Identify the speech code through statistics: a data-driven approach

Andrea Briglia, Massimo Mucciardi, Jérémi Sauvage

Book of Short Papers SIS 2020

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Identify the speech code through statistics: a data-driven approach

Identificare il codice linguistico attraverso la statistica: un approccio empirico

Andrea Briglia, Massimo Mucciardi, Jérémi Sauvage

Abstract Language is what makes humans a unique species of «symbolic animals» by providing them a way to convey meaning through sounds, and it is undoubtedly one of the pillars of our lives, yet we learn it so spontaneously and effortlessly that it is impossible to remember how we came up in its mastery or to give any account on any stage of its acquisition. Thanks to recent advances in data storage, information visualization and automated processing (*e.g.* data mining), there is a growing interest in cutting-edges researches between statistics and linguistics aimed at unfolding the "linguistic genius" of babies by testing hypotheses mining large spoken longitudinal datasets in order to understand - by means of an inductive procedure - the way each of us learnt his/her language without being aware of it.

Abstract *Il linguaggio è ciò che rende gli esseri umani una specie unica nel suo essere degli "animali simbolici" perché ci fornisce un modo di trasmettere significati attraverso suoni. Esso è indubbiamente di fondamentale importanza nella vita di ognuno, ma lo impariamo in un modo così spontaneo che è impossibile ricordare come ne abbiamo acquisito la padronanza così come è impossibile spiegare una qualsiasi delle tappe del suo apprendimento. Grazie ai recenti sviluppi tecnologici nella memorizzazione, visualizzazione e trattamento automatico di grandi quantità di dati è nato un crescente interesse verso studi che combinano statistica e linguistica per spiegare il c.d "genio linguistico" dei bambini verificando tale ipotesi su corpus longitudinali tramite una procedura induttiva.*

Key words: Phonetic Variation Rate; CHAID Model; First Language Acquisition

[1] Andrea Briglia, Univ. of Montpellier "Paul Valery"; email: abriglia@unime.it;
Massimo Mucciardi, Dep. of Cognitive Science, Univ. of Messina; email: mucciard@unime.it;
Jeremi Sauvage, Univ. of Montpellier "Paul Valery"; email: jeremi.sauvage@univ-montp3.fr

1. Identify the speech code through statistics

The so-called “linguistic genius” of babies (Kuhl, 2010) is a matter of long-standing debate in the scientific community: being adults, we realise how much easier is for a toddler to spontaneously learn every kind of language compared to adults’ struggle to maintain a sufficient mastery of a foreign language required – for instance - to attend international conferences. A number of evidences (Saffran, 2003) show that infants are geniuses because it is hypothetically plausible that they can rely on “statistically biased learning mechanisms” (Saffran, 1996) and on an “automatic” pattern recognition neuronal “device”. So, a fundamental question arises: “What is it about the human mind that allows a young child, merely one year old, to understand the words that induce meaning in our collective minds, and to begin to use those words to convey their innermost thoughts and desires?” (Kuhl, 2010). Every infant is facing a huge challenge: learning a sound system made up of many units that - combined together in an almost infinite set of combinations – gives rise to an arbitrary relationship between sounds’ sequences and meaning. According to Saffran’s metaphor, the task is the following: « You must discover the underlying structure of an immense system that contains tens of thousands of pieces, all generated by combining a small set of elements in various ways. These pieces, in turn, can be combined in an infinite number of ways, although only a subset of those combinations is actually correct. However, the subset that is correct is itself infinite. Somehow you must rapidly figure out the structure of this system so that you can use it appropriately early in your childhood» (Saffran, 2003). In fact, the balance between speed and accuracy in learning a language should be of primary importance in the survival of an individual: for this reason we supposed that human brains have been evolutionarily selected to their specific way of detecting regularities and patterns from the external world in order to retroactively syntonize their cognitive potentialities to the environment (Friston, 2010). The literature on « perceptual attunement» (Fort et al., 2017) demonstrates how early children become familiar with their mother language by focalising their speed and accuracy of the recognition task on what they have been experienced to and – symmetrically – by losing the capacity to readily detect and decode unfamiliar cues. So - as language is acquired through cognitive mechanisms that we could consider to be analogous to statistical engines that store probability distributions and formulate predictions based on means and expectancies on what has been previously stored - our attempt is to try to uncover what we have called “statistical learning” by mining a set of longitudinal *corpora* in french language.

