Reproducible Identification of Pragmatic Universalia in CHILDES Transcripts

Daniel Devatman Hromada[1,2,3]

[1] Université Paris Lumières - France

[2] Slovak University of Technology – Bratislava - Slovakia

[3] Berlin University of the Arts – Berlin - Germany

Abstract

This article presents method and results of multiple analyses of the biggest publicly available corpus of language acquisition data : Child Language Data Exchange System. The methodological aim of this article is to present a means how science can be done in a highly positivist, empiric and reproducible manner consistent with the precepts of the “Open Science” movement. Thus, a handful of simple one-liners pipelining standard GNU tools like “grep”, and “uniq” is presented - which, when applied on myriads of transcripts contained in the corpus – can potentially pave a path towards identification of statistically significant phenomena. Relative frequencies of occurrence are analyzed along age and language axes in order to help to identify certain concrete, pragmatic universalia marking different stages of linguistic ontogeny in human children. One can thus observe significant culture-agnostic decrease of laughing in child-produced speech and child-directed indo-european “motherese” occurrent between 1st and 2nd year of age; maternal increase in production of pronoun denoting 2nd person singular “you”; increase of usage of 1st person singular “I” in utterances produced by children around 3rd years of age and marked decrease of the same which takes place around 6 years of age. Other significant correlations - both intra-cultural between English mothers and children, as well as inter-cultural - are pointed down always accompanied with thorough descriptions methodology immediately reproducible on an average computer.

1. Introduction

Reproducibility is one of the hallmark principles of occidental science. Being based upon the philosophy of ancient greeks who were fully aware that only the knowlede of that, which repeats itself in many instances, can lead to generic and transtemporal ἐπίσταμαι, the western scientific method necessarily considers reproducibility as its main *condition sine qua non.* In words of the foremost figure of modern epistemology, "*non-reproducible single occurrences are of no significance to science*" (Popper, 1992).

Hence the primary, epistemological, objective of this article is to show how anyone willing to do so can perform reproducible analyses and experiments regarding the phenomena traditionally falling into the scope of corpus, computational and developmental linguistics. This objective is to be quite naturally attained if ever three precepts are stringently followed :

- use publicly available data
- analyse the data with simple, specific yet powerful tools which are well-known to widest possible public
- faithfully protocol the exact procedure of usage of these tools

In more concrete terms, we promote the idea that - in regards to analysis of statistical textual data - core GNU (Stallman, 1985) utils and commands as well as basic operators and core

functions of open source langages like PERL (Wall, 1990) or R (Team, 2013) indeed offer such "simple, specific yet powerful tools well-known to widest possible public".

When it comes to the precept " faithfully protocol the usage of these tools ", it shall be implemented - in this article and potentially beyond – in a following manner : *every simple transformation of data is to be completely and exhaustively described in a footnote which accompanies the description of the transformation*. By " simple ", we mean such a transformation which can be described as a simple standard UNIX shell[1] one-liner pipelining combining together core commands like " *grep* ", " *uniq* " or " *sort* ".

In case of more complex transformations, the complete source code of program is always to be furnished either in publications's appendix or at least as an URL reference. To assure the highest possible reproducibility of the experiment, the snippet should not call any modules and libraries external to language's core distribution (e.g. no CPAN resp. CRAN).

The most important thing, however, is not to forget that the protocol is to be complete, exhaustive and unambigous. That is, ***.history of all steps*** is to be described in the form which is immediately executable on a standard GNU-positive machine. All means all : from the very fact of downloading[2] the corpus from a publicly available source to the very act of plotting the legend on a figure which is then disseminated among scientific communities.

Given that these precepts are followed and under the conditions that

- the analysis is fully deterministic (i.e. does not involve any source of stochasticity)
- the source corpus has not changed in the meanwhile

it can be expected that the same analysis shall bring the same results no matter whether it is executed in other folder of the same computer (e.g. reproducibility across directories) ; executed on different computers (e.g. reproducibility across experimental apparatus) and|or executed by different experimentator (e.g. experimentator-independent reproducibility).

