# Statistical and computational approaches to first language acquisition. Mining a set of French longitudinal corpora (CoLaJE)

Andrea Briglia

▶ **To cite this version:**

Andrea Briglia. Statistical and computational approaches to first language acquisition. Mining a set of French longitudinal corpora (CoLaJE). Linguistics. Université Paul Valéry Montpellier 3; Università di Messina (Italie), 2021. English. NNT : . tel-03319126

**HAL Id: tel-03319126**
**https://hal.science/tel-03319126**

Submitted on 11 Aug 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Joint PhD program in Cognitive Sciences and Linguistics (XXXIII cycle)

Università di Messina & Université “Paul Valéry” Montpellier 3

# Statistical and computational approaches to first language acquisition. Mining a set of French longitudinal *corpora* (CoLaJE)


Candidate: Mr Andrea Briglia

Supervisors:

Prof. Massimo Mucciardi

*Università di Messina*

Prof. Jérémi Sauvage

*Université “Paul Valéry” Montpellier 3*


| **Jury** | **Affiliation** | |
|---|---|---|
| *Mme A.Morgenstern* | *Sorbonne Nouvelle* | *Rapporteure* |
| *MrYvan Rose* | *Memorial University* | *Rapporteur* |
| *Mr Francesco Cutugno* | *Università “Federico II” Napoli* | *Rapporteur* |
| *Mr J-P Chevrot* | *Université de Grenoble Alpes* | *Examinateur* |
| *Mme Paula Fikkert* | *Radboud Universiy Nijmegen* | *Examinatrice* |
| *Mme Giovanna Marotta* | *Università di Pisa* | *Examinatrice* |
| *Mme Valentina Cardella* | *Università di Messina* | *Examinatrice* |
| *Mr Christophe Parisse* | *Université Paris Nanterre* | *Examinateur* |
| *Mr Francesco La Mantia* | *Università di Palermo* | *Examinateur* |

***9/3/2021***

*Ai medici, agli infermieri e a tutte le persone*
*che lavorano ogni giorno per fronteggiare questa pandemia.*
*Alla scienza, e alla fiducia dell'umanità in essa.*

# Acknowledgements/Ringraziamenti/Remerciements

The first thanks go to Francesco Bianchini, my former supervisor in Bologna. He has no role in this thesis, but he had the opennes to talk with me for hours during spring 2017, when I was only a master student in Anthropology wandering in Zamboni street with no clear direction. Bumping casually in his office after having finished reading Michel Foucalt's "Histoire de la folie" simply changed the course of things in my life. He was one of the fewest to think that I was not mad at the time and the only person that suggested me to apply for a PhD. Because after all, after these three years of Phd, I am still convinced that an "eternal Golden Braid" lies between the form of a *Romanesco broccoli* and human cognition.

I am grateful to professor Massimo Mucciardi and professor Jérémi Sauvage: they have been simply great supervisors. You have thaught me the vast majority of what I have wrote in my thesis, I hope that this will be enough for your expectations. I spent hours with you learning everything I was capable to learn from you, hours behind this screen that I'm becoming convinced that you are wise voices behind this laptop! I hope that this lockdown will end as soon as possible: we will finally able to meet all together, either drinking a glass of *rosé* in Place de la Comédie or tasting some Nero d'Avola on the Strait of Messina!

I am grateful to Giovanni Pirrotta for his expertise on any computer science issue I could ever imagined. My relation with softwares is a quite complicated one, without your speed I would probably end this thesis in some years! I hope to meet you in the real life once this pandemic will be finished!

I am grateful to Christelle Dodane: you've helped me many times and you gave me the opportunity to teach in your Corpus Linguistics class. Thanks for your trust, I hope that I have contributed to the CoLaJE corpus in an interesting way for all the users community

I thank Alì, Cwiosna, Chafik, Bea, Cecilia and all the friends in Montpellier. I am here since a couple of years and I'm feeling at home with you. Thanks for our soirées spent together, for our long coffee breaks and for the happy moments in this beautiful country.

I am grateful to Giancarlo Luxardo and Francesca Frontini for their help and suggestions and for the relaxing moment we spent together walking around in the garrigue and in the Cévennes.

I am grateful to professor Naomi Yamaguchi and professor Christophe Parisse for our meetings in Sorbonne and Nanterre: they know CoLaJE better than anyone else and their advices have been extremely useful. Thanks for having sent me your articles and your thesis: they have been the starting point of what I have wrote here. I hope that my thesis will contribute to the improvement of CoLaJE and I will be ready to continue improving this work and give in an "open access" format all the graphs that you think could interest the research community.

I am grateful to professor Sandra Bringay from Montpellier data science lab and to her master students team: Florian, Marine, Sariaka and Clément. Our collaboration has been long and fruitful, it took time and patience to learn from each other but in the end we succeeded in making a great interdisciplinary work.

I am grateful to professor Arnaud Sallaberry for giving me the opportunity to use his Multiresolution Streamgraph for representing CoLaJE datasets. You spent your time explaining me the basis of this graph and how to put data into it: it has been a quite complicated procedure that still need some improvement, but the current result seems to be promising.

I am grateful to professor Hugo Alatrista for his expertise on data science and for having explained me the basis of some useful application of Python programming on my datasets. I hope to have the opportunity to take a break and join you at the Universidad del Pacifico, sooner or later!

I am grateful to professor Paula Fikkert for her advices during our meeting in Radboud University. You have been helpful and kind and I hope to have an opportunity to visit Nijmegen again. Research excellence in this city is simply wonderful and I hope to improve myself to get there as an active participant of some groundbreaking project.

I am grateful to Matteo Colombo from Tilburg University and Sander Lestrade from Radboud University: I had short conversations with you about your articles but this revealed to be extremely useful for the introductory part of this thesis.

I cannot forget what coming from Emilia-Romagna to Sicily would unexpectedly bring me: here I've found the same willingness to change society that I found in Bologna with On. Elly Schlein. A huge thanks goes to the people who never lose hope and never get tired: Laura Carlino, Paolo Putrino, Alessio Grancagnolo, I loved our time spent together in Catania. My only regret is to not have been with you one year before, during what has become the history of political debate in Italy!

A special thanks goes to On. Claudio Fava and his team for their invaluable support, to professors Mauro Sylos-Labini and Claudio Vannucci for their precious advices, to Leonardo Ferrante for our long-lasting collaboration in fixing every kind of problems and to Dario Montana for his wisdom. "Libera" is a great community, it represents hope and I will always be eager to join you in the next collective effort.

A special thanks goes to ADI, Associazione Italiana Dottorandi/e. Giuseppe Montalbano, Matteo Piolatto, Giuseppe Naglieri and Giulia Malaguarnera. I thank you all for the support in these three years, our Union is stronger and will become more if we gather our forces together. Good luck to Luca dell'Atti for his new mandate, I hope to see you all for our next "Buena Onda" summer camping!

I thank my parents and my sister to the patience they have had during these three years: I know I can become a very susceptible guy, my temperament needs more free time than I actually have. Thanks for the support, thanks for having always granted me the possibility to study whatever I want despite the "curious" professional future that graduate people in Humanities usually have.

A huge hug goes to all my friends in Massa, my hometown. I would like to visit you more times than I did last three years. I always live in a contradiction since I left home in 2012: come back or continue traveling around. Who knows which the next step will be. Thanks for being always there, thanks for our long phone calls, thanks for all the time spent together, thanks for having grown up all together as a great group of brothers: I am who I am thanks to you. You are my backbone.

أشكرك إيمان جزيل الشكر على دعمك خلال هذه الفترة .أنا سعيد بلقائك .آمل أن تكون هذه هي المرة الأخيرة التي استخدم فيها المترجم الآلي للتعبير عن نفسي .الأطروحة القادمة ستكون حتما حول اللغة العربية

## AUTHOR'S DECLARATION

I, Andrea Briglia, declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the Department of Cognitive Sciences at the University of Messina, Italy, is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution.

SIGNATURE:

DATE: November 15, 2020

PLACE: Montpellier, France

# Table of contents

List of acronyms

- CoLaJE (Communication Langagière chez le Jeune Enfant) former ANR research program that collected all the data used in this thesis. P.I Aliyah Morgnestern (University of Paris 3)
- CHAT (Code for the Human Analysis of Transcription)
- CLAN (Computerized Language Analysis)
- CHAID (Chi Squared Automatic Interaction detection)
- EM (Expectation Maximization) the clustering method used
- SPVR (Sentence Phonetic Variation Rate)
- LD Levenshetein Distance
- NLD (Normalized Levenshtein Distance)
- CTWT (Child Total Words Tokenized)
- CTDW (Child Total Distinct Words)

**Key words**: First Language Acquisition, Data Mining; Visual Modelling of Child Spoken Transcripts; Corpus Linguistics; Phonology; Decision-Tree Classification Model; EM Clustering; Parsing;

# Introduction

In a very abstract form, first language acquisition could be viewed as a mixture of deterministic and random processes. This combination results in a probabilistic process.

It is deterministic because rules and constraints applied to human cognition are partly known. The anatomical parts involved in language perception and production, as well as the basic laws of acoustics, influence the way we let sounds carry our inner thoughts in form of conventional meanings without any control over them.

It is partly a random process because the amount of variability between children and within a single child is largely acknowledged (Vihman, 2014, p280) and represents – at the same time – what is interesting and what is difficult in child language studies.

Knowing these rules and constraints does not allow us to predict the outcome of a child beginning to be immersed in his/her native language. All we know is that around the age of 5/6, s/he will master his/her own language/s. We know approximately the learning stages, the date of his/her first word, and the rough order of consonant acquisition. Interesting theories have been developed about the patterns of errors the child will most likely make, such as overregularizations of certain verb forms (Markus & Pinker, 1992) or sequences of varied form of diphthong and coupled consonants (Sauvage, 2015), but it is – to date – impossible to model language acquisition. This is because it is a non-linear process, too complex to be reduced into a black box, made up of a set of algorithms that would – in one way or another – reproduce the inner working mechanisms of our brain.

Nowadays, the verb "to know" is becoming a synonym of the expression "to be able to reproduce": in fact, if you understand how a given entity works, then you will be able to create a model of it – in either an analogical or digital form – and finally formulate a prediction that will be exactly simulated and, *a fortiori*, confirmed by your model.

In this thesis many statistical treatments and computational algorithms will be used: the aim is to find patterns and regularities in longitudinal *corpora* made up of children spoken language transcripts and, by doing so, to account for language acquisition in terms of

describing possible preferential learning paths, outlining constraints by comparing children between them.

In the framework of an inductive reasoning path toward learning, the goal is to see whether and how children from the CoLaJE *corpora* display similar or different developing trajectories of language acquisition.

This is different from using algorithms to find out patterns and regularities in children's spoken language and then claiming that – as these patterns were discovered by using a specific set of algorithms – children's inner learning mechanisms work in the same way (or even a similar one). As the same outcome can derive from different causes; there is no evidence that there must be a relationship whenever two results are qualitatively and quantitatively similar.

For these reasons, it is always better to clearly specify which is the model and which is the reality to be modeled. In this thesis the reality is represented by the recordings taken from CoLaJE. They are temporally ordered samples of an undergoing cognitive development that structures itself month after month.

Chi-Squared Automatic Interaction Detection (CHAID), Expectation Maximization clustering (EM), all the statistical indexes, algorithms, graphs and the simple neural network proposed in my thesis are only plausible models of children's spoken language *corpora* but they are not thought to be a substitute of reality. These are different ways of representing reality: a series of phonetic variation rate means (linguistic errors over time) could be viewed as a simplified version of reality, EM clustering could be seen as a different way of ordering grammatical properties and so on, but they do not represent a way to account for a child's learning mechanisms. This is done mostly by neuroscientists through different epistemological frameworks, data collection and techniques.

The aim of this thesis is to give new methods for evaluating children's language acquisition and new ways of representing it, and not to explain the inner workings of acquisition. Throughout this thesis manuscript, there will be some hypothetical paths proposed in light of the results of automated computing techniques, but these hypotheses will never be tested and validated *strictu sensu,* by using a scientific method.

Clements's "Theory of phonological traits" will be used as a useful tool to better account for quantitative results, as a way to "connect the dots" and form a more coherent image. However, this theory will be used more as a description rather than an explanation.

What we claim is that first language acquisition is an inherently complex phenomenon that we need to understand *via* the lenses of the theory of complex thought. For this reason, the first part of this thesis will briefly propose a set of basic notions about the theory of complex thought by the French philosopher Edgar Morin (2010).

Then the key concepts of Bayesianism in cognitive science will be described as it is an important topic in this domain. Simple analogies will be made between the Bayes' formula and supposed learning mechanisms though this framework was not chosen to be a model of first language acquisition or to describe the results obtained within my work.

Here I provide a general review of the state-of-the-art of cognitive science, which is thought to help the following lines of reasoning and, hopefully, future researches.

# Chapter 1 - “Theory of complex thought”

## 1.1 What is complexity

Giving a definition to a complex system represents a difficult task and it may even reveal unnecessary to have one. The word “complex” being an adjective proper to various scientific domains, every discipline has its own definition. A common denominator to all these versions could be the one proposed by H. Simon, according to which complex systems are:

“made up of a large number of parts that have many interactions”

and

“in such systems the whole is more than the sum of the parts in the […] sense that, given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole.” (Simon, 1981)[1].

Complex systems often display different layers of internal organization, either hierarchically ordered or not. Being these levels mutually interdependent, the typical analytical procedure which consists in dividing the whole in its elementary components to understand – in a bottom-up perspective – the system’s structure and functioning, could not work. This is because the high degree of interdependence makes every component draws its “role” or “meaning” not from itself, but rather from the dynamical interrelation between the other elements, both internal and external to the system.

In his essay “Filosofia della complessità[2]” Italian philosoher G. Gembillo well described the ontological difference between a closed and deterministic system and an open, probabilistic system by giving two examples: a clock can be disassembled to analyse its smaller

[1] Simon, H.A. (1981) « The Sciences of the Artificial ». MIT Press: Cambridge

[2] Gembillo G. (2011) “Filosofia della complessità”. Le Lettere, Firenze.

components and gears and then be re-assembled many times without damaging its structure or its functioning. This does not work in the same way for any living organism: after having been "disassembled" it would be impossible to understand its parts, because each of them would lose the properties it draws from being intimately dependent on the other elementary components. For example, it would not be useful to try to understand the respiratory system if it would be studied separately from the cardiovascular one, being the complementarity among the two systems more than essential.

So, why do living systems are so complex and where does this complexity come from?

It is hard to provide an answer to the existence of this huge complexity, as every answer would be partial and temporary. It may be simply said that the high complexity of living organisms could be an adaptive response, a coping strategy to evolutionary and selective pressures. Survival is often linked to the ability to model the outside world: catching the external information by decoding and recoding it in a different form requires complex cognitive structures whose task is to spot and process huge quantity of information which, in turn, will be elaborated to inform action. This simple draft of a possible retroaction between cognition and environment is hard to frame: is it pre-reflexive, reflexive or metacognitive?

In other words, how much intentionality is there? Can we evaluate the degree of awareness that organisms are known to have? Which is the nature of the force that pushes life to maintain itself despite entropy's relentless growing power?

