogenetic Universalia

Linguistic Universal

A pattern that occurs systematically **across** natural **languages**.

Most common lists of universals, like those of Greenberg (1963), concern syntax, morphology or semantics.

Pragmatic Universal

A L.U. related to pragmatic (extralinguistic context, deictics, etc.) facet of linguistic communication.

Ontogenetic Universalia

Introduce the temporal dimension (age).

Table of Contents

1. Introduction
2. Corpus, Tools and Method
   - Corpus
   - Tools
   - Method
3. Three analyses
4. Conclusion

CHILDES

CHILDES

Child Language Data Exchange System (MacWhinney&Snow, 1985)
http://childes.psy.cmu.edu/data
http://wizzion.com/CHILDES/ (mirror from 6th Feb 2016)

1. more than 50 years of tradition
2. cca 30000 transcripts
3. more than 1.5 GigaBytes of mostly textual data
4. at least 26 languages, dialects or language combinations
5. major terran language-groups (indo-european, ugro-finic, semitic, altaic, east-asian, south-asian) represented
6. Creative Commons BY-NC-SA licence

CHAT format

CHAT system provides a standardized format for producing computerized transcripts of face-to-face conversational interactions. (MacWhinney, 2016; http://childes.talkbank.org/manuals/chat.pdf).

```
@Begin
@Languages:     eng
@Participants:  CHI Eve Target_Child , MOT Sue Mother , FAT David Fathe
@ID:    eng|Brown|CHI|1;6.|female|||Target_Child|||
@ID:    eng|Brown|MOT|||||Mother|||
@ID:    eng|Brown|FAT|||||Father|||
@ID:    eng|Brown|RIC|||||Investigator|||
@ID:    eng|Brown|COL|||||Investigator|||
@Date:  29-OCT-1962
*MOT:   one two three four .
%mor:   det:num|one det:num|two det:num|three det:num|four .
%act:   tests tape recorder
*CHI:   one two three . [+ IMIT]
```

GNU + PERL + R

The idea is to perform the analysis with solely publicly-available open-source **command-line** tools.

GPR combo

- GNU: grep, sort, uniq, sed, wc (runs in bash and connected through pipes)
- PERL: regular expressions are part of language syntax
- R: vectors, matrices, plotting

First command

wget -P CHILDES -e robots=off –no-parent –accept '.cha' -r http://wizzion.com/childes/CHILDES_flat

Pre-processing

Populate filenames with age information

```
mkdir aged; grep -P '\|\d;\d' *| grep Child |
perl -n -e 'chomp; `cp $1 aged/$2-$3-$1`
if /^(.*?):.*0?(\d+);0?(\d+)/;' ; rm *.cha
```

Remove noise

```
perl -ni -e 'print
if $_!~/^\*(MOT|CHI):\t(xxx|www) ?\./' aged/*
```

Extract Child and Motherese utterances

```
mkdir CHI; cp aged/* CHI; sed -i '/\*CHI/! d' CHI/*;
mkdir MOT; cp aged/* MOT; sed -i '/\*MOT/! d' MOT/*;
```

Yields

- 5 833 656 CHI utterances contained in 29180 transcripts
- 3 798 005 MOT utterances contained in 13590 transcripts

Metrics

Main metrics: Probability P_X that signifiant X shall occur in the utterance.

$$P_X = F_X / N_{utterances}$$

where F_X is the absolute number of occurences of X in CHILDES section and the normalization factor $N_{utterances}$ denotes the number of utterances of the CHILDES section.

Probability values are mutually comparable.

Table of Contents

1. Introduction
2. Corpus, Tools and Method
3. Three analyses
   - 1st analysis: Laughing
   - 2nd analysis: Second Person Singular
   - 3rd analysis: First Person Singular
4. Conclusion