2. Corpus & Method

Child Language Data Exchange System (CHILDES) undoubtably belongs among most fascinating language-related corpora. Established by (MacWhinney and Snow, 1985) more than 30-years ago and including transcripts dating back to 1960s, CHILDES does not cease to be the biggest public repository of child language acquisition and development data. Thus, asides huge volumes of audio and video recordings of verbal interactions with children, CHILDES also contains more than thirty thousand distinct transcripts.

Transcript themselves are encoded in UTF-8 compliant plaintext .CHA files. These files follow a CHAT format specified in (MacWhinney, 2012). Every transcript contains a header describing specificities facts concerning the transcribed scenario – e.g. the age of a child, identities of participants (lines beginning with *CHI denote utterances produced by children; lines beginning with *MOT denote utterances produced by their mothers).

Unfortunately, different linguists have followed the CHAT manual in a different manner. For example, some include the timestamp information into their corpus and some not. Some mark the repetition by special tokens like [x 2] (for duplication) or [x 3] (for triplication) and some

[1]$ echo 'All footnote-descriptions of shell one-liners begin with the sign $ and all footnote-descriptions of R commands begin with sign >.'

[2]It is highly recommended to use standard utilities like "wget " or "curl " for that purpose.

transcribe the utterance as such, without using such tokens. And yet another set of differences necessarily originates in transcriber's own perception and habits. For example: while the token "*mama*" is occurrent in 1405 child utterances contained in English sections of the corpus[3], some other English transcribers (e.g. Haggerty or Suppes) apparently prefered to transcribe the mother-directed vocative as "*mamma*" - this occurs in 126 distinct utterances.

Be it as it may, the CHILDES corpus is already so huge that one may except that a well constituted and unbiased quantitative analysis could potentially allow the discovery of phenomena robust to any surface perturbations (e.g. differences in habits and styles of different investigators etc.). In other terms, if *every transcript is understood as a result of a distinct act of sampling*, then it can be expected that the statistical aggregation of such a huge amount of distinct samples (> 30000 distinct transcripts) could let to situation where the noise cancels itself out and statistically significant phenomena emerge.

And individual CHILDES transcripts are indeed distinct. Not only because dozens, if not hundreds researchers and investigators of at least three or four generations had already directly participated on constitution of the corpus. Not only because majority of transcripts were in one way or another related to a specific research project with a goal unrelated to goals of other projects. But also because investigators themselves, as well as the investigated subjects (e.g. children), often stem from huge variety of distinct cultural backgrounds. More concretely: 26 languages are included in the corpus, covering practically majority of main terran language strata (i.e. indo-european languages, asian languages, semitic, altaic and ugrofinic languages etc.). This allows for trans-cultural analysis and such shall indeed be all analysis presented in the section 4.

2.1 Metrics

Results can be mutually compared and communicated only if they are expressed in common units. In case of all experiments presented in this article, the relative frequency - interpreted as the probability of occurrence - of pattern X is such a unit. This is equivalent to absolute frequency of occurrence of F_X normalized by the total number of utterances, i.e.

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

Ideally, for every month mentioned in the CHILDES corpus should correspond one P_X value. To understand our approach more clearly, imagine, for example, in case of hypothethic language whose speakers utter 100 utterances each month since their birth until their tenth birthday. If such speakers utter the token " dog " twenty times every month, than the value of all 120 (i.e. 10 years * 12 months) datapoints describing the time series for this particular token would be constantly equal to 100/20 = 20% = 0.2.

It is principially due to such trivial nature of the calculus hereby presented that the core data-mining procedures can be performed directly on the BASH command-line.