Could all this complexity be the result of a fundamental dialectic between order and chaos? Or between necessity and contingency?

Are all these questions surreptitiously biased by humans' innermost cognitive tendency to teleonomy?

To put this question with the words of J. Monod (Monod, 1970, p256):

« *Nous disons que ces altérations sont accidentelles, qu'elles ont lieu au hasard. Et puisqu'elles constituent la seule source possible de modifications du texte génétique, seul dépositaire à son tour des structures héréditaires de l'organisme. Il s'ensuit nécessairement que le hasard seul est à la source de toute nouveauté, de toute création dans la biosphère. Le hasard pur, le seul hasard, liberté absolue mais aveugle, à la racine même du prodigieux*

*édifice de l'évolution : cette notion centrale de la biologie moderne n'est plus aujourd'hui une hypothèse, parmi d'autres possibles ou au moins concevables. Elle est la seule concevable, comme seule compatible avec les faits d'observation et d'expérience. Et rien ne permet de supposer (ou d'espérer) que nos conceptions sur ce point devront ou même pourront être révisées"*[3] .

For example, language differences could be conceived through these lenses: language families exist as we know them, but they would not be exactly as they are if more language contacts would have taken place if, for example, the Alps did not exist. Without such a huge physical obstacle we could hypothesise that such a sharp division between Romance languages and German languages would not exist, simply because the flow of people would have played a role in mixing the languages.

So how can there be science if almost everything is due to chance?

Animals exchange with each other a large variety of sound, olfactory or visual cues, allowing them to maximise their chance to survive. But it seems that they do signal more than communicate, this because animals' way of exchanging information is almost innate and based on routines, they do not show the creativity humans do (Sievers C. et al., 2017).

Humans produce speech by using two anatomical – physiological structures that – phylogenetically speaking – were there well before we acquired the capacity of "carrying meaning through sounds[4]".

Lungs were firstly designed for breathing and then for singing, glottis was designed to impede food fall into respiratory system and then to module our voice from yelling to whispering. Sphincter vocal folds could be viewed as resulting from a compromise between the innermost ability to chew food and the ability to change the airflow coming from lungs in

[3] Monod J.(1970 ). « Le Hasard et la Nécessité : Essai sur la philosophie naturelle de la biologie moderne ». Editions du Seuil. Paris coll. « Points essais ». p 256

[4] Plebe A.; De la Cruz V. « Neurosemantics. Neural processes and the construction of linguistic meaning ». Springer Studies in Brain and Mind. 10.

order to produce a range of different sounds. From this perspective, accidental choking could be viewed as a "side effect" of this hypothetical form of "exaptation[5] ".

This point will be at the core of the experience-independent *vs* experience-dependent debate (known even as Nature-Nurture debate[6]) on language acquisition for which I will give an introduction providing main authors and references that had opposed their views and arguments during last decades

The dynamical interactions that form the structure of a complex system determine a behaviour (or functioning) that should be considered as an emergent one. It is the non-linear and unpredictable result of the set of internal relations that give rise to occurrences that could not be explained by the single components.

By using the term "non-linearity", also defined as "sensitive dependence on initial conditions[7]" we want to underline the lack of proportionality between causes and effects typical of complex systems, which makes these phenomena probabilistic and, therefore, irreducibly unpredictable.

Complex systems are therefore a mixture of order and disorder, displaying regularity and irregularity (as shown in Figure 1), a temporary balance resulting from the action of numerous interacting parts.

---

[5] Gould S.J.; Lewontin R.C. (1979). « The Spandrels of San Marco and the Panglossian paradigm: a critique of the adaptionist programme ». Proceedings of the Royal Society of London. Series B, Biological Sciences, pp. 581-598 For a precise definition of the concept "exaptation"

[6] Pinker S. (2003). "The blank slate. The modern denial of human nature". Penguin Books, New York. For a state of the art of this debate

[7] Glasner E.; Weiss B. (1993). "Sensitive dependence on initial conditions". Nonlinearity 6, 1067.

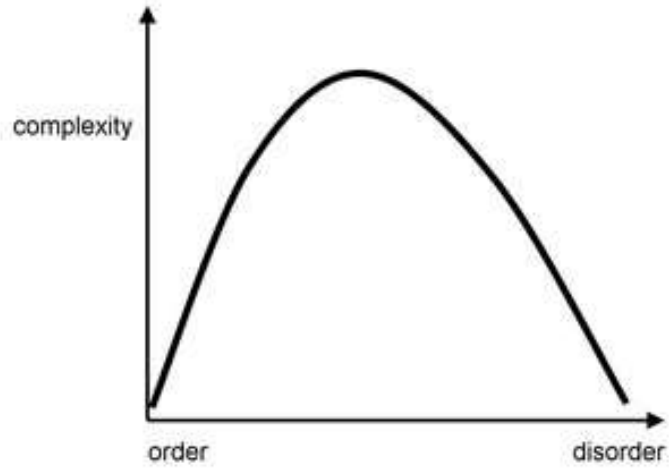



Figure 1[8] : Schematic diagram

*Figure 1* shows a schematic diagram of the shape of such a relation. However, it should be emphasized again that a generally accepted quantitative expression linking complexity and disorder does not currently exist[9].

The relationship between the structure and the functioning of these systems often takes the nature of feedback (positive or negative feedback). Similar dynamics of mutual interactions will be deepened both in the relationship between brain and mind (phenomena such as "selective stabilization" or learning show how a certain use of a system leads to a modification of the "fixed" part (wiring) of the system itself). This is in some ways analogous to the attempt with which neuroscientist Hebb tried to explain the neuronal phenomena of synaptic plasticity through the expression "neurons wire together if they fire together[10]"

It is therefore difficult to measure and to find a suitable unit of measurement that can quantify the degree of complexity of a system made up of many interacting parts, such as language could be considered of.

Some methods try to focus on the analysis of randomness, others on the analysis of regularity (or deterministic processes). In both cases the objective is to try to quantify the amount of

---

[8] Image taken from Scholarpedia http://scholarpedia.org/article/Complexity URL consulted on 22/10/2020

[9] Ibidem, see paragraph “Measures of complexity” URL consulted on 22/10/2020

[10] Hebb, D.O. (1949). “The Organization of Behavior”. New York: Wiley & Sons

information (regularities, sequences, patterns) contained in a system to make them somehow computable in algorithms that will later help to classify the various types of systems and to predict their functioning.

One way to understand complexity is represented by the attempts to model it in topological graphs carried out by network analysis. By spatially visualizing the structure and trying to represent the directions of internal flows, attention is focused on global behaviours: in this way it becomes possible to capture a good part of its functioning, consciously putting in the background the properties of the single elements to shift the focus on the overall interactive dynamics.

A recurring, though not universal, characteristic of complex systems is the scale-invariance form: many complex systems achieve this form through the so-called "phase transition" from a chaotic state to a state called "self-organized criticality" (Bak, 1996).

The study of complex systems through a theoretical framework of a statistical nature has its roots in the first attempts to understand thermodynamic systems: the very high number (*e.g.* Avogadro's number) of molecules present in a fluid makes it impossible to approach it by following the dynamics of each single component, similarly to what happens in social networks or web page networks. The great number of elements and the much greater number of connections between them would put even the most powerful computer in difficult times.

Statistics comes in handy because - through averaging, confidence intervals, analysis of variance and other related concepts - it solves the problem of processing large amounts of information.

The average of a quantity - despite the simplicity that a mean represents in itself - has resulted of interest in estimating the uncommon path of what we have called Sentence Phonetic Variation Rate (SPVR), and the subsequent *intra*-child and *inter*-children comparisons made based on this value. As already pointed out (Sauvage, 2015), first language acquisition is not an incremental process: a child can properly pronounce a given word correctly and incorrectly in the same sentence, the same child can correctly pronounce a given phoneme at 3 year old and make a mistake on it at three and a half, proving the instability of learning.

## 1.2 The hallmark of complexity: Scale-free phenomena

Scale invariance is an "irregular regularity" which, in very different phenomena, shows a self-similar structure in which one part repeats the shape of the whole in which it is contained.

This iterative process consisting of repeating the same scheme at different scales generates structures called "fractals", as the Franco-Polish mathematician B. Mandelbrot, its discoverer, called them.

Scale-free phenomena can be of two types: spatial (or topological) and temporal.

In the former, it is an abstract geometric form (*e.g*. Julia's set) or a natural form (the branches of trees, bronchioles and pulmonary alveoli or tributaries of large rivers) that repeats itself in space so that, if a magnifying glass would be placed over a part, it would be approximately equal to the structure that contains it.

In the latter, irregular regularity is expressed in the form of temporal dynamics in which a portion of time displays oscillations of variables analogous to those of the long period of which it is part. One of the first studied examples was the fluctuation of cotton prices in the U.S. stock market highlighted by Mandelbrot (Mandelbrot, 1963[11]). What was observed was that the daily, weekly, and monthly fluctuation curves were statistically similar. The fact that similar proportions between variations are traceable at different time scales leads one to think that a fractal structure is present.

A network that has a power law probability distribution of the degree - regardless of the type of structure it may assume - is called a “scale-free” network.

The degree is a fundamental property of the elements that form a network; it indicates the number of adjacent connections of a specific element with other elements of the network under examination. In scale-free networks, few elements (hubs) hold many connections and, vice versa, many elements hold few connections.

[11] This regularity was questioned with other data that would rather show that variations are more unpredictable than what was initially thought

It can be expressed by the following mathematical formula:

$$P(k) \sim k^{-\lambda} \tag{1}$$

Called "power-law"

Where $k$ is the degree (number of connections) of $P$ and $\lambda$ an exponent[12] derived from a specific theorem.

From the relationship between the formula and the shape of the distribution we can see how, as the value of a variable increases, the value of the other progressively decreases by forming the so-called "fat tail" effect: a long tail that runs along the abscissae axis indicating rare and significant occurrences of the phenomenon.

When $\lambda<2$ the average degree diverges, while when $\lambda<3$ the standard deviation of the degree converges. Most scale invariance networks have an exponent between 2 and 3.

## 1.3 An example of emergent scale-free phenomena in human language: Zipf's law

An example of a power-law probability distribution is Zipf's law, named after the philologist and linguist at Harvard University who discovered it while examining in a comparative way a huge number of literary texts: he noticed that this regular recurrence was a common denominator between many different languages[13]. Today he is considered one of the fathers of computational linguistics.

[12] Caldarelli G. (2006). "Scale free networks". Oxford University Press. P224

[13] G. K. Zipf. (1949). "Human behavior and the principle of least effort. An introduction to human ecology"", Addison Wesley press, Cambridge, Massachussets.

This systematic distribution of words shows how the frequency of a term is inversely proportional to its rank (decreasing order of appearance of words in a text).

It can be expressed by the formula:

$$f(z) = \frac{C}{z^a} \tag{2}$$

The frequency of the word of rank *z* is equivalent to the ratio between *C*, a constant that depends on the length of the corpus and its vocabulary (approximately a type/token ratio index), and *z* raised to *a*, an inverse index of the lexical richness of the *corpus*.

If we put $a=1$[14] , we obtain an ideal series that approximates the frequency distribution of words in various tested samples: the second word occurs half the times of the first, the third a third of the times of the first, the fourth a quarter and so on.

As rank increases, the difference between the frequency of the previous word and the next word will progressively decrease: by doing so, successive rank increases results in a progressively slower decrease of frequency.

The final part of this series is called "tail". It is at the bottom-right of the second quadrant of the Cartesian axes of the graph where this correlation is visually plotted. It often presents words with only one occurrence: these are semantically important words called *hapax legomenon*.

The reason at the base of this recurrent frequency distribution was enunciated by Zipf in the so called "principle of least effort" (Zipf, 1949): the speakers and the hearers - while communicating - try to maximize the result (the mutual understanding) by using a quantity of cognitive resources that is just enough to reach the purpose.

For this reason, few words are pronounced many times and, vice versa, a lot of words barely occur.

[14] As described in Lenci A. 2010. Course materials “Parole e frequenze”. Università di Pisa

According to the author:

“In simple terms, the principle of least effort means, for example, that a person in solving his immediate problems will view these against the background of his probable future problems, as estimated by himself. Moreover, he will strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems. That in turn means that the person will strive to minimize the probable average rate of his work-expenditure (over time). And in so doing he will be minimizing his effort, by our definition of effort. Least effort, therefore, is a variant of least work[15]”

Another correlation linked to the "principle of least effort" highlighted by Zipf explains how, in human semiotic systems in general, most frequent words are shorter and less frequent words are longer:

$$f_v \propto \frac{1}{l_v} \tag{3}$$

[15] Zipf G. K. (1949). “Human behavior and the principle of least effort”. Addision- Wesley press, Cambridge, Massachussets. P 1

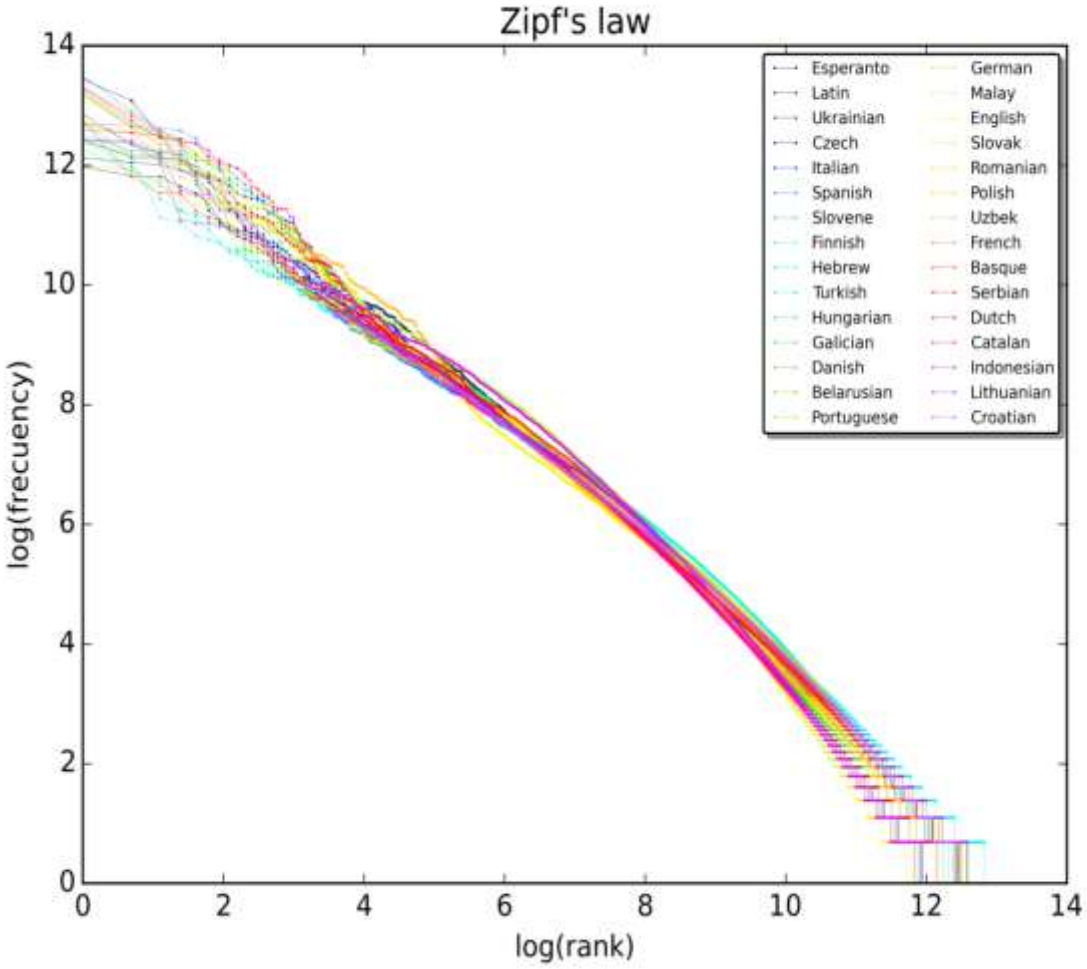



Figure 2[16] : Zip's law

In *Figure 2* a double logarithmic plot of many languages showing a substantial overlap between them, confirming in this way Zipf's law universality. Results displayed in this graph have been obtained by using Swadesh lists (Calude & Pagel, 2011)[17]

Words can be linked by semantic, positional, sintactycal or grammatical relationships. We can transform a text into a network in which each word is a node and each relation a line. In many of these graphic representations built in this way, the network of words will result to

[16] Scholarpedia, lemma "Zipf's law"

[17] Piantadosi S. (2014). "Zipf's word frequency law in natural language: A critical review and future directions". Psychon Bull Rev.; 21(5): 1112–1130. P6

have a scale-free form deriving from the characteristic power law frequency distribution previously explained.