2. Data structure and model

Colaje-Ortolang (2020) is an open access french database, part of the broader CHILDES project (2020): seven children have been recorded in a natural setting one hour every month, from their first months of life approximatively until six years old.

Data are available in three different formats: IPA, orthographic norm and CHAT (acronym for Code for the Human Analysis of Transcription), each of them is aligned to the correspondent video recording, allowing researchers to see the original source and to eventually reinterpret every utterance on their own. The main coding structure of the database consists in the fundamental division between "*pho*" (what the infant really says) and "*mod*" (what the infant should have said according to the adult's standard phonetic/phonological norm): we define "variation" every occurrence in which "*pho*" differs from "*mod*". How much the density of the sampling can influence the range of deductions and generalizations that we could draw from data is a debated question: is there a threshold beyond which the sampling is sufficiently representative and, *a fortiori*, any logical implication from it will be empirically valid? The answer depends on the level of analysis, in other words: the scale at which we want to focus on (Tomasello, 2004). In linguistics there are many scales: from the most basic units such as vowels and consonants to complex syntactical constructions. In a *corpus* such as the MIT Media Lab's pioneering "Human Speech Home project" (Roy, 2006) that consists in 400'000 hours of audio and video recordings, every level of analysis will be granted by a strong empirical support, as nearly everything the infant said has been recorded. The sampling of the "Paris Corpus" (Morgenstern et al., 2012) is obviously many times less statistically representative (one hour every month, that is roughly 0.5 – 1% of what the infant listen and say during the sample period, assuming he is awake about ten hours per day) so it probably could not provide sufficient empirical support to highly specific research on particular lexical phenomena or the emergence of specific syntactic structures, but on the other hand we think it suffices to provide a fair statistical support in order to account to the more general phonetic units' level, as well as the emergence of word categories such as pronouns, articles and determinants, being the probability of finding at least one target from any given sample higher for more basical units (Tomasello, 2006). Further, the age span is wider and – having been recorded 7 infants by using the same research protocol – it gives to researchers an easy way to compare development's intercourses between infants. Goal is to verify whether and how "any variation does not randomly vary into any other, but it rather should follow an underlying pattern, as every variation has an order in itself" (Sauvage, 2015). We first import 4 *corpora* of a single child named "Adrien" at 3 years and 1 month of age (time 22), 3 years and three months (time 24) and then time 27 and time 34. To turn raw data in a computationally and statistically tractable format we unbundle them into a data structure in which every sentence appears on the row side and every word on the column side. In table 1 are summarized the main statistics for 4 *corpora*: we can see how a quantitative increase in the number of words and length of sentences in which these words are combined causes an increase in S.D. that is due to a parallel increase in the lexical variability (type/token ratio) that – in turn - expands the range of possible variations a child can utter.

Table 1: Corpus statistics

Time	Mean	Length	S.D.
22	2.64	343	1.80
24	2.80	324	1.76
27	3.34	580	2.39
34	5.89	641	4.28
Total	3.98	1888	3.32

Mean = average number of words within a corpus; Length = length of the corpus; S.D.= standard deviation of the number of words within a corpus

Consequently, considering a single phrase of a *corpus*, we define "phonetic variation rate" (PVR) the ratio between the number of phonetic variations (NPV), that is the number of differences detected between "pho" and "mod", on the total numbers of words (TNW). In formula, for the phrase "i" and the total numbers of words "j": $PVR_{ij}=NPV_{ij}/TNW_{ij}$. In this way, by appropriately setting the subscript "j", we obtain for each corpus the PVR_j which represents the phonetic variation rate considering a definite number of words "j". Table 2 summarizes the results of the PVR considering j = 1, 2, 3,4,5 and 20 (max number of words in a single sentence.) From table 2 we can see how nonlinearity affects language acquisition: globally, PVR decreases over time but counterintuitive phenomena such as regressions (Morgenstern et. al, 2012) are frequent: it could happen that a child mispronounces something that he had previously correctly pronounced. The same holds for PVR over sentence's lenght: we expect (and observe) that rate increase as the lenght increases, but there are some exceptions to the norm that could require a specific account.