3.2 Preprocessing

Four hundred and sixty-seven megabytes of data compressed in 983 zip files are obtained after the corpus has been downloaded from its original source[4] or from a mirror site which

---

[3] `$ grep "mama" child/*Eng* |wc -l; grep "mamma" child/*Eng* |wc -l`

[4] `$ wget -P CHILDES -e robots=off --no-parent --accept '.zip' -r http://childes.psy.cmu.edu/data/`

represents state of CHILDES as of February 6th 2016[5]. After these files are recursively decompressed[6], the CHILDES arborescent structure is flattened so that all .CHA files are contained within one sole directory[7]. A following one-liner subsequently "peeks into" each .CHA file, retrieves child's age from it and puts this information into files' name[8].

Utterances containing only *xxx* and *www* tokens – which, according to CHILDES manual denote "unintelligible words with an unclear phonetic shape" resp. "untranscribed material" - are removed from all child and mother transcripts[9]. Next step is executed only to speed-up following pattern extraction processes: child utterances are funnelled into simplified transcripts stored in "CHI" subdirectory and maternal utterances are funnelled into "MOT" subdirectory [10] . Translocutory information is thus lost but this is allowed for the purpose of this article in which we shall focus solely on relative frequencies of certain tokens and not on more complex discourse units.

All this yields 5833656 lines (e.g. utterances) contained in 29180 non-empty simplified transcripts stored in "child" directory and 3798005 lines contained in 13590 non-empty simplified transcripts stored in the "mother" directory.

Note that metadata like age (years and months), language group, language and CHILDES investigator's identity are stored directly in the simplified transcript's filename. Workbench common to all following analyses can be thus considered as ready.

3. Analyses

3.1. First Analysis – Laughing

It has been recently indicated that English mothers interacting with children younger than 16 months tend to laugh significantly more often than mothers which interact with children between 16-31 months of age (p.222, Hromada, 2015). Our 1st analysis will use CHILDES to address this hypothesis from a trans-cultural perspective.

It may be surprising to use a dataset, which is essentially a linguistic corpus for, a purpose of study of such a non-verbal means of communication as laughing definitely is. But the very CHAT manual (p.62, MacWhinney, 2012) explicitely specifies the **&=laughs** marker as a most common standardized spelling denoting a specific extralinguistic event.

Unfortunately, within the totality of CHILDES corpus, the marker itself **&=laughs** is not the only standardized form denoting the phenomenon and some authors prefered to use markers

[5]$ wget -P CHILDES -e robots=off --no-parent --accept '.zip' -r WILL-BE-GIVEN-IN-CAMERA-READY-VERSION

[6]$ find CHILDES/data -name "*.zip" | while read filename; do unzip -o -d "`dirname "$filename"`" "$filename"; done

[7]$ mkdir CHILDES_flat; find CHILDES/data -type f |perl -n -e 'chomp; if (/\.cha/) {$f=$_; s/\//-/g; s/\.-data-//g; `cp $f ./CHILDES_flat/$_`;}'; cd CHILDES_flat;

[8]$ 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

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

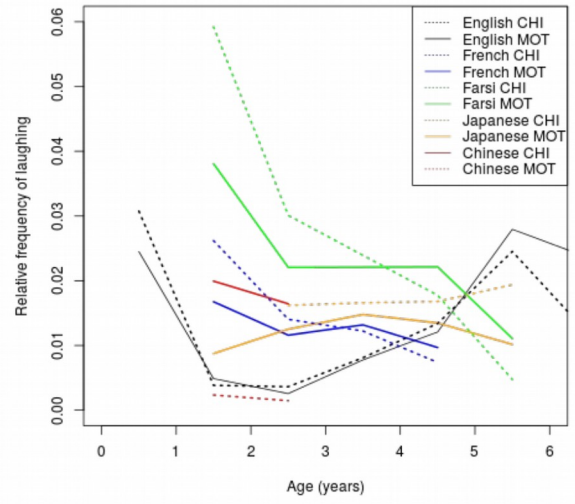
[10]$ mkdir CHI; cp aged/* CHI; sed -i '/*CHI/! d' CHI/*; mkdir MOT; cp aged/* MOT; sed -i '/*MOT/! d' MOT/*;

like **[=! laughing]**. Hence, for a purpose of our 1st analysis, we have simply used the token **laugh** as the one whose frequencies of occurrence we have decided to measure.