This argument would apparently seem not correlated to the main topic of this thesis, but if we look at the graph below, we can observe how POS tags (which will be used as a standard of reference to automatically tag and parse Adrien and Madeleine *corpora* in Chapter 8,9,10) are characterised by a near-Zipfian distribution. This means that few POS tags highly occurr and many POS tags are quite rare: this may partly depends on language typology and other language specific constraints, but it is widely acknowledged that word classes differ in size: closed classes (such as adpositions, auxiliaries, determiners, numerals and pronouns) are magnitude of orders smaller than open classes (such as adjectives, adverbs, noun and proper noun, verbs).

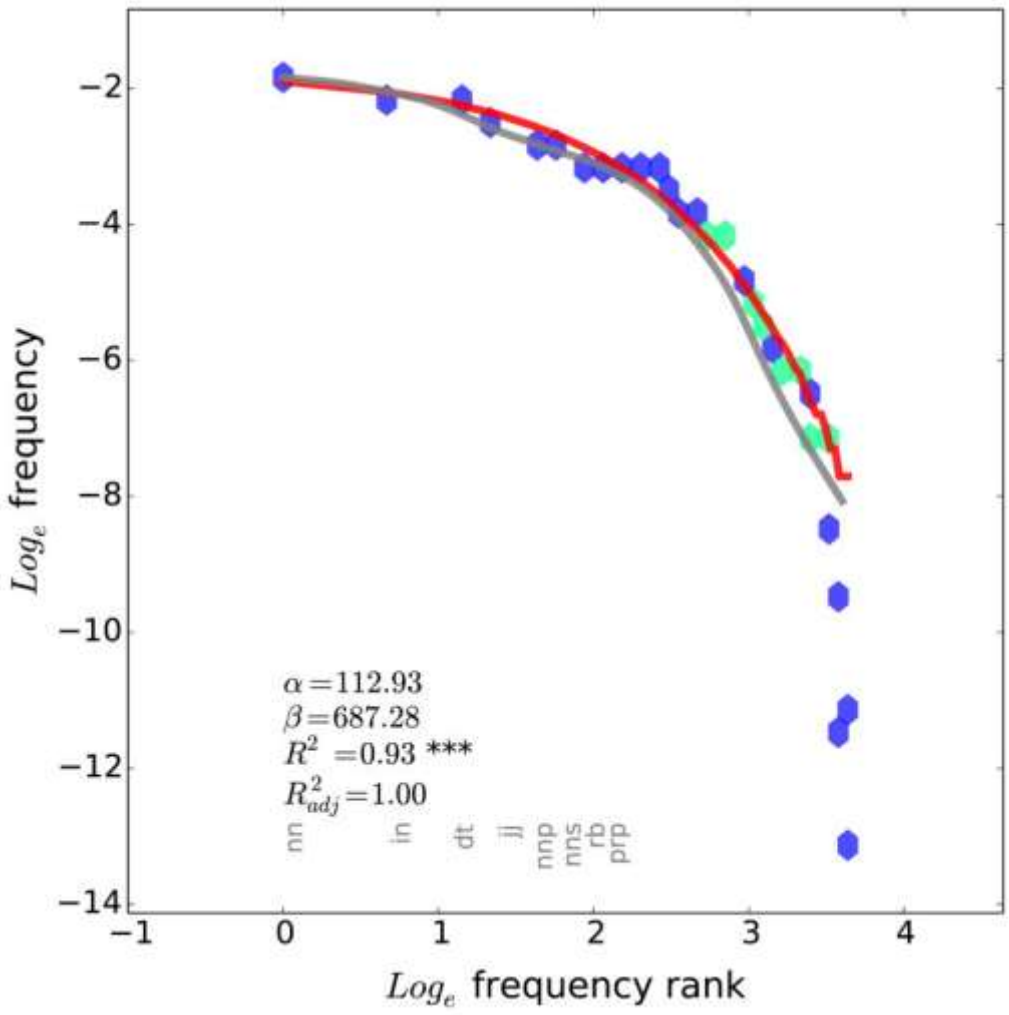



Figure 3: Frequency distribution of syntactic categories from the “Penn Tree Bank”

For a review article on the interesting role of Zipf's law in current linguistics see Piantadosi (Piantadosi, 2014)[18].

The american author argues that:

"For language in particular, any such account of the Zipf's law provides a psychological theory about what must be occurring in the minds of language users. Is there a multiplicative stochastic process at play? Communicative optimization? Preferential reuse of certain forms? In the face of such a profusion of theories, the question quickly becomes which—if any—of the proposed mechanisms provides a true psychological account of the law. This means an account that is connected to independently testable phenomena and mechanisms and fits with the psychological processes of word production and language use" (Piantadosi S., 2014)

An answer has probably been given by Dutch linguist Sander Lestrade (Lestrade, 2017), who may have found a balance between two competing factors in language structure and use:

"Words shouldn't be too general, however, as this would lead to ambiguity. In order to become frequent (within a word class), a word should be specific enough to single out its referent in context and general enough to be applied to different referents"[19].

The driving force behind the recurring probability distribution represented by Zipf could then be unveiled by a "compromise" between syntax and semantics based on a logical principle similar to the "least effort" initially proposed by Zipf himself:

"Words are from different parts-of-speech classes, which differ in size by orders of magnitude. Within classes, words differ in meaning by being differentially specified for a number of meaning dimensions. If a word is specified for a few dimensions only, it becomes ambiguous; if it is overly specific, it will hardly ever be applicable. It was shown that neither of these ingredients suffices to produce Zipf's law, but together they can. Where the results differ from the Zipfian ideal, they do so in the way natural language does. Thus, the model does not "overfit" Zipf's law but really seems to capture the underlying language mechanisms that drive it[20]".

---

[18] Piantadosi S. (2014). "Zipf's word frequency law in natural language: A critical review and future directions". Psychon Bull Rev.; 21(5): 1112–1130.

[19] Lestrade S. (2017). "Unzipping Zipf's law". PLoS One. Aug 9;12(8). p2

[20] Ibidem, p9

What I have written until now has to do with cognitive sciences?

How is it possible that humans effortlessly and (seemingly) uncosciously learn how to speak and then, once they became aware of their language in adulthood, they continue to use it without being aware of such an important underlying principle?

It seems that humans, by default, once get used to a given language, they would prefer to use it forgetting the nature of its rules and constraints, like if humans were used by language instead of using it. The difference between signifier and signified and the general arbitrariness of language are at the core of human communication but they need to be conceptualized in order to be learnt, otherwise – by default – it seems that humans would conceive language as granted, as if it was fallen from the sky, a statical object rather than a dynamical one.

For these reasons it is important to point out the existing connections between evolution and language: in this interplay probably lies the answer to much of the previous questions as well as to the following one: to what extent language acquisition and language use are intentional?

According to American philosopher D. Dennett

"minds evolved and created thinking tools that eventually enabled minds to know how minds evolved, and even to know how these tools enabled them to know what minds are" (Dennett, 2017, p1).

The point is that the only way to get knowledge of minds (and, therefore, of language) is by using minds (and language): this causes a paradox, a "strange loop" (Hofstadter, 2007) because the subject under examination corresponds to the object itself, and vice versa. To have an image representing this loop, the best way is to admire M.C Escher artworks and lithographs, for instance "Print Gallery" and "writing hands".

To solve this puzzle, a starting point could be framed in a way that would reconsider the degree of intentionality in many cognitive tasks: language – in some ways – could be learned and used by

"reason without reasoners, competence without comprehension[21]"

[21] Dennett D. (2017). "From Bacteria to Bach and back. The evolution of minds". Penguin books. New York. P4

It is uncomfortable to accept that we are are writing and reading this thesis and – at the same time - admitting that these capacities allowing us to question ourselves and read other people questioning on it are simply due to chance. In a similar way, it is hard to accept to be made of a complexity that seems to be far beyond our cognitive capacities, especially the low-effort ones.

As a Danish physicist pointed out when writing about complexity:

"Psychologically, we tend to view our particular situation as unique. It is emotionally unacceptable to view our entire existence as one possible fragile outcome among zillions of others. The idea of many parallel possible universes is hard to accept, although it has been used by several science-fiction writers. The problem with understanding our world is that we have nothing to compare it with. [..] We cannot overcome the problem of unpredictability. [..] So how can there be a general theory or science of complexity? If such a theory cannot explain any specific details, what is the theory supposed to explain? How, precisely, can one confront theory with reality? Without this crucial step, there can be no science.

[..] Fortunately, there is a number of ubiquitous general empirical observations across the individual sciences that cannot be understood within the set of references developed within the specific scientific domains. These phenomena are the occurrence of large catastrophic events, fractals, one-over-*f* noise and Zipf's law. A litmus test of a theory of complexity is its ability to explain these general observations. Why are they universal, that is, why do they pop up everywhere?[22]".

## 1.4 Some reflections on Free Energy Principle and the Bayesian brain hypothesis as overarching principles of cognition

The starting question that gives the general framework of this inquiry is:

« what it is to be alive »?

[22] Bak P. (1996) "How nature works. The science of self-organized criticality". Springer-Verlag New York. p12

What behaviour must a living system adopt to resist to dispersive forces over time?

Biological systems seem to resist to the second law of thermodynamics because they maintain their physical integrity in the face of random fluctuations in the environment (Friston, 2010).

This capacity for « negative entropy [23] », in other words acting selectively upon the environment and metabolizing food, distinguish living from non-living systems.

Free energy theory assumes that any living system possesses a random dynamical attractor, that is a set of states towards which a dynamical system tends to evolve for a wide variety of initial conditions of the system's state. Despite non linearity and « butterfly effects [24]», living systems spontaneously tend to a relatively narrow range of critical states.

This could be considered as analogous to the Saussurian dichotomy expressed in French as « *langue-parole* », to the extent that a norm (*i.e* a set of established conventions and rules) exists, so every language learner and users should comply to this norm and acquire it, and norms are of course independent from the individual user. Yet, every one of us, both adult and children, finds his/her own way to use a « parole » path through the « langue » system. «langue » and « parole » will never become identical, otherwise humans will probably be robots.

A probabilistic path turns a narrowing constraint (the « langue ») to be equivalent to a range within which an infinite amount of mutually intelligible possibilities (users' « parole », *i.e* every instance of a « langue » differing from it to some extent) can find a specific and unrepeatable form. Every path relating norms to usages is similar to every other one but, at the same time, irreductibly different. There is only one « langue » and there are as many « parole » as speakers. This is probably linked to the intrinsic power of creativity related to language.

Ending this short hypothetical analogy between language and physics, we can continue by saying that, under appropriate conditions, any system possessing a random dynamical attractor can be shown to be formally equivalent to any system at a steady state far from

---

[23] Friston, K. (2010). "The free-energy principle: a unified brain theory?". *Nat Rev Neurosci* **11,** 127–138

[24] In the original english form ""does the flap of a butterfly's wings in Brazil set off a tornado in Texas?". In Lorenz E.N. (1963) "Deterministic non periodic flow". J. Atmosph. Sci. 20, 130–141

equilibrium, where the system's « characteristic » variables are within homeostatic bounds (Friston, 2009 ; Colombo, 2018).

Borrowing a mathematical model from physics, free energy theorists claim that « the paths of the processes of adaptive (living) systems fall within a specific, relatively narrow region of all possible states in their phase space » (Colombo, 2018). It follows that from free energy's perspective « survival is equivalent to the system's being in that narrow region » (Ibidem, 2018)

A second analogy: could this observation be in principle similar to what it is defined in language acquisition as « perceptual attunement [25]» ?

This process consists in a fundamental retroaction between perception and articulation in which children progressively improve their ability to perceive every detail and shade of native constrasts while progressively losing their ability to spot non-native constrasts, as the input for non-native contrasts does not shape children cognitive system to fine-tune their related sensorimotor abilities. Adults' ability to send and receive messages in a mutual understandable way could be viewed as the final achievement of this process.

Friston gives a smart example to describe what adaptivity is meant to be: if we put a drop of ink in a glass of water, we expect that it will disperse in a few seconds. But if this drop of ink would start – afer an initial dispersion movement – to move backward and gather its molecules to the initial concentrated state, by countering in this way the dispersive force of water, we would then begin to think that this drop of ink is a living entity as it is striving to maintain itself in a stationary state (Friston, 2010).

From this example we can draw two conditions on system's adaptive behavior:

1) To behave adaptively is to preclude phase transitions and stay away from thermodynamic equilibrium
2) To keep physiological variables within certain homeostatic bounds is to change a system's relationship with its own environment

[25] Fort. M; Brusini P.; Carbajal M.; Sun Y.; Peperkamp S. (2017). "A novel form of perceptual attunement: context-dependent perception of a native contrast in 14-month-old infants". Developmental Cognitive Neuroscience 26. 45-51

This could be considered as a physical definition of what Troubetzkoy defined « crible phonologique » (Troubetzkoy, 1949): to give an example, an Italian native speaker will have difficulties to perceive all nasal features in French : for the sake of its speed-accuracy balance ( or trade-off, if you prefer) s/he would unconsciously set a threshold of tolerance that would initially prevent him/her to clearly perceive French nasals as his/her experience (allowing him/her to maintain cognitive effort at acceptable levels) does not recognise a sound that is not part of his/her native language. By doing so, the Italian native speaker will have to make considerable effort to perceive (and subsequently produce) nasals and this effort often concretise itself in L2 language courses.

### 1.4.1 What is the « free energy principle » and how does it relates with Bayesianism ?

Since its formulation, FEP has been used to explain either organism's cognitive functions such as action, perception and attention as well as organisms' evolution and development (Friston, 2010).