Laughing

Objective

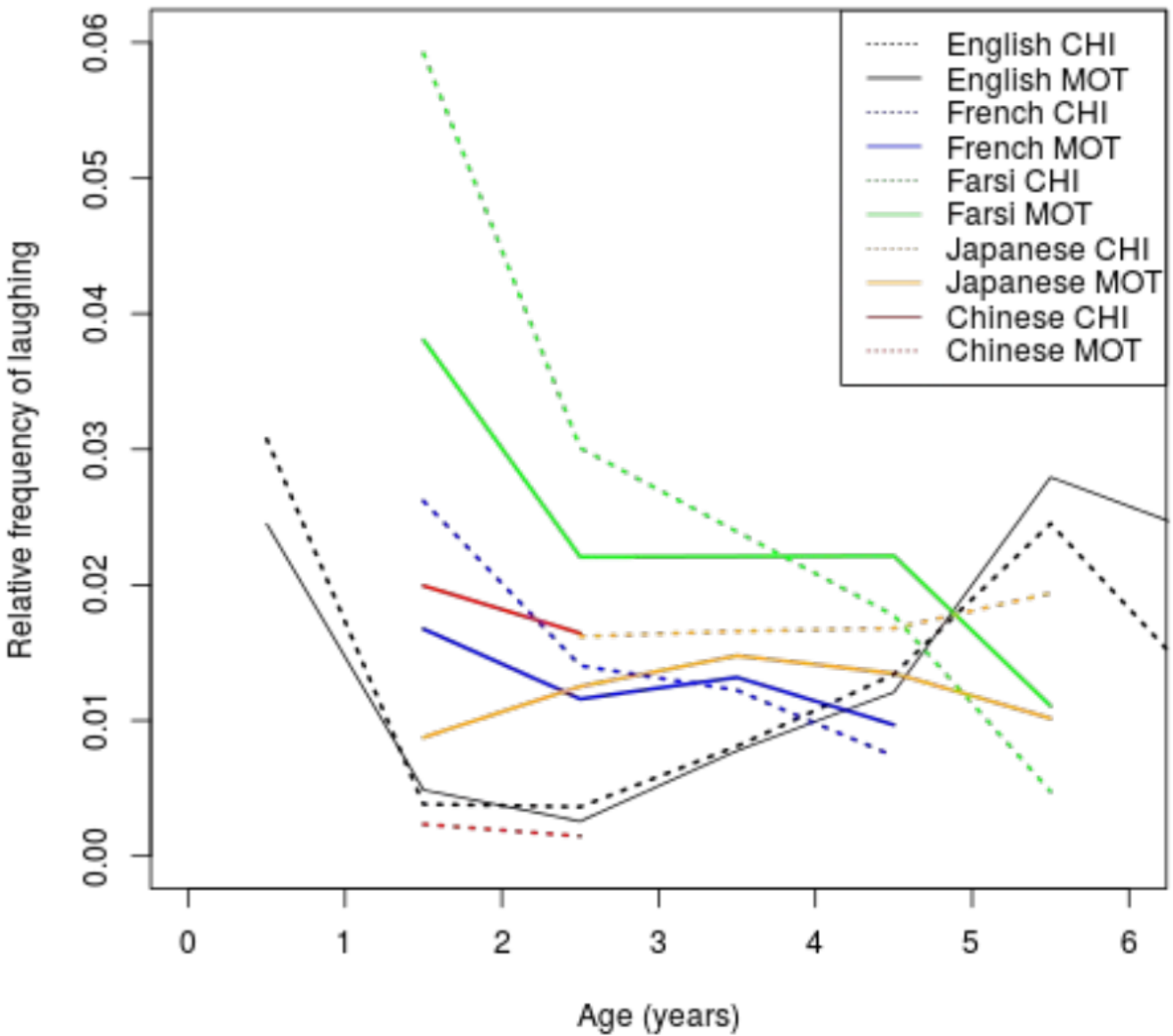
Verify whether observed tendency (Hromada, 2016, Conceptual Foundations) of mothers to laugh less is in interaction with older toddlers is specific to English, or whether it is a culture-independent invariant.

Both **&=laughs** and **=!laughing** tokens are used by diverse CHILDES transcribers, so we simply use for occurences of **laugh** token.

```
grep laugh MOT/*French*|grep -o -P '\-French\-.+\-'|
sort|uniq -c;grep laugh MOT/*Farsi*|grep -o -P '\-Farsi\-.+\-'|
sort|uniq -c;grep laugh MOT/*Japanese*|grep -o -P '\-Japanese\-.+\-'
|sort|uniq -c;grep laugh MOT/*Chinese*
|grep -o -P '\-Chinese\-.+\-' | sort | uniq -c ;

wc -l MOT/*Eng*|perl -e
'while (<>){s/MOT\///;/(\d+) (\d+-\d+)-/;
$h{$2}+=$1; } for (sort keys %h) {/(\d+)-(\d+)/;
print "$h{$_} $1 $2\n";}' >MOT.Eng.N
```