Three indo-european (english, french and farsi) and two non-indo-european languages (japanese and chinese) were chosen in order to address the developmental trajectory of laughing from a trans-cultural perspective. For each among these langages, a *target investigator* was identified as the one who most frequently used the marker **laugh** in his transcripts of motherese[11]. Corpus subsections " Farsi-Family ", "French-MOR-York ", " Japanese-MiiPro " and " Chinese-Beijing " were thus identified as such target subsections. All English-language transcripts (i.e. such files whose filename contains the token " Eng ") were also taken into account.

The core of the procedure is as follows: total amount of utterances is obtained, for each month and each target subsection of the corpus, by a one-liner[12] which redirects its output into a file whose every row contains three space-separated columns: first column denotes the denotes the value of $N_{utterances}$ and second and third column denote the year resp. month. The procedure is to be repeated ten times alltogether, five for each target corpus subsections multiplied by two possible locutor values of the locutor variable (MOT[13] or CHI[14]).

Follow ten executions of a command sequence which generate 10 files containing absolute frequencies of occurrence of the token laugh within five different corpus sections – and again for both MOT[15] and CHI[16] locutors - which are aggregated according to child's age in the moment when laughing was noted down by the CHILDES investigator. And that's it: all result-containing files can now serve furnish input datasets for the R code which produces a plot displayed on adjacent figure.



Probability that laughing accompanies or substitutes an utterance produced by, or directed to, a child of specific age.

[11]$ 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 ;

[12]$wc -l MOT/*Farsi-Family* |perl -e 'while (<>) { s/MOT\///; /(\d+) (\d+-\d+)-/; $h{$2}+=$1; } for (sort keys %h) {/(\d+)-(\d+)/; print "$h{$_} $1 $2\n";}' >exp1.MOT.Farsi-Family.N

[13]$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";}' >exp1.MOT.Eng.N

[14]$wc -l CHI/*Eng* |perl -e 'while (<>) { s/CHI\///; /(\d+) (\d+-\d+)-/; $h{$2}+=$1; } for (sort keys %h) {/(\d+)-(\d+)/; print "$h{$_} $1 $2\n";}' >exp1.CHI.Eng.N

[15]$grep laugh MOT/*Eng* |perl -n -e '/MOT\/(\d+)-(\d+)/; print "$1 $2\n"' |uniq -c >exp1.MOT.Eng.F

[16]$grep laugh CHI/*Eng* |perl -n -e '/CHI\/(\d+)-(\d+)/; print "$1 $2\n"' |uniq -c >exp1.CHI.Eng.F

Potentially the most salient phenomenon is a marked decrease in production of laughs which occur between birth and second year of age. This could be potentially explained in terms of gradual switch from non-linguistic means of communication towards more verbal interactions. However, in case of child-directed speech of Japanese motherese the relative frequency of laughing seems to increase during the same period and in case of chinese, the decline is much less marked than in case of indo-european langages. This may potentially suggest an intercultural difference – a hypothesis which is further corrobated by the fact that it is only in case of indo-european langages that the " dotted " lines cross with " solid " lines. Id est, little english-, french- and farsi- speaking children tend to laugh more often than their mothers but older children seem to laugh less frequently than their mothers.

This quiproquo notwithstanding, relative frequencies of CHI time series significantly correlate with MOT time series in both English (Pearson's correlation coefficient 0.933, t = 7.36, df = 8, p-value = 7.886e-05) and in Farsi (corr. coef. 0.972, t = 5.9224, df = 2, p-value = 0.02735). In French correlation is quite close to significancy threshold (t = 4.1692, df = 2, p-value = 0.053, cor. coef = 0.947) when data is aggregated in year-sized packages but is insignificant (t = -1.1598, df = 27, p-value = 0.2563) when time series are correlated with monthly granularity. No statistically significant correlation between child-produced and mother-produced laugh time-series has been observed in case of Japanese or Chinese.