This was in part a response to the problem of « handling uncertainty » (Colombo, 2012, p698): as the external world is a combination of regular and irregular events, being able to infer the causes of sensory inputs in an « optimal[26] » way would be an adaptive asset for a given organism.

Learning causal regularities from the seemingly chaotic storm of events and phenomena that uninterruptedly pop up before our senses seems to be a fundamental driving force for evolutionary adaptive organisms. We could say that this is the main assumption of FEP: brains are statistical models of the worlds in which they live in (Friston, 2010), evolution seems to have selected the ones that could best represent the outside world in their nerve

[26] Here the meaning of the word "optimal" means that it must follow the rule of conditionalization (Bayes's rule), in which the iterative substitution of prior and posterior probabilties in light of new evidence or data constitutes what is defined as inference.

cells' structure (anatomy) and functioning (physiology), reflecting the main statistical tendencies and forces that shape physical world.

Friston's claim is strong, according to his framework the anatomy of every system has to contain within it a model of the environment in which that system is immersed.

As organisms live in a world that has some deep hierarchical structure in which there is action at a distance - for example in which the colors of the objects surrounding them is determined by the instant light as it comes to their eyes or- to give another example, the general effects of gravity on every object having a mass, these fundamental forces have influenced - along the constant retroactive feedback between organisms and environments - nerve cells to recapitulate these external causal structures.

Brains look what they are, networks with long connections connecting every element between each other at a distance in an approximatevely self similar fashion because - according to FEP - these are direct effects of the external world in brain circuitry.

According to the author, FEP offers a « framework within which to explain the constitutive coupling of the brain to the body and the environment » which provides « a normative, teleological essence to the synthesis of biology and information[..][27] »

Trying to anticipate what will goes on next from the basis on what has happened before « the nervous system would encode probabilistic models » (Colombo, 2012, p698)

By doing so, Bayesian models « provide us with one class of method for producing an estimate of a stimulus variable in function of noisy and ambiguous sensory information » (Colombo, 2012, p698)

Going back to the example of the drop of ink, remembering that we can clearly distinguish life in it by his way of resisting to external conditions, we can add – in the light of what I have explained so far – that this is a fundamental way to counter dispersive forces, in other words to keep a certain internal order against the external growing disorder (in a Markov blankets framework).

[27] Colombo M. ; Wright C. (2018). « First principles in the life sciences : the free-energy principle, organicism, and mechanism ». Synthese. Springer. P2

If we would give a mathematical formalization of this countering of dispersive forces, we could say that the internal states of an organism would form a self-organized process that can be showed to be a flow capable of changing a given probability distribution into another one by updating the internal representation of external reality.

This probability distribution, in mathematical terms, functions as a Bayesian model evidence: a defining dynamic of every living system that does not disappear over time in which the flow of the internal states would move as to maximize Bayesian model evidence, that implies that every living organism has – with a large range of degrees – a model of the world it inhabits. By doing so, active and sensory states tend to function in a way that maximize the existing model of the world [28].

In simpler words, this process will cut-off new information that is beyond a certain threshold of consistency with the previously stored set of information (namely, the structure of the Bayesian model evidence)

The brain in fact is an organ that seems to be actively constructing explanations for its own sampling of the world: in other words, the brain has not only to gather and explain all the sensory inputs, but it also has to choose which sensory input is consistent with its own belief and prediction of the world.

This tendency could be considered as a sort of « unconscious cherry-picking »: for merely homeostatic reasons (or, in Zipfian terms, « least effort » reasons) humans tend to confirm their certainties instead of looking to integrate new evidence from the outside world.

To conclude, any system that exists would behave as if it has a model of the world and it is trying to gather evidence for its own model of the world (Friston, 2010), this is what I believe is in place when children are temporarily looking and trying to put adult complex words (too much complex for them) in children's templates (Fikkert, 1994, p13 and Figure 10 « The output as input model » at chapter 2)

[28] This means to maximise marginal likelihood or minimize free energy. From Friston's class on FEP on British council youtube's channel https://www.youtube.com/watch?v=NIu_dJGyIQI URL consulted on 22/10/2020

### 1.4.2 FEP's general framework. Limits and advantages on adopting this broad perspective

Let's try to go into details and formalize what I have explained so far: according to professor Colombo, a prominent critics of FEP, "under any formulation, the reasoning leading to FEP has the form of a trascendental argument for the conclusion that FEP is a condition on the very possibility of existence of adaptive systems[29] ».

By following his analysis, there are six main steps to deduce FEP from external observations:

1. If a system **Σ** acts selectively on the environment to avoid phase transitions and is in a non-equilibrium steady state, then **Σ** behaves adaptively
2. **Σ** behaves adaptively only if **Σ** preserves its physical integrity by maintaining its « characteristic » variables within homeostatic bounds despite environmental fluctuations (the so-called « extended phenotype of the organism »)
3. **Σ** acts selectively on the environment to avoid phase transitions and is in a non equilibrium steady-state just in case **Σ** preserves its physical integrity by maintaining its « characteristic » variables within homeostatic bounds despite environmental fluctuations
4. **Σ** preserves its physical integrity by maintaining its « characteristic » variables within homeostatic bounds despite environmental fluctuations just in case **Σ** places an upper bound on the informational entropy (*average surprise*) of its possible sensory states
5. If **Σ** minimizes the free energy of its possible sensory states, then **Σ** places an upper bound on the informational entropy of its possible sensory states
6. Any system **Σ** that places an upper bound on the informational entropy of its possible sensory states will preserve its physical integrity by maintaining its « characteristic » variables within homeostatic bound despite environmental fluctuations
7. Any system **Σ** that minimizes the free energy of its possible sensory states will preserve its physical integrity by maintaining its « characteristic » variables within homeostatic bounds despite environmental fluctuations (Colombo &Wright, 2018)

[29] Colombo M.; Wright C. (2018). " First principles in the life sciences: the free energy principle, organicism and mechanism". Synthese. p3

All these necessaries and sufficient conditions are listed together to answer a question:

« what characteristics must biological systems possess to maintain their path within a specific (homeostatic) region that precludes phase transitions ? » (Colombo & Wright, 2018)

Before focusing on the fundamental correlation between surprise and upper bound, I would like to provide a definition of another strictly related concept: entropy.

In common sense, we use this term to refer to disorder or chaos, that is something that is far from being predictable. Entropy as decay of diversity or entropy as tendency toward uniform distribution of kinds are two useful version of this concept that could be interesting to the aim of my research (see the paragraph on « power law probability distributions »)

There are several definition of entropy depending on which domain this concept is used, it may be fair to say that the first formal definition of entropy may have been used to describe the second law of thermodynamics, which states that in any isolated system (a system that has no exchanges of matter or energy with other systems different from itself) any kind of activity (for instance, metabolization) unavoidably increases the quantity of energy that is no longer available to do any physical work, because of a lack of « order »[30].

In probability theory, the entropy of a random variable measures the uncertainty about the value that might be assumed by the variable (*e.g* 1/6 for dice).

As we previously said (example of the drop of ink) homeostasis in biological systems works as an attractor that recursively but not identically moves over and over its state space revisiting a limited set of states over time, then we could plot as a power law probability distribution all these possible state space finding that a small amount of them is highly probable compared to the rest (the so-called « fat tail » graphic plot effect)

Yet, how these models that have been conceived in physics (and sound a little bit abstract) could improve current knowledge in cognitive sciences ?

Let’s ask ourselves how sensory inputs are interpreted by our senses and processed by our cognition.

[30] “entropy” in Scholarpedia

To give a clear definition on that, I have to explain what Markov blankets are:

« given a set of random variables N, the Markov blanket for a variable x ∈ N is the subset M containing all random variables that « shield » *x* from all the other variables in N. Fixing the values of the variables in M leaves *x* conditionally independent of all other random variables; hence, the Markov blanket of a random variable is the only knowledge one may need to predict the behavior of that variable » (Colombo, 2018, p10)

This way of modeling reality is useful whenever it comes to try to explain into details what is going on in a complex set of relations between an organism and its *milieu*. To help describe this relation, we need to quantify reality to better track and evaluate what happens inside and outside these blankets.

FEP's theoretical claim is to conceptualise four basic types of quantities:

1) external states $\psi = \{ \psi_1, \ldots, \psi_n \}$ standing in for the environmental causes of sensory states;
2) active states A= $\{a_1, \ldots, a_n\}$ that change what external states the system samples;
3) sensory states (or samples D= $\{d_1, \ldots d_n\}$ that depend upon active and internal states of the system;
4) a generative model M defined in terms of its parameters or sufficient statistics, and would be « encoded » by the system's internal states (Colombo, 2018, p9)

According to FEP, an organism is made up of its defining dynamics of recurrent state spaces (*i. e* the « extended phenotype » as we initially stated) and, around itself, functioning as a bridge between internal and external states, are active and sensory states that could be modelized as Markov blankets.

What happens inside the internal states is not anymore considered independent and *apriori* from what happens in the external states: as we initially said, the constant retroactive feedback between organisms and the environment in which they live in is – in FEP assumption – what makes inner neural circuitry an anatomical/physiological structure adapted to catch statistical tendencies from the outside world and to draw inferences from them.

It follows that the first step is to determine how sensory states are generated by external states.

According to the *Bayesian brain hypothesis*, brain is akin to a Bayesian machine and « the function of this machine would be to infer the causes of sensory inputs in an « optimal » way. Since sensory inputs are often noisy and ambiguous, this requires representing and handling uncertainty [31]», so the core question is the following:

> how brain extract useful information for the organism's survival (*e.g* regularity detection or pattern recognition) from noisy and ambiguous sensory information?

« Statistical inference is the process of drawing conclusions about an unknown distribution from data generated by that distribution. Bayesian inference is a type of statistical inference where data (or new information) is used to update the probability that a hypothesis is true. To say that a system performs Bayesian inference is to say that it updates the probability that a hypothesis H is true given some data D by applying Bayes'rule[32]:

$$P\{h|d\} = \frac{P(d|h)\,P(h)}{\sum_{h\ in\ H} P(d|h)\,P(h)} \qquad (4)$$

We can read this equation in this way: the probability of the hypothesis given the data $P\{h|d\}$ is the probability of the data given the hypothesis $P(d|h)$ times the prior probability of the hypothesis $P(h)$ divided by the probability of the data $\sum_{h\ in\ H} P(d|h)\,P(h) = 1$

This theorem is known as the « rule of conditionalization » because, expressing the relationship between conditional probabilities and their inverses, it gives to the agent a way to reallocate probabilities in light of new evidence or data [33].

For this reason, Bayes' theorem is currently used to model learning mechanisms and, in more general terms, to give a formalization to many cognitive processes that change over time and that are characterized by a constant modification derived from external stimuli.

---

[31] M. Colombo ; P. Seriès ; « Bayes in the brain. On Bayesian modelling in neuroscience". British Journal for the Philosophy of science, 63 (2012), 697 – 723.

[32] Ibidem, p699

[33] M. Colombo ; « Bayesian cognitive science, predictive brains, and the nativism debate ». Synthese (2018) 195 : 4817-4838

Bayesian statistics, named after Bayes in 1763, has been a key innovation because it shifted the way to conceive the role of chance from an aleatory perspective to « epistemological » uncertainty.

Being chance represented through probability distribution gives a way to consider lack of knowledge and the possibility of gathering new information over time as two parameters that form a « dynamical » formalization for the assessment of uncertainty about unknown quantities to be able to provide more precise inductions and predictions.

Some authors[34] claim that researchers are already able to draw correspondences between behavior and brain in terms of a unique Bayesian functioning acting at different levels: it means that a change in certain perceptual tasks directly corresponds to changes in computationally-based brain mechanisms, and viceversa[35].

These authors are among those who have formulated the *Bayesian coding hypothesis*, according to which:

« the brain represents information probabilistically, by coding and computing with probability density functions or approximations to probability density functions » (Knill and Pouget, 2004, p713)

In light of this statement, how Bayesian coding and computing is considered statistically optimal?

Let's follow Colombo's reasoning:

"Call S a random variable that takes on one of a set of possible values $S_1, \ldots S_n$ of some physical property – e.g colour, lenght, or velocity. A physical property of an object is any measurable property of that object. The value of S at a certain time describes the state of that object with respect to that property at that moment in time. Call M a sequence of measurements $M_1, \ldots M_n$ of a physical property. M can be carried out through different

[34] For instance: de Petrillo et al. on animal decision making « Emotional *correlates of probabilistic decision making in tufted capuchin monkeys (Sapajus spp.)»* in *Animal behavior* 129:249-256 · July 2017.

A.Seth *« Interoceptive inference, emotion, and the embodied self* » in Trends in cognitive sciences. Vol 7, Issue 11, 2013. Pages 565-573

[35] Knill DC ; Pouget A. ; « The bayesian brain : the role of uncertainty in neural coding and computation ». Trends in neuroscience. 2004 Dec;27(12):712-9

measurement modalities. Call $M_i$ a sequence of measurements obtained through modality *i*. Measurements $M_i$ are typically corrupted by noise. Noise might cause a measurement $M_i$ to yield the wrong value for a given S. An estimator *f* $(M_i)$ is a deterministic function that maps measurements $M_i$ corrupted by noise to values of the physical property S. If we assume that $M_i$ is the measurement carried out by sensory modality $i$ – e.g vision or touch – then perception can be modeled as Bayesian inference.

Given a sequence of measurements $M_i$ , the task of a Bayesian sensory system is to compute the conditional probability density function P (S | $M_i$ ). We can then restate Bayes' rule (1) in this way:

$$\mathrm{P}\ (\mathrm{S} \mid M_i\ ) = \frac{P\ (M_i|S)\ P\ (S)}{P\ (M_i)} \qquad (5)$$

Where $P\ (M_i|S)$ specifies the likelihood of the sensory measurements $M_i$ for different values of the physical property S, P(S) is the prior probability of different values of S and P (S | $M_i$ ) is the posterior density function. Bayesian inference here is concerned with computing the set of beliefs about the state of the world given sensory input[36].

« Specifically, the surprise of sampling some sensory outcome ( or experiencing some sensory states) can be represented with the negative log probability : - log $p\ (\ D= d_{t+1} \mid a_{t,} M)$

This measure quantifies the probability that a sensory $d_{t+1}$ is sampled, given action $a_{t,}$ and the generative model M. If the sensory outcome is « incompatible » with M and $a_{t,}$, then the sensory sample $d_{t+1}$ is surprising. If there is a high probability that biological systems are found at any point in their lifetime in homeostatic states, then environmentally-generated

[36] M. Colombo ; P. Seriès ; « Bayes in the brain. On Bayesian modelling in neuroscience". British Journal for the Philosophy of science, 63 (2012), 697 – 723. P701-702

sensory samples will be unsurprising. Sensory samples generated by all other external states in the environment will be highly surprising [37].

The way that we perceive objects of the external reality is heavily influenced by the the way we can act upon them (see the concept of *affordance*[38]), for instance : something that can be seen is exclusively perceived in virtue of how it can be manipulated by a given organism ; to give an example, if I see a fruit, I « spontaneously » see in which ways this fruit can be useful to me (an affordance in the embodied cognition and enactivist's literature) and consequently in which ways I can afford it by acting upon it (*e.g* by grasping it).

Would it be possible to draw a parallel with sounds' perception and production?