Table 2: Main statistics for PVR by time and number of words

Time	Statistics	PVR_1	PVR_2	PVR_3	PVR_4	PVR_5	PVR_20
22	Mean	0.477	0.556	0.655	0.513	0.667	0.577
	Length	*132*	*62*	*56*	*40*	*18*	*343*
	S.D.	*0.501*	*0.416*	*0.311*	*0.299*	*0.322*	*0.415*
24	Mean	0.494	0.528	0.525	0.538	0.608	0.553
	Length	*79*	*90*	*68*	*39*	*26*	*324*
	S.D.	*0.503*	*0.362*	*0.322*	*0.247*	*0.268*	*0.371*
27	Mean	0.558	0.532	0.563	0.471	0.440	0.483
	Length	*154*	*108*	*87*	*86*	*50*	*580*
	S.D.	*0.498*	*0.388*	*0.284*	*0.247*	*0.239*	*0.359*
34	Mean	0.305	0.281	0.244	0.278	0.260	0.246
	Length	*82*	*57*	*71*	*89*	*63*	*641*
	S.D.	*0.463*	*0.341*	*0.270*	*0.208*	*0.232*	*0.266*

In a second step, we used CHAID (Kass G., 1980) to get a general insight on how PVR changes over time and which kind of phonetic units are correctly articulated and which are not. From the results obtained[2], we can clearly see how time is the main regressor because it splits most part of the *corpus*, then the lenght of sentences

[2] All statistical analyses were performed using R, Excel and SPSS. In the CHAID model, cases are weighted by TNW.

plays a role as well, as we can observe in the *corpus* "time 34", where the fourth word causes the formation of an additional branch to the tree. The main pattern CHAID has detected in a "blind" way is the morphological difference between phonemes: as we can see from the tree table of the CHAID model (table 3), in the node 15 (PVR_20 mean 0.971, variation rate very high) words are longer and contains many "r" and couples of consonants, sounds typically learnt later in development.

Table 3: Tree table for CHAID model (main results - first and last three PVR_20 values)

Node	PVR_20 (Mean)	N	Primary Independent Variable	p-value	Split values
15	**0.971**	68	w_mod_1r	0.000	ɑ̃kɔʁ; sɛlsi; spidœʁma; isi; vjɛ̃; pʁefɛʁ; boku; bɔ̃ʒuʁ; vwatyʁ; vɛʁt; kaʁgo; osito; ɛskaʁgo; pjɛʁo; by; tʁwa; katʁ; sɛ̃k; sis; sɛt; ɔ̃z; duz; tʁɛz; katɔʁz; kɛ̃z; sɛz; dissɛt; dizɥit; diznœf; vɛ̃; vɛ̃teœ; vɛ̃tdø; vɛt; vɛ̃tkat; tɛ; tetɛ; kwɛ̃kwɛ; kwɛ̃; ʁjɛ̃; kɔʁnəmy; flœʁ; vɛʁ
20	**0.918**	255	time	0.000	22
4	**0.880**	490	w_mod_1r	0.000	etɛ̃; ty; sɔʁ; muje; lə; ɑ̃kɔʁ; lwi; sɛlsi; spidœʁma; akʁoʃe; otuʁ; sali; tɔ̃be; uvʁ; dɛʁjɛʁ; pɔʁt; isi; sɥisi; alɔʁ; ɑ̃; adʁijɛ̃; aj; tɛ̃kjɛt; naomi; puʁ; lotʁ; metɛ; zafiʁa; sypɛʁ; desine; mɔ̃tʁ; nunuʁs; dɔʁmevu; ʒak,
30	**0.079**	165	w_mod_1r	0.000	wi; la; ø; œ̃; bɛ̃; komɑ̃; dø; ba; duz; tʁɛz; katɔʁz; dɑ̃; noemi; tɛl; twa; kwa; ə; tjɛ̃; konɛ; ɛm; ka; pe; y; ve; igʁɛk; zɛd; ɛn; potiʁɔ̃; kɑ̃guʁ; s; sɥila; paʁl; tʁo; tabul; taʁɛt
24	**0.033**	152	w_mod_2r	0.000	la; vø; papa; apɛl; bum; dudu; mamɑ̃; sa; lə; akemi; dɔn,
27	**0.025**	119	w_mod_2r	0.000	nɔ̃; le; papa; lə; isi; bys; ʒoli,