3.2. Second Analysis – 2nd person singular

It has also been indicated that English mothers interacting with their children tend to use the pronoun for 2nd person signular " you " much more frequently than is the case in standard linguistic communication (p.218, Hromada, 2015).

Similiarly to our 1st analysis, our 2nd analysis uses CHILDES to address this hypothesis from a trans-cultural perspective. The procedure is thus very similar to the one already presented with one major difference : we do not focus on assessement of occurrences of one standard marker (e.g. " laugh ") which is present in different corpus sections ; but rather look for, in each specific subscorpus, for a specific Perl Compatible Regular Expression, a ($PCRE_{2p.sg}$) which matches nominative forms of 2nd person singular in the langage of subcorpus under study. Following table lists 6 cases of such PCREs for matching 2p.sg. in 6 languages.

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

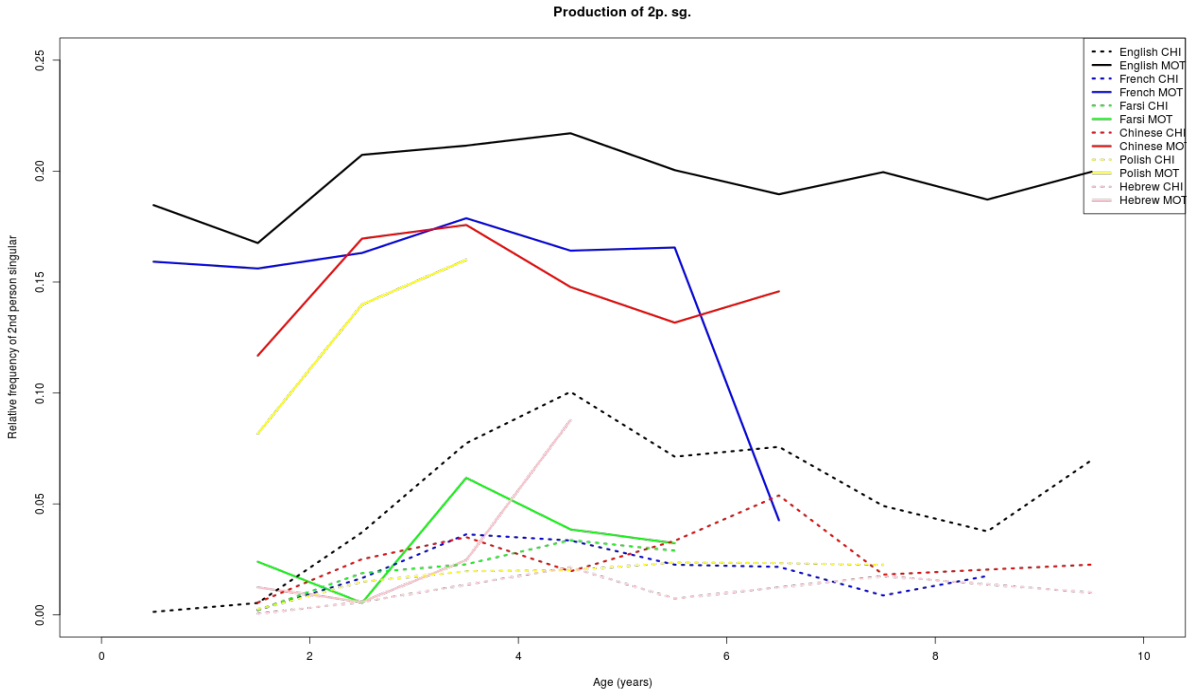
Usage of these regexes within one-liners using the case-insensitive " grep " allows us to obtain distributions of relative frequencies independently for MOT[17] and CHI[18] utterances.

Command sequence yielding distributions of $N_{utterances}$[19] is practically the same as in first analysis (c.f. footnotes 13 & 14), the only difference being due to the fact that this time we do not focus on subcorpora which represent transcripts done by specific target investigators, but

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

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

rather process much bigger datasets containing all transcripts representing the langage under study. $F_{PCRE2p.sg}$ and $N_{utterances}$ distributions are subsequently processed by the R code which is, mutatis mutandi, identic to R code snippet used in analysis 1. This yields Figure 2.