According to Friston, a lot of perceptual capabilities would be based on the opportunities in terms of actions that a given organism would be able to perform on a given part of external reality: we only see through the eyes of our muscles in terms of what it means for our behaviour (Friston, 2010) in the sense that when we perceive something, the opportunities for actions that our percept forwards directly influence our perception.

These considerations go directly to the one of the key theoretical innovations of current cognitive science: the brain is not anymore conceived as an *apriori* black box that independently process external stimuli and produce certain outputs, but it is rather viewed (in a more objective way according to the supporters'claim) as an extended organ that constantly act on and react to the environment in which is it immersed, trying to establish a dynamic causal modeling (Friston, 2003) between action and perception.

To express this framework in his words:

---

[37] M. Colombo, C. Wright ; "First principles in the life sciences : the free-energy principle, organicism, and mechanism". Synthese, 2018, p 10.

[38] "The affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill. The verb to afford is found in the dictionary, the noun affordance is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment". J. J *Gibson (1979), The ecological approach to visual perception. Houghton Mifflin Harcourt. Boston. p. 127*

« the environment is acting upon you, and you are acting upon the environment » (Friston, 2010)

This circle of causality could also be reversed in this way:

« your action upon the world becomes the world's way of perceiving you and the world acts upon you through your perception of the world » (Friston, 2017)

To conclude this introductory part on the theoretical framework that I would like to adopt to study first language acquisition, it seems to me fair to underline some critical aspects of this theory. I suppose tthis theory to be one of the fittest to try to model in the more objective way (and then simulate in the most realistic way) a puzzle of human cognition such as the acquisition of phonetic patterns and their relative phonological value.

According to Colombo, "FEP epistemic status is opaque" (Colombo & Wright, 2018): reflecting on its logic structure, we could even claim that FEP is a plausible model that – striving to be as universal as possible - lacks some clarity in its application to different domains (Colombo & Wright, 2018)

It is difficult to give a unique and constant account of FEP's mechanisms because its broad formulation seems to provide a way to « personalize » it based on the biological phenomenon that is analyzed through this lens. In addition, despite there is a growing consensus by the scientific community on the fact that many cognitive processes can be modelized in a fair way by the « Bayesian brain hypothesis », it is still hard (maybe even impossible) to be able to directly observe network of neurons « encoding » and « updating » their evaluation and prediction schemes over time through the support of new data coming from the environment.

To conclude: is Bayesian modeling biologically feasible?

Can first language acquisition be modeled by this framework?

### 1.4.3 A couple of examples

Understanding language through mathematical means is an increasing domain that gather more disciplines in a collective effort to solve puzzles concerning the way by which

we, as humans, are capable of « convey meaning through sounds[39]» and it is reductible to broad background questions such as:

« is meaning related to computation? », "which is the nature of the relation between computation and representation? » (Plebe & de la Cruz, 2018)

Language intimally shapes the way we think, there is a complex link between perceptual categories and linguistic categories that is hard to frame and investigate (Stapel & Semin, 2007). Language learning is a spontaneous and effortless process and this makes difficult almost every metacognitive reflection on it: it is so embedded in our way of knowing the world (including ourselves) and acting upon it that we are not ready to have a step back from it in order to appreciate its contingent and conventional nature. For this reason, we claim that studying its underlying quantitative structure could be a path toward a better understanding of language's aspects that humans use while communicating with each other without being aware of using it.

To give an example on how computational tools can shed some light on language I have summarized the way Zipf's law has been discovered[40].

Zipf may in fact be considered the forerunner of many following researches combining statistics and linguistics. Human language follows a power-law probability distribution in words' frequency (Ferrer i Cancho, 2001): if this relation is plotted, any given written text as a romance, or any written transcription of an oral discourse will show a scale-free probability distribution of words' degree (the number of links each word has with other words).

In recent years there has been a growing number of projects and papers on modelling first language acquisition by using statistical and/or computational tools (Wintner, 2010, for a review).

One of the first and most important experiment showed how specifics "innately biased statistical learning mechanisms" are activated during *in vitro* settings in which children easily learn how to keep memory of the transitional probability between syllables to spot words'

[39] Plebe A.; De la Cruz V. (2015) « Neurosemantics. Neural processes and the construction of linguistic meaning ». Springer.

[40] Ferrer I Cancho R. ; Solé. R.V. (2001) « The small world of language ». Proc. R. Soc. London B (2001) **268** 2261- 2265

boundaries (Saffran et al., 1996). Similar findings demonstrate that is worth trying to explore databases made up of transcripted infant spoken language to verify whether and how underlying patterns and recurrent sequences of learning stages are at work during acquisition.

These researches - in a nutshell – try to tackle the following question: can we infer some trace of the structure and functioning of language by finding recurrent statistical pattern from it?

Which empirical value could have a statistical measurement for such a qualitative phenomenon language seems to be?

I would like to cite what I may define a "thought experiment" that I consider well representative of what I have said so far:

« Imagine that you are faced with the following challenge: 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[41] ».

Seen from this perspective, statistics could probably help to understand first language acquisition. For this reason, I think is important to summarise with a short description the experiments on transitional probability that led to the discovery of « statistically biased learning mechanisms » by Saffran.

She demonstrated how powerful and accurate children are in detecting the probability between pairs of syllables in their language.

According to her, the first step is as follows:

"in every language, infants must determine where one word ends and the next begin without access to obvious acoustic cues [..] Given the statistical properties of the input language, the ability to track sequential probabilities would be an extremely useful tool for infant learners" (Saffran, 2003, p110)

[41] Saffran J. (2003). " Statistical language learning: Mechanisms and Costraints". Current directions in Psychological science. Vol.12 No 4. P 110-114. P111

In the framework of the long-standing debate between Nativism and Constructivism (see Chomsky – Piaget debate in the review article by Piattelli-Palmarini, 1983), statistics provides grounded arguments to reshape the proportion between the roles of experience-independent and experience-dependent processes.

So, how this hypothesis was tested?

From a general observation "within a language, the transitional probability from one sound to the next will generally be highest when the two sounds follow one another within a word, whereas transitional probability spanning a word boundary will be relatively low" (Saffran, 2003, p111)

Here is the transitional probability formula:

$$\left(\frac{Y}{X}\right) = \frac{frequency\ of\ XY}{frequency\ of\ X} \tag{6}$$

She found that the little amount of time exposure (two minutes) has unconsciously triggered a significant co-occurrence detection mechanism in every infant, so we could say that:

"some aspects of early development may turn out to be best characterized as resulting from innately biased statistical learning mechanisms rather than innate knowledge" (Saffran, 2003, p112)

Statistical regularities and probability distributions are not the only cue children have while learning their native language: prosody plays a great role too (Dodane, 2009, 2010, 2012). Supra-segmental cues such as pauses, intonation and stress are acquired very early and are often presented in an exaggerated form in child-directed speech to focus attention on a desired target such as syntactically dependent links or question marks.

Keep track of statistical regularities is one of the keys to understand the segmentation problem: how do children identify words from continuous speech?

Yet, knowing the transitional probability is necessary but not sufficient: children are helped by stress, the use of words in "carrier sentences" such as "il y a XXX", "regarde XXX", "papa fait XXX" and pauses. "carrier sentences" and "pivot schema" rely on statistical in an indirect way: as the first part of sentences remains the same, it is supposed that for children it

would be easier to spot as different the string of syllable following the fixed one. Gestuality is also a cue to the segmentation task but it is not a topic of this thesis.

First words appear around 12 months in parallel with the articulation of bisyllabic proto-words ("papapa", "dadada" and other CVCV patterns, especially those who contain voiceless consonants because bilabials are learnt earlier than other consonants), produced with rising and rising-descending intonational (stress) contours: it can be observed that children acquire very early the sound patterns of their native language as French children begin to put stress on the last syllable of words while Italian children start to put stress on the second-last. This temporal and prosodic forms will progressively become analogous to the adult form.

In current literature, prosodic bootstrapping theories (Soderstrom M et al., 2003) would put the importance not on the linear structure of speech (as Saffran did) but rather on the hierarchy on which it would be hypothethically built on.

This thesis will not contribute to reconsider the long-lasting debate between nativists and constructivists or looking to reframe it with renewed empirical bases and an evolutionary-adaptive framework.

What interests me here is to understand as much as I can the main focal points of debate, such as the relation and the proportion between experience-dependent and experience- independent mechanisms in human learning skills (both domain-general and domain-specific). Despite the lack of prosodic cues in her account, I think that Saffran's results represent a milestone in the history of language acquisition studies.

Free energy principle, as it takes in account sensory inputs and statistical information, could be considered as a unified brain theory that would allow us to reshape the debate between these two opposite theories with new tools and insights (Colombo, 2015).

Bayes's formula would seem to be a tool able to represent the process of the neutralization of phonetic variation, in other words : the way through which a child reach the ability to pronounce words in a « correct » way ( this is what Sauvage defined as « le rapprochement à la norme[42]»), because it seems a flexible mathematical model in which prior knowledge influence both the marginal likelihood and the posterior knowledge, giving to the researcher the opportunity to try to track a possible learning path of a language's sound patterns.

[42] Sauvage, 2015. P130

It could be plausible that the task of learning a language's sound patterns (its basic constrastive phonemes and phonotactic constraints for instance), would be an analogous process to the one consisting of building up a Bayesian model evidence. In other words, learning a language (here conceived as a statistical structure of the environment) is in some ways a process that bring a child to minimize long-term prediction error (refer to the schemes in the previous chapter): my claim is that a similar procedure has been demonstrated, for instance, by Patricia Kuhl in her well known experiment on japanese-american children[43].

To conclude, exploring the hypothesis by which Bayesian predictive coding could be a way to improve current explanations on how our inner cognitive percept influence our perception of external stimuli has been explained in this introductory chapter as a review of current literature on cognitive issues. I hope to will be able to take a step forward and make an attempt in testing how this framework could possibly be adapted to and tested on CoLaJE data as well as other longitudinal corpora of child spoken language in different idioms.

## 1.5 Data mining. A short summary

"Data" has become a buzzword used in almost every scientific domain and applications using "data" to do a number of different activities are growing exponentially. "Data" are supposed to represent something that is objective rather than subjective, for this reason "data-driven" analysis are thought to be an improvement from previous researches (despite nobody specifies on what kind of empirical basis were previous scientific claims or discoveries based on).

The point is that technology improvements give us a way to store and explore a huge quantity of data that was unthinkable even ten years ago: for these reasons academics started to talk of

[43] Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S. & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Developmental Science, 9, F13-F21

a “paradigm shift” in science, going from a theoretical research to a *data-driven* or exploratory science (Hey et al. 2009; Kitchin, 2014).

In corpus linguistics these technological improvements led to many changes: nowadays there are plenty of algorithms allowing linguists to mine large quantity of structured or unstructured data looking for the desired information or the desired analysis. In the past this was done manually and was a boring and time-consuming task.

CLAN has been the first automated tool to analyse longitudinal *corpora* in a quantitative basis: commands of this software are able to give a huge range of different results regarding lexical and grammatical abilities, MOR could be considered as a forerunner of current NLP automated parsing tand tagging (in particoular, POS tagging) techniques and a query such as the one provided by CoLaJE’s website[44] is an important information-retrieval tool to look for specific occurrences in any given session.

In this thesis children sentences have been parsed by using a Python-based NLP toolkit named “stanza” [45]. This tool provides linguists a way to automatically tag hundreds of thousands of words gramatically ordered in sentences within a few seconds.

The accuracy of this tool kit grows proportionally to its use growth: the more data it processes the more efficient its processing will be, and if it is increasing in efficiency, it will be more likely used again, and so on (as the so-called “snowball effect”). This positive retroactive feedback is probably at the core of recent advances in computational linguistics.

Machine learning, artificial intelligence and data mining are at the core of a revolution in language and cognitive sciences: their computational power and their ability to model in a biologically inspired way aspects of human mind will shape the near future linguistic inquiry.

Data science can be defined as “the selection and retrieval of variables to analyse and their preparation and subsetting; analysis techniques like text classification, regression modelling,

[44] https://ct3xq.ortolang.fr/ct3xq/interro URL consulted on 5/11/2020

[45] Qi P.; Zhang Y.; Zhang Y.; Bolton J.; Manning C. D. (2020). “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages”. Association for Computational Linguistics (ACL) System Demonstrations.

clustering and outlier detection; as well as the interpretation of all preceding steps to draw insights from the data or to reformulate the research question and refine the process[46]"

While data mining could be defined as "the process of finding useful, previously unknown patterns and relationships in datasets" (Witten et al., 2016, p13), text mining refers to a "similar field of application dealing with unstructured data, that means a text representing a novel or an e-mail, but that has nothing to do with a structured database (Frey, 2019, p14).

This implies many difficulties inherent to the arbitrary and ever-changing nature of language, a "multi-layered and ambiguous phenomenon" (Frey, 2019, p13) subject to many different and non-mutually exclusive interpretations.

[46] Frey J. C. (2020). " USING DATA MINING TO REPURPOSE GERMAN LANGUAGE CORPORA. An evaluation of data-driven analysis methods for corpus linguistics". PhD thesis. Università di Bologna. P12

# Chapter 2 - Phonological theories: an overview

## 2.1 A state of the art

According to Vihman, a phonological theory needs to address a set of core questions aimed at making explicit which is the "developmental source of the linguistic system, the hierarchically structured set of categories and constraints on patterning in each domain of language which make up the native speaker's (and listener's) knowledge of language" (Vihman, 2014, p246). This process is what will structure child's perception into categories and – later in adulthood – will work as a filter to the processing of different languages (see experiments on "r" and "l" perception in Japanese-English children by Kuhl P., 1993)

Here below the list of core questions:

1) What is the role of biology, or of "preprogramming" in guiding phonological development?
2) How does the child develop phonetic categories from the speech signal? And what is the role of frequency in shaping the child's phonological learning? ( this second question will be adressed in this thesis)
3) "Is there a difference between phonetic and phonological development? Or, alternatively, how can we account for apparent *discontinuities*, or reorganisation, in the child's phonological representation of knowledge?"

   This question seems to be similar to the one proposed by Sauvage (Sauvage, 2015) « It has been demonstrated that any onset does not randomly vary in any possible other onset, further, it has been observed that the process of neutralisation of this kind of variation was based on a parallel process: the building of the

representation of a phonological system mainly driven by adults' intervention[47] ». In fact, given a target word (here conceived as a sequence of phonetic units) the particular sequence of variated phonetic/phonological forms that a given child pronounces before he completed his learning could be viewed as a sequence of temporary achieved phonological structures that influence the variations across the ages. This topic will be explored at page 91.

4) "How similar is phonetic and phonological development cross-linguistically and across individuals of the same language?"

5) "What is the role of attention and effort in early phonological learning? Is language learning as effortless for the child as it seems (and as is sometimes claimed)?"

6) "What mechanism(s) could account for both lexical learning and the construction of grammatical knowledge?" Do these learning mechanisms are consciously-directed or passively absorbed by exposure to input?

We may should consider Jakobson's structuralist viewpoint of phonology as the first attempt to account to the acquisition of phonology by the child.