while in the node 11 (PVR_20 mean 0.267 – not shown) words are shorter and contain more vowels and bilabials (e.g. "ma","ba") and - more generally - sounds pronounced by using the external part of mouth (easier to learn because infants can spot them by seeing them and thus providing cues for imitation, unlike sounds such as "r" or "l" who are articulated at the bottom of the throat and thus they have to be deducted by the child). We wrote "blind" because CHAID cannot distinguish morphological differences between phonemes, yet it performs a remarkable result simply by calculating interactions between occurrences. In conclusion, in this paper we have shown how the use of the CHAID model could provide us a way to analyse and evaluate child language development in a quantitative manner. Our results are sufficiently coherent to the state of the art of phonetic units acquisition (McLeod et

al. 2018). The main limit is that this technique doesn't take into account morphological differences, as the PVR is calculated on the difference between "pho" and "mod", regardless of what they represent linguistically: in order to overcome this limit we start to use Python to analyse *corpora* according to a predetermined list of phonetic units to track and quantify every variation, then we turn them into a "Multistream graph" (Cuenca et.al, 2018). These are the future directions of our research, once again we are trying to combine statistics and linguistics to try to test whether and how "any variation does not randomly vary into any other, but it rather should follow an underlying pattern, as every variation has an order in itself" (Sauvage, 2015).

References

1. Cuenca E., Sallaberry A., Wang Y., Poncelet P. « *MultiStream : A Multiresolution Streamgraph Approach to explore Hierarchical Time Series* ». IEEE Transactions on visualization and computer graphics, vol.24, no. 12. (2018)
2. Fort M.; Brusini P. ; Carbajal J. ; Sun Y. ; Peperkamp S. "*A novel form of perceptual attunement : Context-dependent perception of a native contrast in 14-month-old infants* ». Developmental cognitive neuroscience 26 , 45-51. (2017)
3. Friston. K. "*The free energy principle: a unified brain theory?".* Nature reviews. Neuroscience. Vol 11. February 2010. 127. Ref to the "Bayesian brain hypothesis"
4. http://colaje.scicog.fr/index.php/corpus cited 20 Feb. (2020)
5. https://childes.talkbank.org/ cited 20 Feb. (2020)
6. Kass, G.V.:. An Exploratory Technique for Investigating Large Quantities of Categorical Data". App. Statist 29(2):119-127 (1980)
7. Kuhl P. K. « *Brain Mechanisms in Early Language Acquisition* ». Neuron 67, (5). 713-727. (2010)
8. McLeod S.; Crowe K.. "*Children's Consonant Acquisition in 27 Languages: A Cross-linguistic Review*". American Journal of Speech-Language Pathology. 1-26. (2018)
9. Morgenstern A.; Parisse C. (2012), « *The Paris Corpus* ». French language studies *22*. 7-12. Cambridge University press. Special Issue. P11
10. Roy. D et al. « *The human speech home project* ». International Workshop on Emergence and Evolution of Linguistic Communication. Springer. Heidelberg. (2006)
11. Saffran J. "*Statistical language learning: Mechanisms and Costraints*". Current directions in Psychological science. 2003 Vol.12 No 4. P 110-114. (2003)
12. Saffran J. R ; Aslin R. N ; Newport E. L ; « *Statistical learning by 8-Month-Old infants* », Science, vol. 274, december. 1926-1928 (1996)
13. Sauvage J. « *L'acquisition du langage : un système complexe* ». L'Harmattan, Louvain la neuve. P103. (2015).
14. Tomasello, M. and Stahl, D. « *Sampling children's spontaneous speech: How much is enough?* » . Journal of Child Language, 31:101–121. (2004).