A phenomenon common to all languages under study can be observed practically immediately. That is, on all six solid MOT lines, one can observe, between first and fourth year of child's age, a marked increase in maternal usage of 2nd. person singular. Sometimes such an augmentation is less marked (as in french), sometimes it comes later (between 2nd and 3rd year of age in case of farsi and hebrew), but it always comes. And it always reaches all-time-heights before fifth year of age, after which the maternal usage of "you" tends to slowly converge back to its "normal" levels.

Note also that in English motherese, " you " is used in approximately every fifth utterance. What is also striking in regards the English language - which is definitely the biggest CHILDES subcorpus - is quite significant correlation between time-serie representing the usage of 2p. sg. by mothers and time-serie representing the usage of 2p. sg. by children themselves (Pearson's cor. coeff. = 0.768, t = 3.393, df = 8, p-value = 0.009451; Kendall's τ = 0.6, T = 36, p-value = 0.01667[20]; Spearman's ϱ = 0.733, S = 44, p-value = 0.02117).

3.3. Third Analysis – 1st person singular

Our 3nd analysis is identic to the second, the only thing which changes are the PCRE patterns which are this time supposed to match nominative forms of pronous denoting the 1st. person

[19]$wc -l CHI/*Farsi*|perl -e 'while (<>){s/CHI\///;/(\d+) (\d+-\d+)-/;$h{$2}+=$1;}for (sort keys %h){/(\d+)-(\d+)/;print "$h{$_} $1 $2\n";}' >exp2.CHI.Farsi.N

[20]>cor.test(aggregated_**mot_lang1**[,6]/aggregated_**mot_lang1**[,3],aggregated_**chi_lang1**[,6]/aggregated_**chi_lang1**[,3],method="kendall")

singular. Id est the ego, the self-reference, the "I". Following table lists 7 cases of such PCREs matching 1p. sg. in their respective CHILDES subcorpora.

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

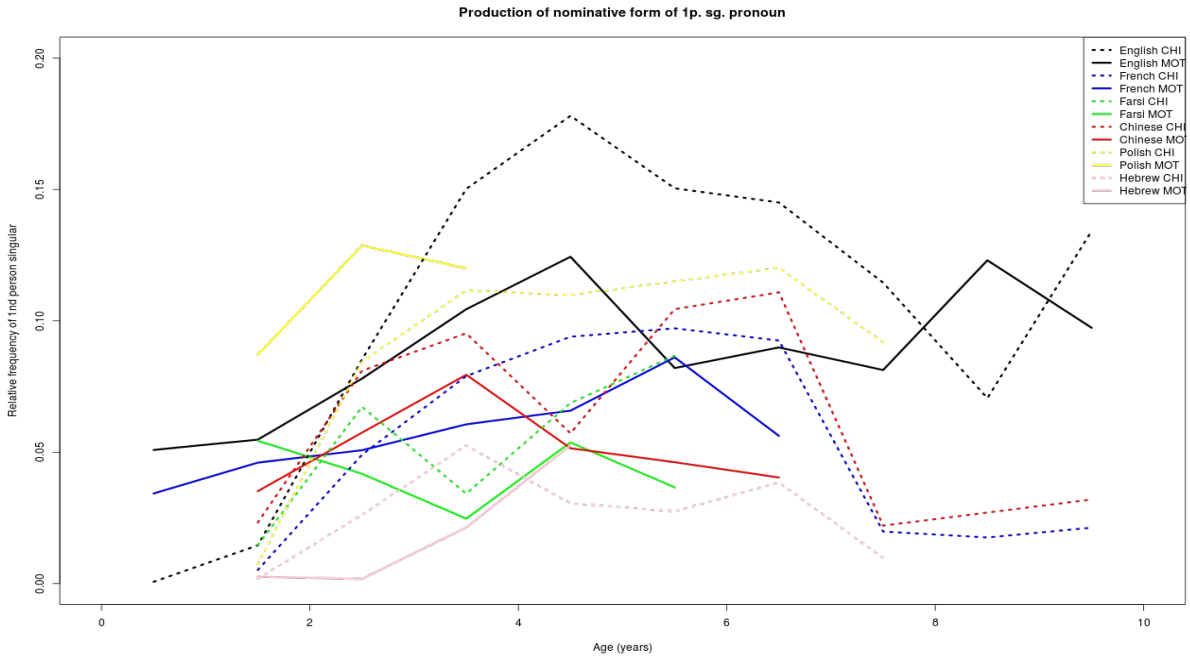
Everything else - from extraction of absolute frequencies of forms matched by PCREs all the way to aggregating, normalizing and plotting - is, mutatis mutandi, identic to 2nd analysis.