Phonological theories on language acquisition are a debated topic: assumptions and frameworks on which these models rely on differ a lot, they usually draw on different empirical data (different languages, different sampling techniques etc), and – *a fortiori* – they do arrive to different – sometimes incompatibles – theories.

Here, goal is to summarize three main types of competing models: formalist, perceptionist and functionalist/emergentist.

[47] Sauvage, p103, this is the original extrait in French « « Il a pu ainsi être montré que n'importe quelle attaque ne variait pas en n'importe quelle autre, et que le processus même de neutralisation de ce type de variations faisait appel à une représentation du système phonologique résultant de l'action d'autrui » (Personal translation)

The first attempt to study longitudinal *corpora* on language acquisition by having the explicit target of outlining a set of general laws governing phonological structure and its acquisition was made by Jakobson, who claimed that phonology is essentially made up of contrasts between fundamental units and acquisition corresponds to the universally progressive order of acquisition of these contrasting units:

"[..] the unfolding of a phonological system is the progressive differentiation of oppositions affecting successively smaller sound classes, based on the principle of maximum contrast and corresponding to the implicational universals of adult phonological systems[48]".

Technological improvements such as video and audio recordings, as well as data storage facilities gave an opportunity to design experiments aimed at confirming Jakobson's hypotheses on general tendencies and costraints on language acquisition. One of his main claims was that most widely used phonemes were the first learned by children of all known languages: by combining longitudinal *corpora* and quantitative techniques it is possible to test such kind of hypotheses.

For example, according to Adda-Decker[49], in French adult language most widely used phonemes are "r", "l", "s", "t", but it is obviously clear that the minimal pair "r"-"l" is far from being the first learned consonantal opposition. While for vowels, "a", "e", "schwa", "i" are the most common and the results seem more coherent to Jakobson's prediction as reported in Vihman (Vihman, 2014).

"The first consonantal opposition is predicted to be oral vs. nasal ([ba] : [ma] or [da] : [na]), then labial vs. dental ([ba] : [da]); the first vocalic opposition, high vs. low ([i] : [a]), then high – mid – low ([i] : [e] : [a]) or high front – high back – low ([i] : [u] : [a])[50]"

The same test can be done on other languages by using frequencies and statistics freely available on the Maddieson's UPSID website (and in the "Lyon-Albuquerque Phonological Systems Database[51]" a French- American improvement of the previous project)

---

[48] Vihman, 2014,p250

[49] Adda-Decker M.; (2006). "De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux". JEP2006 - XXVI Journées d'Étude sur la Parole, 12-16 juin 2006, Dinard (Proceedings)

[50] Vihman, 2014,p250

In the final part of the thesis it will be provided a graphical and interactive way to test hypotheses related to phonemes' frequencies and ages (from 1 to 5 years old). See chapter 11 "Data mining")

In this thesis there will be any comprehensive historical reconstruction of the Nativism-Constructivism debate because there are plenty of essays and papers that have already provided a summary and a viewpoint better than I would be able to do here.

What it is important to say is that the author and his academic environment place theirselves on the constructivist area, in the sense that they do not think that there are sufficient findings confirming a language acquisition device (LAD, Chomsky, 1986) or a universal grammar (UG, Chomsky, 2007) already wired in our brain at birth in light of an evolutionarily inherited ability.

It could be fair to state that the authors belong (or think to belong) to a more constructivist and usage-based approach to study first language acquisition, in which the extraordinarily ability that infants and children show in regard to language acquisition should be rather explained by experimental evidence such as J. Saffran's experiments on how infants can rapidly find out and keep memory of the transient probability between syllables' sequences (Saffran, 1996, 2003).

Despite empirical evidence on statistical learning, a question still remain unanswered: how do children represent these set of probabilities?

How do they code that – to give an example – « ma » comes usually after « ta » and quite rarely after « tr » ?

Through which process do they become able to recognize words' boundaries so efficiently (in terms of speed and accuracy) as adults do?

These experiments showed that there is probably nothing "already available" prior to any exposure to ambient language except for "innately biased statistical mechanisms" (Saffran, 1996). More up-to-date findings regarding similar questions focused on bi-multilingual children's ability to recognize and integrate segmental, prosodic and phonological cues can

[51] http://www.lapsyd.ddl.cnrs.fr/ URL consulted on 12/7/2020

be found in the work by Mehler (Mehler et al., 2014) as well as in the "Laboratory Phonology Conference Papers" (Ramus F. et al. 2010)

In light of these discoveries, it could be said that learning is – at the same time – a powerful and constrained process.

Similar statistically-driven studies to highlight the brilliant performance of what it may be called "children pattern-recognition ability" have been done also on syntax (Gomez & Gerken, 1999) and speech categories such as consonants and vowels (Maye, Werker, & Gerken, 2002).

Despite these empirical evidences would seem to undermine Nativism, it is important to retain some important aspects of the huge contribution that Chomsky leaved to linguistics and cognitive sciences, that in the case of this thesis is especially about "generative phonology".

In the end, Nativism would still be considerable as the most elegant and simple way to account for a fundamental characteristic of language that learning-based theories seem to be not able to account for: languages of the world, despite huge surface differences, are based on deep-rooted common structural properties and they do not vary in random ways at all. For these reasons, it would be comprehensible to retain as scientifically plausible (although hard to demonstrate) the presence of innate knowledge of language already hard-wired in us.

The point is that the "constrained statistical learning framework" (Newport & Aslin, 2000) seems to be able to account for these deep similarities although it has an experience-dependent root: what would rather be common at birth is not a "module", an already established packed knowledge, but a "mechanism" able to turn some specific kind of input into a progressively structured output.

The readiness of the activation of this mechanism is debatable, as well as the intrinsic "learnable" structure of natural languages.

Where does this "learnability" come from?

"human languages have been shaped by human learning mechanisms (along with constraints on human perception, processing, and speech production), and aspects of language that enhance learnability are more likely to persist in linguistic structure than those that do not.

Thus, according to this view, similarities across languages are not due to innate knowledge, as is traditionally claimed, but rather are the result of constraints on learning[52]".

After this short sum up, it could be said that the new framework proposed by Saffran and colleaugues would represent a way to focus on an ongoing process instead of a fixed state: this is in line with the epistemological reframing of many disciplines proposed by Morin:

"*La nécessité de penser ensemble, dans leur complémentarité, dans leur concurrence et dans leur antagonisme, les notions d'ordre et de désordre nous posent très exactement le problème de penser la complexité de la réalité physique, biologique et humaine*[53]"

In fact, human language is an ever-changing phenomenon that is shaped by acoustic features (some sounds are more easily heard than others) that play a role – for instance - in terms of place and manner of articulation[54]; biological constraints such as different voice pitch and tones between males and females[55] (and between homosexual males and females[56]) and their subsequent preferences in mating choice; and – last but not least – social factors such as social status, gender, formal education, bilingualism and many others.

Morin, by underlining the importance of "connecting knowledges[57]" meant even to focus on the interaction between competing and/or contrasting disciplines or theories in the same domain: a bridge between nature and nurture in the long-lasting debate between experience-independent ( e.g Chomsky; Pinker) and experience-dependent (e.g Piaget; MacWhinney) theories on language is thus welcomed.

---

[52] Ibidem, p110-111

[53]Morin E. "La complexité humaine". 1994, p. 301

[54] Alwan A, Jiang J, Chen W. Perception of place of articulation for plosives and fricatives in noise. Speech Commun. 2011;53:195–209

[55] Suire, A., Raymond, M., Barkat-Defradas, M. (2019). "Male vocal quality and its relation to females' preferences". Evolutionary Psychology

[56] Suire, A., Tognetti, A., Durand, V., Raymond, M., Barkat-Defradas, M. (2020). "Speech acoustic features: a comparison of gay men, heterosexual men, and heterosexual women". Archives of Sexual Behaviour

[57] The original French is "relier les connaissances". Personal translation

Seeking complementarity and try to understand which are the conditions that bring a given researcher (or a given school of thought) to think a particular way about a given topic, instead of simply say that that particular way is far from yours and, thus, not worthful to be taken is an effort that – hopefully – would bring science to improve itself through a meta-analysis of its preliminary conditions.

"*Il n'est donc pas question d'opposer les cadres théoriques existants pour tenter d'expliquer comment les enfants apprenent à parler mais bien d'essayer d'articuler ces cadres théoriques, chacun étant susceptible d'alimenter la réflexion globale d'une approche plus complexe*[58]"

Making complementary what initially was thought to be irreductible to one's own theory and background led Chomsky to literally "broaden" his perspective: his vision of language as a merely formal set of rules and principles already encapsulated in our brains developing *apriori* from any external sensory input and any anatomical/physiological motor/perceptual constraints has been opened up in one of his later reformulation of his works:

" 'faculty of language - broad sense' (FLB) which includes both the FLN ("faculty of language narrow sense", which is the 'abstract linguistic computational system alone') *and* the sensorimotor (or phonetic) and conceptual-intentional (or semantic and pragmatic) systems with which it "interacts and interfaces""[59].

[58] Sauvage, 2015, p9 "what is at stake is not to oppose different existing theoretical frameworks to explain how children acquire their native language but rather to highlight what could be complementarity between these theoretical frameworks, each of them being potentially able to contribute to the overall reflection in a more global approach". Personal translation

[59] Hauser; Chomsky & Fitch (2002, pp 1570-1571) in Vihman M. Ibidem, p255

## 2.2 Prosodic Phonology

This theoretical approach has been mainly developed by Paula Fikkert in her doctoral dissertation on the acquisition of prosody by observing a sample of Dutch speaking children during naturalistic longitudinal observation (Fikkert, 1994).

Her claim consists in describing how though syllable structure and stress could be learned lexically, children rather tend to learn it through a parametric basis:

"they have to learn what the parameter values are of the language they are learning. Since adult speakers have more or less clear intuitions about what constitutes a possible syllable or word in Dutch, and, furthermore, have intuitions about where to place stress in nonsense words, this view of acquisition is also preferable, because it explains why adults have these intuitions, which they would not have displayed if syllable structure and stress were lexical properties[60]"

Drawing from Chomsky and Nativism more generally speaking, she focuses on the importance of learnability of languages (see chapter 7 of her thesis, especially the paragraphs "Metrical theory as part of UG" and the subparagraph "the form of the input") and on the suprasegmental level rather than segments (see chapter 6 on stress acquisition):

"[..] there are well-defined parametric theories for both syllable structure and stress [..] it is generally assumed that parameters are innate. This means that we have some ideas about the innate properties with which the child is born, namely, the principle and parameters of UNIVERSAL GRAMMAR (UG), which is of considerable help for the understanding of the developmental stages [..]. Moreover, it is possible to write algorithms which, when applied to machine learners, are able to choose the correct parametric values for stress and syllable structure in a given language (Dresher & Kaye 1990, Dresher 1991, 1992)[61]"

[60] Fikkert P. "On the acquisition of prosodic structure". Radboud University Nijmegen Phd thesis Repository. 1994

[61] Fikkert, 1994, p2

This means that a child learning his/her language has an hypothesis space in which s/he tests the degree of fitness between expectations based on this model and data arriving in form of input from the environment.

The point is that it is not clear the extent to which this structuring process in which a child define his/her preset parameters (UG) in order to specify and complete his/her language knowledge is consciously-directed or not:

" [..] the claim that development reflects a learning process is controversial (Atkinson, 1982). According to Chomsky (1987) knowledge of language grows in the mind of a child. It therefore involves no learning. An alternative to learning is TRIGGERING. The recognition of some cues in the data, the child's "trigger experience" (Lightfoot, 1989), is said to trigger parameter setting. Triggering is a relationship between the data and their consequences for parameter setting. This involves non conscious learning. Rather, in this view, UG must also contain the cues for which the language learner looks in the data, and which may trigger parameter setting. In this sense, hypothesis testing is looking for cues in the data which may trigger parameter setting [..][62]"

This could be considered analogous to what Sauvage defines "*le rapprochement à la norme*" (Sauvage, 2015) : children – step by step – express theirselves in a way that is more and more similar to the adult one, but it is not specified by the author whether he considered this process being consciously-directed or triggered by the external input.

This difficult question should be addressed both from the Constructivists and the Nativists because it is posited *a priori* from the experience-dependent or independent framework we choose to account for language acquisition.

Fikkert suggested a framework on which studying prosody in words from an analysis of its constituents:

- *Prosodic Hierarchy*
- Prosodic Word Wd
- Foot F
- Syllable σ
- Mora μ

[62] Fikkert, 1994, p 10

and a hypothesis to integrate acquisition and prosodic theory:

"the child's template is determined by the prosodic parameters, all of which are in the default setting at the early stages of acquisition. On the basis of evidence in the data the child sets parameters to the marked values. The child's template is thus defined in terms of the authentic prosodic units. It determines the relation between the input (target or base) and the output (the child's production form). The child maps the input forms, *i.e.*, (part of) the adult target forms, onto his or her template, in such a way that the template is maximally satisfied[63]".

The core syllable (or the default value) is considered to be CV: a selection strategy is applied by the child to the complex adult forms that need to be mapped in his/her template, this causes different phenomena such as coda reduction, deletion (according to Ito's prosodic licensing. Ito, 1986), expansion of a target form.

For instance, Théophile at 2;07;28 after having heard the noun "crocodile" in the correct form from his mother, he pronounces it in a reduced form[64]:

| | | | |
|---|---|---|---|
| CHI | 0:48:33 | 0:48:35 | un crocodile ! |
| pho | | | ɛ̃ kodi:l |

It could be explained by the fact that plosive-liquides phonemes are often reduced to CV forms at that age: a two and a half year old child can probably perceive "crocodile" as it is but – in the articulation phase – he had to map the adult form into his template "under construction", in which a CCVCVCVC turns into a more easy-to-articulate CVCVC form.

One year later, at 3;05;11, the same child (who probably loves reptiles) has improved his articulatory abilities[65]:

---

[63] Fikkert P., 1994, p18

[64] http://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data/colaje/theophile/THEOPHILE-26-2_07_28/THEOPHILE-26-2_07_28.tei_corpo.xml&m=/data/colaje/theophile/THEOPHILE-26-2_07_28/THEOPHILE-26-2_07_28-480p.mp4&time=2913.733 URL consulted 22/10/2020

CHI 0:06:27 0:06:29 +❮ yyy c'est le mot de crocodile .

pho kʁɔ se lə mɔ də kʁɔkogil

Now a CCV sequence is mastered, but it is not clear whether this probable effort could have influenced the prononunciation of “d”: this consonant was correctly pronounced at 2;07;28 but now turned into a “g”. This is probably due to the fact that “d” and “g”, in French language share the following features:

“sonant”, “approximant”, “continuous” (all unmarked traits) and “voiced” (marked trait)

and they do differ only regarding the place of articulation (Coronal), as we can observe in « Clements & Hume, 1995 » (see Annex 8). “d” and “g” are both plosives but they do differ in the fact that the first is an alveolar and the second a velar (see Annex 7).

Another similar example is provided by Anaé at 2;01;05:

CHI 0:22:32 0:22:36 le crocodile .

pho lə kugil

The mother points to a figure of a crocodile and pronounces this name correctly and the child shows these variations: maybe because of her younger age compared to the two previous examples, she combines the two previous variations made by Théophile.