Plot


English CHI
English MOT
French CHI
French MOT
Farsi CHI
Farsi MOT
Japanese CHI
Japanese MOT
Chinese CHI
Chinese MOT
Relative frequency of laughing
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
1
2
3
4
5
6
Age (years)

Some observations

For english, french and farsi children:

- marked decrease of maternal laughing between first and third year of age (english, french, farsi)
- little children laugh more often than their mothers but older children laugh less frequently than their mothers
- significant correlations between MOT and CHI in English (Pearson's cor.coeff 0.933, p = 7.886e-05) and in Farsi (corr. coef. 0.972, p-value=0.02735). Almost significant in French (p=0.053, cor. coef = 0.947)

In regards to laughing, Indo-European mothers and children seem to follow different ontogenetic trajectories than their Japanese and Chinese counterparts

$\Rightarrow$

no culture-independent Universal ?

2nd Person. Sg. Pronouns

Language-specific CHILDES sub-corpora are matched by following Perl-Compatible regular expressions (PCREs):

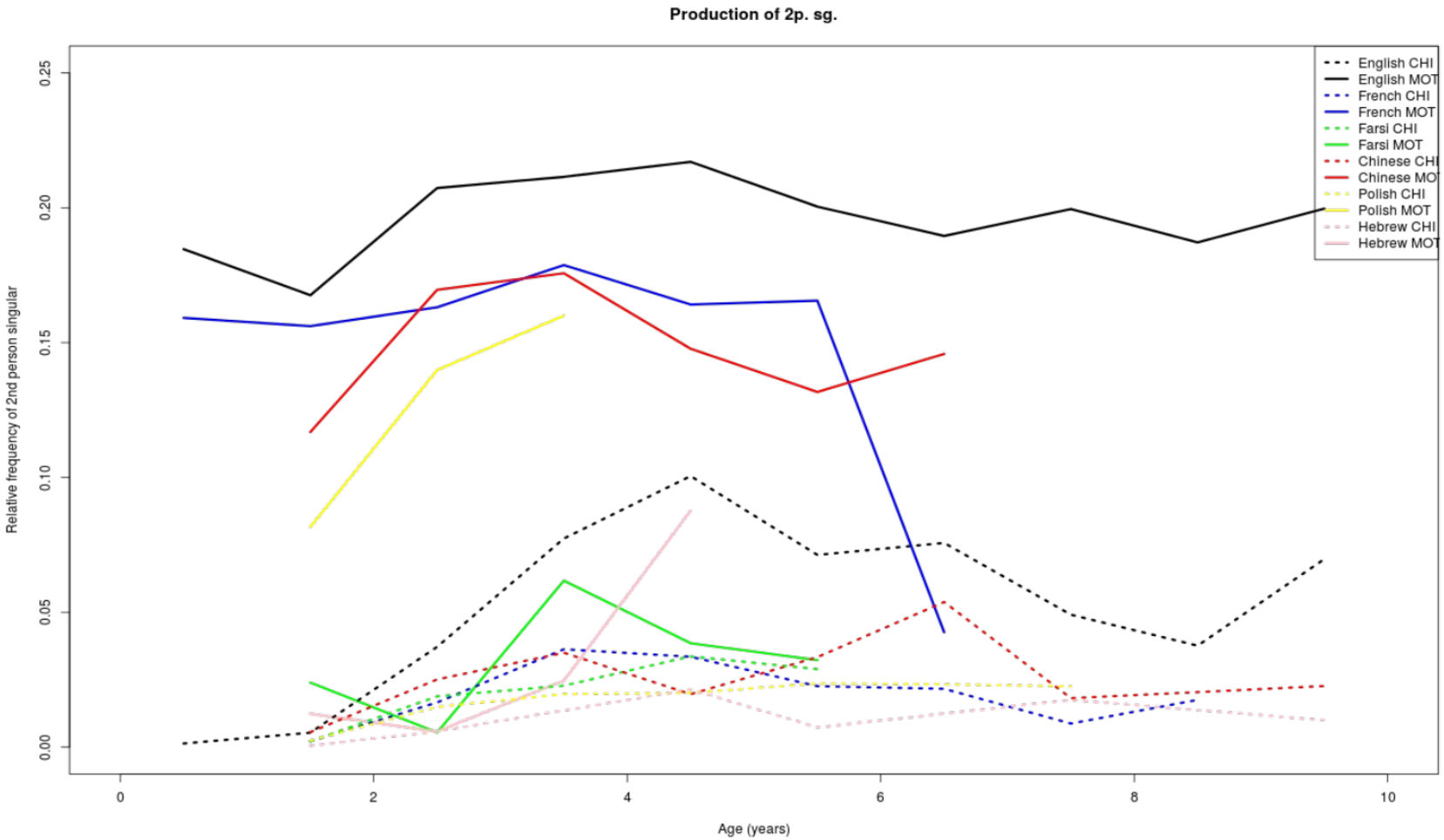
Language	**English**	**French**	**Farsi**	**Polish**	**Chinese**	**Estonian**	**Hebrew**
$\mathbf{PCRE}_{\mathbf{2p.sg}}$	[\t]you[']	[\t]t(u |oi |')	[\t]to	[\t]ty	(你|ni3)	[\t]s(in)?a	[\t]ata?