This leads to visualisation presented at the bottom of this page. An interestant phenomenon can be noticed: while in early infancy, mothers of all language backgrounds use 1p.sg. much more frequently than children (probably because children are still in a pre-linguistic stage), the difference is being switfly and strongly counteracted. Hence, around three years of age, children of all[21] cultures tend to produce 1p. sg. much more frequently than their mothers.

But not only augmentation of use but also diminutions are of certain scientific interest. Hence, a steep decline in use of 1p.sg. can be observed between 6th and 7th year of age. That is, during the period when children and enter school and which markes the offset of that ontogenetic stage which (Piaget, 1951) labeled as "egocentric".

Similiary to 2nd analysis, a significant correlation between time serie representing the production of "I" by english-speaking mothers and production of "I" by english-speaking children can be observed (Kendall's $\tau = 0.555$, $T = 35$, p-value = 0.02861).

What's more, the plot indicates a path towards identification of **statistically significant inter-cultural correlations**. Thus, after filling the gap[22] in the Chinese dataset related to the fact



[21] With exception of Polish language where we unfortunately lack motherese data from 3rd birthday onwards.

[22]>aggregated_chi_lang4[9,]=(aggregated_chi_lang4[7,]+aggregated_chi_lang4[8,])/2

that CHILDES does not seem to contain transcripts of chinese 8-year olds, one shall observe a correlation[23] between time-series of relative frequencies of 1p.sg produced by french and chinese children (Kendall's τ = 0.511, T = 29, p-value = 0.02474). Idem for english and french (Kendall's τ = 0.777, T = 32, p-value = 0.**00**2425), for polish and hebrew (Pearson coef. = ; Kendall's τ = ; Spearman's ϱ = 0.786, S = 12, p-value = 0.04802) and if one stays faithful to canonic $p<0.05$ precept (Fisher, 1925) and opts for Spearman's rho or Pearson's coeff rather than for Kendall's tau, then, for example then also for french and polish (Pearson coef. = 0.837, t = 3.4219, df = 5, p-value = 0.0188 ; Kendall's τ = 0.619, T = 17, p-value = 0.06905 ; Spearman's ϱ = 0.785, S = 12, p-value = 0.04802) as well as for polish and hebrew (Pearson coef. = 0.759, t = 2.6117, df = 5, p-value = 0.04757; Kendall's τ = 0.619, T = 17, p-value = 0.06905 ; Spearman's ϱ = 0.786, S = 12, p-value = 0.04802[24]) .

4. Discussion

It is a common practice in contemporary Corpus Linguistics in general and in Natural Language Processing in particular, to focus fully on formal and theoretical properties of one's model or analysis. Thus, majority of publications in these domains limit themselves to dissemination of few core formulas behind the analysis which is presented + results which were obtained (F-scores etc.). In atmosphere where sharing the code with the community is more an exception than a rule, it is not surprising that majority of publications disregard the concrete aspects of implementation and execution of one's analysis as unworthy of interest.

Such an attitude can be excusable when one attacks a highly specific engineering problem. But in regards to analyses aiming to attain the general knowledge - id est, when *doing fundamental research or exploratory science* – such an approach is to be discarded as inconsistent with the ideal of experimentator-independent reproducibility.