It is important to point out that individual differences are of great importance[66]:

---

[65] http://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data/colaje/theophile/THEOPHILE-34-3_05_11/THEOPHILE-34-3_05_11.tei_corpo.xml&m=/data/colaje/theophile/THEOPHILE-34-3_05_11/THEOPHILE-34-3_05_11-480p.mp4&time=391.265 URL consulted 22/10/2020

[66] http://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data/colaje/madeleine/MADELEINE-25-3_03_02/MADELEINE-25-3_03_02.tei_corpo.xml&m=/data/colaje/madeleine/MADELEINE-25-3_03_02/MADELEINE-25-3_03_02-480p.mp4&time=2071.466 URL consulted le 20/10/2020

| | | | |
|---|---|---|---|
| CHI | 0:34:31 | 0:34:34 | j'ai eu le crocodile ! |
| pho | | ʒɛ y l kʁɔkɔdil | |
| mod | | jɛ y lə kʁokodil | |

In this case Madeleine at 3;03;02 is younger than Théophile but she is able to properly pronounce CCV effortlessly.

The same holds for Anae at 3;05;22[67]

| | | | |
|---|---|---|---|
| CHI | 0:59:03 | 0:59:06 | ah non t(u) as eu un crocodile [=! sourit] ! |
| pho | | | a nɑ̃ t a y ɛ̃ kʁokodi:l |

These examples – though very shortly explained – seem to confirm the tendency already demonstrated by Morgenstern & Parisse in their graphs (Morgenstern & Parisse, 2012), where they show how girls on average develop faster than boys.

[67] http://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data/colaje/anae/ANAE-21-3_05_22/ANAE-21-3_05_22.tei_corpo.xml&m=/data/colaje/anae/ANAE-21-3_05_22/ANAE-21-3_05_22-480p.mp4&time=3543.908 URL consulted le 20/10/2020

# Chapter 3 - The corpus

## 3.1 Introduction

The open access database CoLaJE/Ortolang[68] has been collected thanks to a French national research fund « ANR » during 2009-2012 and it is described as follows:

« *L'objectif du Projet ANR CoLaJE est de reconstituer l'émergence et le développement de la communication langagière chez le jeune enfant, avec une approche pluridisciplinaire et multimodale. L'analyse simultanée de la phonologie, la prosodie, la morpho-syntaxe, le dialogue et le mimo-gestuel nous offre une perspective enrichie du développement linguistique de l'enfant. Notre travail s'appuie sur une base de données commune, comportant pour la première fois des suivis longitudinaux de productions spontanées de 7 enfants, de la naissance jusqu'à l'âge de 7 ans. Les données peuvent être directement visualisées sur le site Thématique 3 d'Ortolang*[69] »

CoLaJE is part of a broader international set of data sets called « Child Language Data Exchange System ». It was created in 1984 by Brian McWhinney, professor of Psychology at Carnegie Mellon University. This project is currently the largest and most known multilingual repository of first language acquisition data[70].

CHILDES is in turn part of a wider and multidisciplinary project named « Talkbank » that is inpired by similar principles and organized according to the same ground rules for data-usage and data-sharing:

---

[68] Here is the link to new website http://vheborto-ct3.inist.fr/ct3/toppage/ and to the still functioning old one https://www.ortolang.fr/market/corpora/colaje URL consulted on 22 october 2020

[69] https://www.ortolang.fr/market/corpora/colaje URL consulted on 22 octobre 2020

[70] https://childes.talkbank.org/

“TalkBank is a project organized by Brian MacWhinney at Carnegie Mellon University with the support and cooperation of hundreds of contributors and dozens of collaborators. The goal of TalkBank is to foster fundamental research in the study of human communication with an emphasis on spoken communication. Currently, TalkBank provides repositories in 14 research areas, as represented by the links on this page. Data in TalkBank have been contributed by hundreds of researchers working in over 34 languages internationally who are committed to principles of open data-sharing. These data are used by thousands of researchers resulting in many thousands of published articles. Data in TalkBank use a consistent XML-compatible representation called CHAT which facilitates automatic analysis and searching, using open-source and free programs we have developed[71]».

Talkbank comprises datasets on a wide variety of topics different from child language learning such as « AphasiaBank », « BilingualBank », « ClassBank », « DementiaBank » etc.

Nowadays CHILDES counts more than 4500 members around the globe, more than 130 *corpora* and over three thousand published articles[72].

Other examples of French *corpora* close to CoLaJE in terms of data collection, transcription and coding are for instance the « corpus de Lyon » and « GoadRose » available on the same branch of CHILDES https://childes.talkbank.org/access/French/

According to his creator, pillars of CHILDES are:

- Data sharing and informed consent
- Multimediality
- Open Access, Web Access and the the possibility of having a space for Commentary
- Interoperability (see TEI Text Encding Initiative format)
- Community integration

Concerning the informed consent process, before submitting their data a research group or institution must make every set of video recording and its related transcripts comply to a set of norms named « IRB approval [73]», mainly consisting of a permission for data sharing, a

[71] https://talkbank.org/ 24/06/2020

[72] Data taken from a .ppt document freely available on CHILDES website

[73] https://talkbank.org/share/irb/ CHILDES website. URL consulted on 24 june 2020

varying level in participants' anonymity and more generally the level of access to the data that these latter agreed to grant.

As linguistic interactions have been studied and can be studied in a huge variety of ways depending on the aim of the researches in question, CHILDES provides a set of tools to tackle typical issues of conversational interactions in a customizable manner: coding, quantitative analysis with CLAN, systems for audio and video linking such as PRAAT and ELAN.

CoLaJE (Morgenstern A.; Parisse C.; 2012) is a database composed of seven children that have been video recorded *in vivo* approximately one hour every month from their first year of life until they were five. Data is transcripted in three forms:

- **CHI** is what the child says in the orthographic form,
- **pho** what the child really says
- **mod** what he should have said according to the adult norm.

## 3.2 General reflection on interpretation

« Tout corpus est une construction, au sens où il est toujours le produit des analyses du chercheur [74]»

Interpreting child language is a hard task: there is an irreductible "distance" between adult's cognition and child's cognition, perceiving the world in a different way implies explaining it a in a consequently different way.

To put it in the words of the people who have mostly contributed to CoLaJE:

"*Ce processus d'interprétation est largement dépendant des locuteurs et des circonstances, et son résultat est éminemment variable. On voit ceci de manière exacerbée dans les échanges*

[74] « Every corpus is a man-made production, in the sense that it derived always from the researcher's analyses» Ochs E., 1979

*avec de jeunes enfants car ceux-ci sont souvent difficiles à comprendre même pour les adultes qui les entourent. La « distance » entre les productions vocales et le résultat de l'interprétation peut être très grande en ce sens que le résultat peut être très largement « fabriqué » par l'interlocuteur adulte de l'enfant*[75]"

It happens to be difficult to choose between two possible interpretations deriving from a two different – yet logically grounded – disambiguation paths: linking the possible meaning with the context is far to be trivial

Here some examples:

| | | | | |
|---|---|---|---|---|
| 4316 ADRIEN- 25 | ADRIEN3_04_14 | hippopotame | ipopotam | eopo |

It is intuitive to interpret "eopo" as the target word "hippopotame" because the context directly suggests that this two-syllables sequence "eopo" was meant to be the target word: in the previous turn of speech the mother said exactly "hippopothame" and was indirectly seeking to check whether the child was able to articulate this fairly difficult noun at three year old:

| | | | | |
|---|---|---|---|---|
| 4348 ADRIEN- 26 | ADRIEN3_05_14 | vais demander à papa | vɛ dəmɑ̃de a papa | be mone a papa |
| 4418 ADRIEN- 26 | ADRIEN3_05_14 | veux du jus grenadine | vø dy ʒy gʁənadin | vø dy zy toladin |

Example for the OL (plosive-liquides) variations at three years old

| | | | | |
|---|---|---|---|---|
| 4457 ADRIEN- 26 | ADRIEN3_05_14 | elle est garée la voiture | ɛl ɛ gaʁe la vwatyʁ | el e gae la fɔty |

"r" avoidance and substitution at three year old

[75] Morgenstern A. ; Parisse C. (2007). « Codage et intérpretation du langage spontané d'enfants de 1 à 3 ans ». Corpus 6. Intérpretation, contextes, codages, pp 55-78

The uvular rhotic /{/ appears as the most difficult consonant for most children. When it does not undergo deletion altogether, this consonant can be produced in several different forms, as a stop, a fricative, or substitutions, however, appear to be systematic and driven by the child's phonological system (Rose, 2000)

As reported in CoLaJE website[76], it takes 45 work hours in order to transcribe every hour of videorecording : as there are around 30 records per infant, this means that it took approximately nine thousand hours to transcribe all the seven *corpora* !

A final remark before ending these considerations on data is a personal reflection: I am studying and elaborating transcripted data for which I have not been involved at all during the collecting phase and this could be considered a possible source of bias because – ideally – it would be better to work on data you collect on your own.

Another further bias could be that French is not my mother tongue, so my interpretation would more likely to be influenced by my mother language. This second bias should be lowered by the fact that my French is good enough to attend university classes in this language and another mitigating point could be that many phonological features of French are shared with Italian because of their common Romance root.

Collecting data is of primary importance for any scientific advance and I regret to not having contributed at all to this collective effort. Here below a consideration from researchers who collected CoLaJE:

"*Il convient toutefois de ne pas surestimer la valeur d'une base de données comme CHILDES. Avec un seul clic de souris, il est tentant de l'utiliser comme unique source et de se dispenser de constituer sa propre collecte de données: on pourrait finir par ne travailler que sur des transcriptions écrites sans jamais entendre le discours des sujets vivants. Or, se confronter directement au langage d'un enfant grâce à l'observation ou à l'expérimentation est absolument primordial pour tous ceux qui étudient l'acquisition du langage. Travailler uniquement sur des enfants virtuels peut amener à oublier le caractère très interprétatif des phénomènes langagiers et les limites de tout type de transcription. Il est important à la fois*

[76] http://colaje.scicog.fr/index.php/recueil-et-transcriptions URL consulted on 3/06/2020

*d'utiliser mais aussi de participer, de contribuer à l'échange des données au niveau international, et de rester constamment en contact étroit avec les enfants*[77]".

I agree with what it is stated by the authors and I could only say that I hope to have the opportunity to give my contribution to the CHILDES project by starting a new longitudinal corpus in the near future.

## 3.3 Transcription Norms in CHAT

Aim of the program was to « brought together specialists from various fields of language acquisition to study language development in the same longitudinal corpus from a multimodal and interdisciplinary perspective. The analyses aimed to find regularities in acquisition for each child and across the children[78] »

To do so, researchers have created a uniform standard, let's call it a convention, in order to have a common empirical ground to look at raw data. In fact, collecting *in vivo* data from spontaneous speech by adhering to the same set of rules allow in a second step researchers to compare in a coherent and rigorous way children development over time.

As we can see, data are available in two different formats: video recordings have been translated in **CHAT** (an acronym for Code for the Human Analysis of Transcripts). This code is a widely used standardized format for producing conversational transcripts. It has no language specific requirements and could be used together with CLAN (Child Language ANalyses) to improve the quality and accuracy of transcription, as well as for phonological and morphological tasks.

The choice of the transcription mainly depends on the scope of the research:

- for child language acquisition, CHAT is probably the most used convention

[77] Morgenstern A.; Parisse C. (2007). "Codage et interprétation du langage spontanée d'enfants de 1 à 3 ans". Corpus 6. Bases ; corpus et langage - UMR 6039. P 59

[78] Morgenstern A. ; Parisse C. (2012), « *The Paris Corpus* ». French language studies *22*. 7-12. Cambridge Universitypress. Special Issue .P7

- for domain-general studies, a phonetic and/or phonological transcription based on the ISO-10646 standard UniCode norm (that includes IPA characters too) could be considered a reference
- for huge adult conversation interaction corpus, especially concerning monolinguals, there is not a real need to transcribe them in IPA characters, thus a simple standard orthographic transcription in the given language is provided, providing a way to have final results in a shorter time period

CoLaJE has been transcribed by using CLAN (acronym for Compurized Language Analysis):

1) First, researchers went to children houses (or eventually in public gardens or on the street) to film it through a videocamera. Before starting, they agreed a consent form with both parents.

2) Then they listen to these data and transcribe them in standard French orthographic (for adults) and in IPA characters (for children) (see annex 1). Depending on the hypothesis to test, other information can be coded, especially non verbal cues (such as in Madeleine, where non verbal cues as « smiling » and « crying » are coded by using brackets [] ) . Pragmatic context is important because it contributes to the interpretation of meaning (Grice P., 1957): for this reason, transcribers choose – depending on the context – to provide a short summary of what is going on, that could be hidden by a particular camera angle or because it could refer to something that child's parents and the recorder could have discussed before turning on the video. Prosody is not coded, there are only questions and exclamations that could implicitly indicates that the intonation is rising in questions[79] and falling in answers. To ensure a flawless trancription, this first version undergoes a kind of peer-review done by another fellow linguist by using « CHECK », a specific software provided by CLAN

3) Finally, researchers write down metadata in order to have an overview on what they have transcribed: they often point out details about the transcription, some exception to a rule they were obliged to made to account for a particoular phenomena not listed in the CHAT handbook. A short summary of salient developments occurred compare to those of the previous record are often provided too, as well as remarks on eventually detected changes in child's development of non-verbal abilities which are not directly transcribable.

[79] Dodane C. & Martel K. (2012) have provided transcriptions where prosody is integrated, but they do not appear on CHILDES.

More generally speaking, transcriptions in IPA are provided when there is the need to have the original oral form, which is not always the goal of researchers. It is also possible to transcribe the parents' speech into IPA, but since this takes a long time, the spelling is supposed to be sufficient (especially since with CHAT elisions are noted). The IPA notation depends on the discipline of the researcher (whether he is a phonetician or not, in conversational analysis, for instance, IPA is usually not used) and on the extent to which speech is comprehensible or not (this depends on speakers' voices, recording quality and so on).

By following this procedure, researchers complied to the inter-transcriber reliability and agreement (Vihman et al., 1985), for this reason data from different children transcribed by different researchers can be compared in a rigorous base, reducing as much as it is possible any kind of bias derived from subjective interpretation.

There are three types of information coded:

- General information on the whole recording session. They are mainly headers beginning with a @ symbol
- Transcription for each sentence, they begin with a * symbol
- More detailed information regarding a sentence related to a specific context, coded with a % symbol, such as for "%xgestes" (for gestures) "%sit" (for situation) and %xpnt (for pointing gesture)

According to the CoLaJE project members, to obtain a complete and readable CHAT transcript there are six steps to be followed:

1) Define the headers
2) Divide spoken language in sentences and turn of speech
3) Transcribe principal lines
4) Add dependent lines (such as % contact information lines)
5) Check the transcription
6) Fill what done so far with all the other describing elements needed

Here is an example of a transcription coded with CHAT


Figure 4 Transcription coded with CHAT

Here is an example copy and pasted from an extract of a file (Adrien-31-4_00_15.cha):

@Begin

@Languages: fra

@Participants: CHI Adrien Target_Child, FAT Father Father, MOT Mother Mother, OBS Observer Observer

@ID: fra|Yamaguchi|CHI|4;0.15|male|||Target_Child|||

@ID: fra|Yamaguchi|FAT|31;7.11|male|||Father|||

@ID: fra|Yamaguchi|MOT|27;1.8|female|||Mother|||

@ID: fra|Yamaguchi|OBS|28;1.29|female|||Observer|||

@Media: ADRIEN-31-4_00_15-480p, video

@Date: 13-JAN-2009

@Time Duration: 3521.44

@@Birth] of C: of CHI: 28-DEC-2004

@Comment: [@Font] Arial Unicode MS:18:0

@Comment: Chat

*CHI: www .