The absolute frequency F_X of cases when $PCRE_X$ matched is assessed as usually:

```
grep -i -P "[\t ]you[' ]" MOT/*Eng*|
perl -n -e '/MOT\/(\d+)-(\d+)/;
print "$1 $2\n"' |uniq -c >exp2.MOT.Eng.F
```

Subsequently, $F_X/N_{utterances}$ division and plotting are realized in R. (c.f. http://wizzion.com/code/jadt2016/childes.R for the trivial R-code snippet)

Plot


Production of 2p. sg.
Relative frequency of 2nd person singular
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
Age (years)
English CHI
English MOT
French CHI
French MOT
Farsi CHI
Farsi MOT
Chinese CHI
Chinese MOT
Polish CHI
Polish MOT
Hebrew CHI
Hebrew MOT

Some observations

One can observe, in English

- in motherese, "you" is used in cca every fifth utterance
- significant correlation between CHI and MOT time series (Pearson's cor. coeff. = 0.768, t = 3.393, df = 8, p-value = 0.009451; Kendall's tau = 0.6, T = 36, p-value = 0.016671; Spearman's rho = 0.733, S = 44, p-value = 0.02117)

One can observe, in all languages

- Marked increase in maternal usage of 2nd. p. sg. between 1st and 4th year of age has been observed in case of all six studied languages (representing three distinct language groups).
- children use 2nd. p. sg. less often than mothers (only exception: Farsi between 2 and 3)

$\Rightarrow$ ontogenetic Universal ?

1st Person. Sg. Pronouns

Language-specific CHILDES sub-corpora are matched by following Perl-Compatible regular expressions (PCREs):

	English	French	Farsi	Polish	Chinese	Estonian	Hebrew
$\mathbf{PCRE}_{1p.sg}$	[\t]I[']	[\t](j(e |')|moi)	[\t]m[aæe]n	[\t]ja	(我|wo3)	[\t]m(in)?a	[\t]ani

The absolute frequency F_X of cases when $PCRE_X$ matched is assessed as usually:

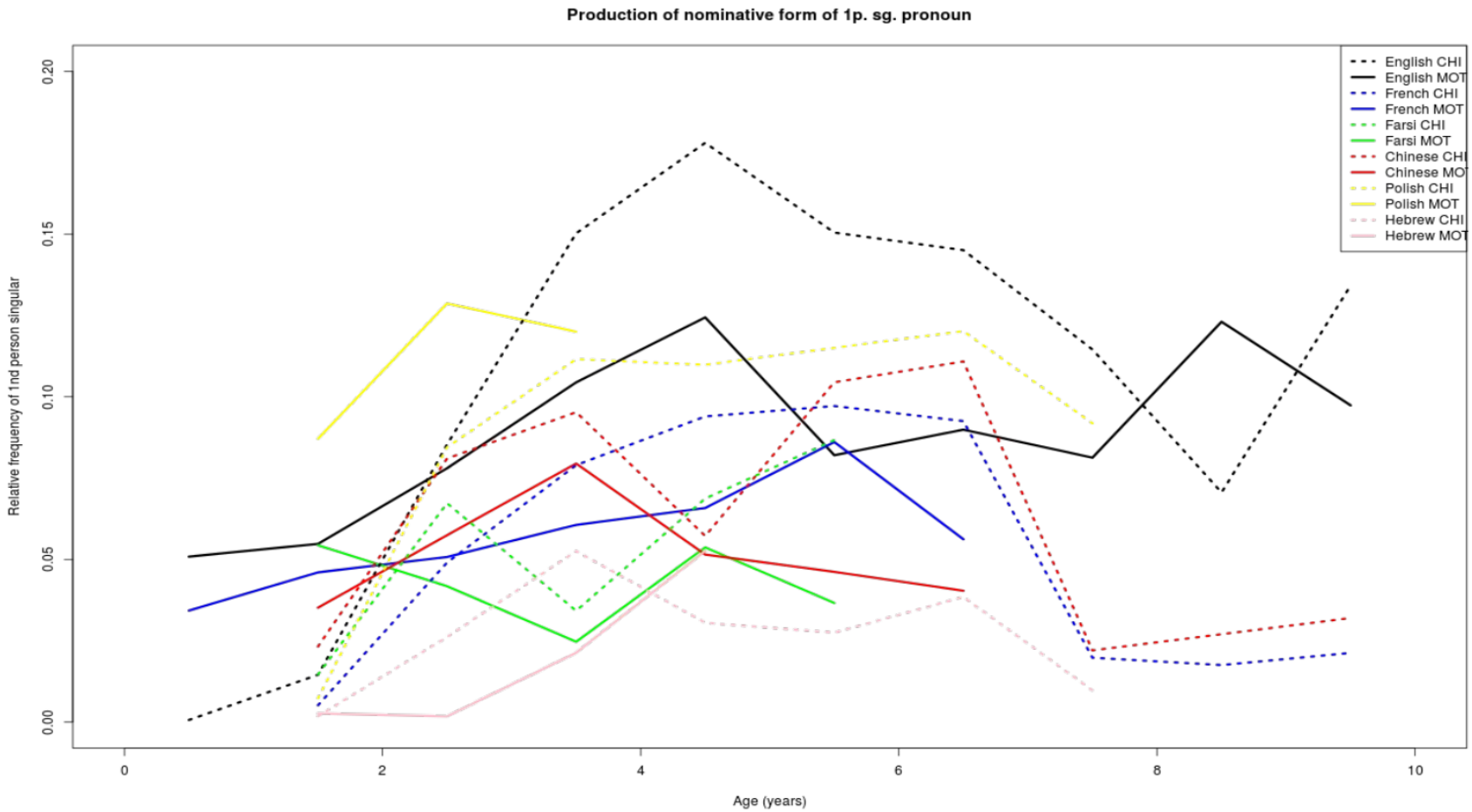
```
grep -i -P "[\t ]I[' ]" MOT/*Eng*|
perl -n -e '/MOT\/(\d+)-(\d+)/;
print "$1 $2\n"' |uniq -c >exp3.MOT.Eng.F
```

Subsequently, $F_X/N_{utterances}$ division and plotting are realized in R. (c.f. http://wizzion.com/code/jadt2016/childes.R for the trivial R-code snippet)

Important: focus on ALL transcripts of a given language.

3rd analysis: First Person Singular

Plot



Some observations

- ALL: around 3 years of age, children tend to pronounce 1.p.sg much more frequently than their mothers
- ALL: steep decline between 6th and 7th year of age (offset of "egocentric" stage?)
- ENGLISH: significant correlation between usage of mothers and children

Significant intercultural correlations

- french and chinese children (p=0.02474)
- english and french children (p=**0.00**2425)
- polish and hebrew children (p=0.048)
- polish and french children (p=0.048)

$\Rightarrow$ language-independent ontogenetic trajectory of usage of 1.p.sg?

Table of Contents

1. Introduction
2. Corpus, Tools and Method
3. Three analyses
4. Conclusion

Methodological conclusion

Combination of

- command-line (no GUI!)
- open-source (for free!)
- fast *
- deterministic

utils (grep, uniq, ...) and languages (PERL, R) yields a 100% reproducible methodology for very little cost.

Experimental protocol automatically stored in **.history** (or .bash_history) and **.RHistory** files: no need to reinvent the wheel!

* 3rd analysis executed on one sole core of 3.2 Ghz PC with 8GB RAM and CHILDES data stored on a SSD disk was over in less than 15 seconds

Epistemological conclusion

Developmental Psycholinguistics + Natural Language Processing
+ Big Data + OpenScience
=
la textometrie psycholinguistique

Manifest:

- to perform state-of-the-art research without expensive tools and apparati
- to study ontogeny of soul and language in a non-invasive fashion
- to share all that can be shared

Psycholinguistic conclusion

Piaget eut raison.

Merci pour votre attention.
Questions ?