In this article, we have explained how cost-efficient (i.e. as free as open source software), reproducible and transparent science can be performed at the very border of corpus and developmental psycholinguistics. More concretely, in footnotes of this article, we have presented less than two dozens one-liners which pipeline and combine PCREs (Wall, 1990; Hromada, 2011) with core GNU utilities like “grep”, “uniq”, "wc" and “sort”. Asides this, a snippet of few dozen lines of beginner-level non-optimized R code is hereby being published[25] in order to furnish complete – i.e. from downloading the corpus from publicly available source all the way to final plots and correlation coefficients - description of three experiments hereby performed.

Common to these three experiments was a preprocessing phase which purified and repartitioned hundreds of megabytes of data contained in CHILDES. Result of this phase were two directories, CHI which contains utterances produced by children and MOT which contains motherese utterances (cf. section 2.2). Principal motivation behind this repartitioning

[23] >cor.test(aggregated_**chi_lang2**[,6]/aggregated_**chi_lang2**[,3],aggregated_**chi_lang4**[,6]/aggregated_**chi_lang4**[,3],method="kendall")

[24] >cor.test(aggregated_chi_**lang6[**,6]/aggregated_chi_**lang6**[,3],aggregated_chi_**lang5**[,6]/aggregated_chi_**lang5**[,3],method="spearman")

[25] http://wizzion.com/code/jadt2016/childes.R

was a speed-up of any subsequent analysis. For example the 3rd analysis - when executed on one sole core of 3.2 Ghz PC with 8GB RAM PC and CHILDES data stored on a SSD disk (a fairly standard configuration) - didn't last more than 15 seconds. All the way from matching the first regular expression on the first line of first transcript to R's final plotting.

Mentioning regular expressions, we consider it as important to reiterate that regexes, like those implemented in Perl or PCREs, seem to us to be much more than impressive yet weird character sequences that no neophyte can read. Unambigously denoting what they should denote - i.e. a specific set of character sequences, a specific pattern, schema and form - *PCREs are formalisms in their own right* (Hromada, 2011). Idem for shell commands and PERL or R instructions - they also are unambigous formalisms and for purposes of NLP, they can turn out to be *at least* as worthy as other formalisms.

Formalisms, tools and methodology being thus defined by a concrete example, a question can be posed: "What should be the name of a discipline which uses implemets such a method and uses such tools ?" And given that what was done used techniques common to textometry in order to address topics common to developmental psycholinguistics (Tomasello, 2009), an answer could potentially sound: "**Textometric Psycholinguistics**".

It is only now - with toolbox specified and reproducible method and scope of interest of discipline properly delimited - that a discussion about culture-independent anthropological constants occurent in adult-child verbal and pre-verbal interactions - id est a discussion about "linguistic universalia" and their meaning, a discussion among *savants* can, hopefully, begin.

References

Fisher, Ronald Aylmer. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

MacWhinney, Brian & Snow, Catherine. (1985). *The child language data exchange system*. Journal of child language, 12(02), 271-295.

MacWhinney, Brian. (2012). *The CHILDES Project Tools for Analyzing Talk–Electronic Edition Part 1: The CHAT Transcription Format*.

Piaget, Jean. (1951). *Principal factors determining intellectual evolution from childhood to adult life.* Columbia University Press.

Popper, Karl. (1992). *The Logic of Scientific Discovery.* Routledge, London.

Hromada, Daniel Devatman. (2011) *Initial Experiments with Multilingual Extraction of Rhetoric Figures by means of PERL-compatible Regular Expressions.* RANLP Student Research Workshop, 85-90.

Hromada, Daniel Devatman. (2015). *Conceptual Foundations: Intramental Evolution & Ontogeny of Toddlerese. In press.*

Stallman, Richard. (1985). The GNU manifesto.

Team, R.Core. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013.

Tomasello, Michael. (2009). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Wall, Larry. (1990). PERL: Practical Extraction and Report Language.