%com: Adrien- 13 janvier 2009

*MOT: ça sert à ça les xxx ?

*MOT: tiens regarde on fait comme ça .

*MOT: le doudou , tu le portes là , Allez va jouer , zou !

%xgestes: pose le doudou sur l' épaule d' A et ferme la porte de la cuisine

*MOT: xxx .

*OBS: ça va Adrien ?

*FAT: Adrien !

*FAT: il est où le programme ?

*CHI: yyy .

%pho: ada

%mod:*

Video is aligned to the transcriptions, as showed here

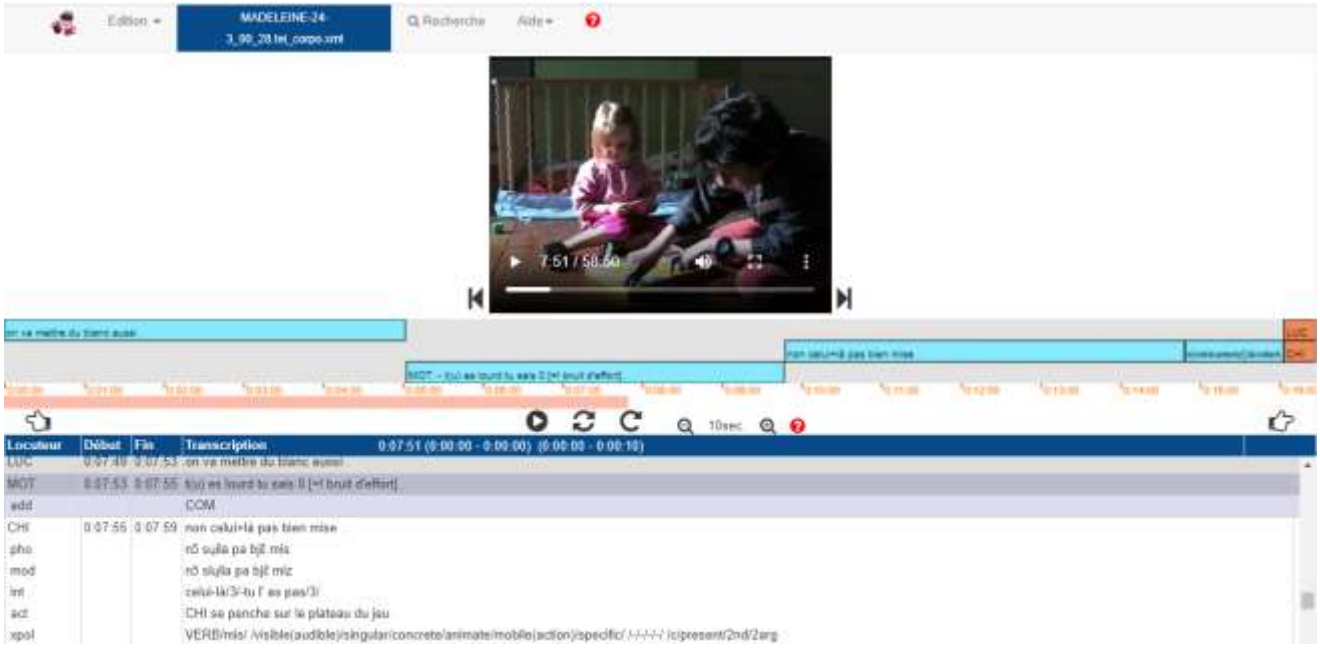



Figure 5 Example of a video[80]

As we can see, it is possible to watch simultaneously the video and what the participants are saying in an orthographic transcription by rolling the clear blue strip. The TIERS " %sit " helps pragmatic inferences.

This is in line with the best pratices in research on spontaneous longitudinal *corpora*:

" [.. ] visual context is vital for understanding language development. For example, to study how a child learns the meaning of "thank you", investigators need to know the non-linguistic contexts in which the phrase was heard and used to understand how the child generalizes from particular instances to new contexts" (Roy D., 2006)

From a technical point of view, it is necessary to give oneself the means to describe sounds, gestures, contexts and every possible situation in a sufficiently precise way to be able to share the data and analyses with people who are not involved in the original data collection.

To reproduce the data collection situation as much as possible, it is necessary to include the recorded videos in the corpus. These video must always be complemented by textual

80 http://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data4/colaje/madeleine/MADELEINE-24-3_00_28/MADELEINE-24-3_00_28.tei_corpo.xml

descriptions that allow for a better specification of the original image or to present the context of the collection.

The transcriptions can also be produced and/or converted into other formats by means of softwares allowing more precise analyses of certain linguistic parameters such as PRAAT (for phonetic and prosodic analysis), PHON (for phonological analysis), ELAN (for mimo-gestual analysis), etc.

Here is a detail of the transcription as showed in the website

| | | | |
|---|---|---|---|
| FAT | 0:00:28 | 0:00:29 | pourquoi tu veux pas l'enlever ? |
| CHI | 0:00:29 | 0:00:31 | ‹non je pas enlever› ! |
| pho | | | ‹nɔ̃ j e pa ɑ̃ve› |
| mod | | | ‹nɔ̃ ʒə pa ɑ̃lve› |
| com | | | N' a pas le micro sur lui |
| FAT | 0:00:31 | 0:00:32 | non tu veux pas l'enlever ? |
| CHI | 0:00:32 | 0:00:33 | non ! |
| pho | | | nɔ̃ |
| mod | | | nɔ̃ |
| FAT | 0:00:33 | 0:00:34 | pourquoi ? |
| CHI | 0:00:34 | 0:00:37 | ‹je veux pas enlever à maison› ! |
| pho | | | ‹de de pa ɑ̃ve a byzo› |
| mod | | | ‹ʒə vø pa ɑ̃lve a mɛzɔ̃› |
| com | | | N' a pas le micro sur lui |

Figure 6 An extract[81]

« xxx » stands when a child utters something that it is impossible to transcribe in any form, it could be considered as noise.

While « yyy » stands when a child utters something that does not make sense in a given language, but it is nevertheless transcribable, such as « bibibipapa ! ». Who knows what an infant would have wanted to express by saying such a word?

[81] The extract is taken from this session http://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data/colaje/adrien/ADRIEN-19-2_10_14/ADRIEN-19-2_10_14.tei_corpo.xml&m=/data/colaje/adrien/ADRIEN-19-2_10_14/ADRIEN-19-2_10_14-480p.mp4&time=3024.117

It may be interpreted as an invention or as an exercise on bilabials (what we may call a "diversified" or "non reduplicated" babbling).

These two forms ( "xxx" and "yyy") are quite common during the first three years of life, and they tend to decrease over time. The limit of the interpretation of the interpreter/transcriber/reader are well represented: it is hard to evaluate the intentionality behind certain utterances, as well to what extent adult's subjectivity is influencing the interpretation.

Others useful and recurrent coding symbols used in CHAT are for compounded words, elisions, fillers, repetitions and partial repetitions of words not being finished, para-linguistic occurrences, overlap (when two or more participants are speaking simultaneously).

In this thesis we have focused in particoular on two longitudinal data: « Adrien » and « Madeleine », because they are the most complete in terms of transcriptions.

The first corpus has been collected in a natural-spontaneous setting one hour every month since the child was fifteen months-old until she was almost five year old (4;11,20). The second corpus has been collected one hour per month since the child was 10 months-old until he was four-years old, then one hour every three months after his fourth-year-old birthday.

We have extracted all the Adrien and Madeleine sentences (CHI-pho-mod tiers) in .csv format and – on this basis – we created a spreadsheet like this one:

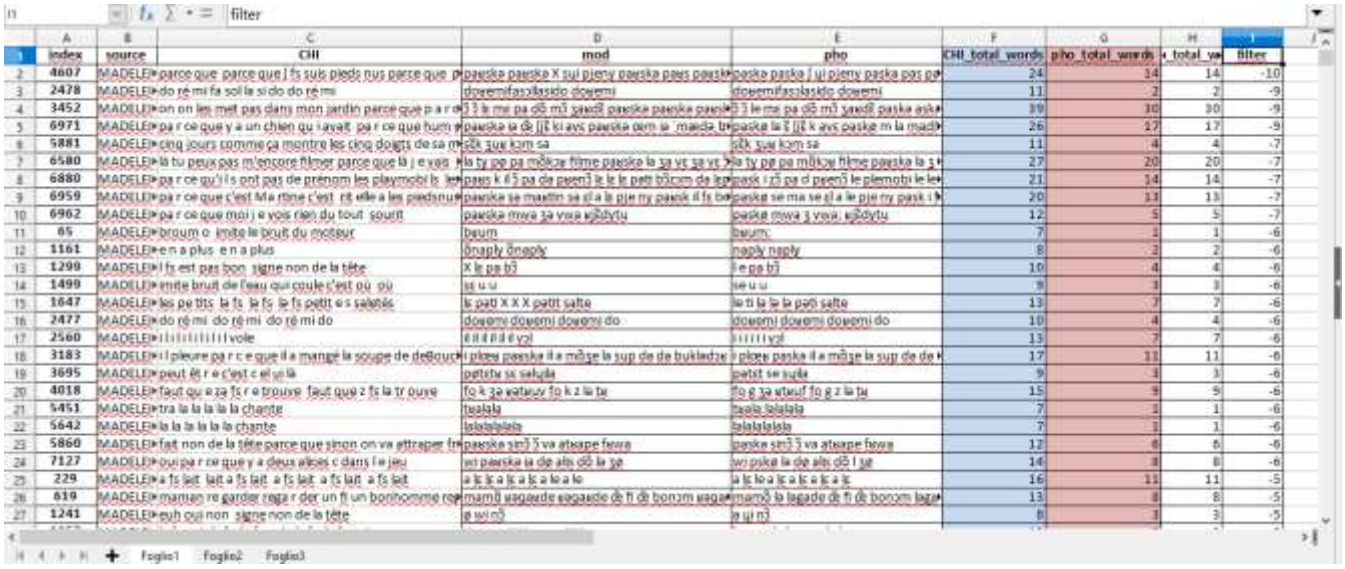


Figure 7 Example of the data set

*(See supplementary file for a better image quality).

We counted *mod* and *pho* total words and we realised that many times these two counts do not correspond one each other: this is normal because mod represents what the child should have said according to the adult norm.

This has been a huge difficult: to get the value (absolute or relative) of variation the algorithm needs an equal number of corresponding words ordered in exactly the same way in the two tiers, otherwise the output value will not be exact.

To do so, we began by cancel words in [] parentheses in which nonverbal cues such as [sourit]  [pointe] were coded.

A number of choices have been made to make pho and mod lines "fit" with each other, thus providing a way to the algorithm to calculate the number of variations.

For instance, Madeleine at 2_07_07 row 4607 (see supplementary file named "CHI&mod.allineamento.xls")

CHI parce que  parce que ʃ fs suis pieds nus parce que  parce parce que  parce que z fs ai enl e vé mon collant

Mod paʁskə paʁskə X sɥi pjeny paʁskə paʁs paʁskə paʁskə X jɛ ɑ̃lve mɔ̃ kolɑ̃

Pho paskə paskə ʃ ɥi pjeny paskə pas paskə paskə z ɛ ɑ̃lve mɔ̃ kolɑ̃

| CHI_total_words | pho_total_words | d_total_wo | filter |
|---|---|---|---|
| 24 | 14 | 14 | -10 |

Figure 8

First, for a reason still unkown, a lot of words are in a bizarre splitted form once exported in .xls, here is the verb "enlevé" which appears as "enl e vé" and of course the system recognises it as three different words as they are separated by a space.

If we look at the row and count 24 minus 14 gives 10, this is how the filter works. To avoid this problem we create a routine in Python able to recognise these differences and bridging

the gap (see supplementary file for the code). Despite so, some sentences were too difficult to keep equal and for many different reasons we decided to take off these sentences as they were less than the 5% of the total amount.

```
for index, row in df.iterrows():
    mod_list = row['mod'].split()
    pho_list = row['pho'].split()
    if len(mod_list) == len(pho_list):
        num_cols = len(mod_list)
        for num_col in range(0, num_cols):
            df.loc[index, f'w_mod_{num_col+1}'] = mod_list[num_col]
            df.loc[index, f'w_pho_{num_col+1}'] = pho_list[num_col]
            if mod_list[num_col] == pho_list[num_col]:
                df.loc[index, f'phonetic_variation_{num_col+1}'] = 0
            else:
                df.loc[index, f'phonetic_variation_{num_col+1}'] = 1
    else:
        Qui entra nei casi in cui la cardinalità di pho e mod è diversa e applica una euristica per l'incolonnamento
        comparing = compare_mod_pho(mod_list, pho_list)
        for idx_mod, mod_mapper in comparing['mod'].items():
            if mod_mapper['matched']:
                df.loc[index, f'w_mod_{idx_mod+1}'] = mod_mapper['word']
                df.loc[index, f'w_pho_{idx_mod+1}'] = mod_mapper['word']
                df.loc[index, f'phonetic_variation_{idx_mod+1}'] = 0
            else:
                df.loc[index, f'w_mod_{idx_mod+1}'] = mod_mapper['word']
                df.loc[index, f'phonetic_variation_{idx_mod+1}'] = 1

        mod_idx_not_matched = [idx_mod_false for idx_mod_false, mapping in comparing['mod'].items() if mapping['matched'] == False]
        pho_idx_not_matched = [idx_pho_false for idx_pho_false, mapping in comparing['pho'].items() if mapping['matched'] == False]

        if len(mod_idx_not_matched) == 0 and len(pho_idx_not_matched) == 0:
            pass
        elif len(mod_idx_not_matched) > 0 and len(pho_idx_not_matched) == 0:
            for idx_not_matched in mod_idx_not_matched:
                df.loc[index, f'w_pho_{idx_not_matched+1}'] = "SILENCE"
        elif len(mod_idx_not_matched) == len(pho_idx_not_matched):
            for i in range(0, len(mod_idx_not_matched)):
                df.loc[index, f'w_pho_{mod_idx_not_matched[i]+1}'] = comparing['pho'][pho_idx_not_matched[i]]['word']
        elif len(mod_idx_not_matched) > len(pho_idx_not_matched):
            for i in range(0, len(pho_idx_not_matched)):
                df.loc[index, f'w_pho_{mod_idx_not_matched[i]+1}'] = comparing['pho'][pho_idx_not_matched[i]]['word']
            for i in range(len(pho_idx_not_matched), len(mod_idx_not_matched)):
                df.loc[index, f'w_pho_{mod_idx_not_matched[i]+1}'] = "SILENCE"
        elif len(mod_idx_not_matched) < len(pho_idx_not_matched):
            for i in range(0, len(mod_idx_not_matched)):
                df.loc[index, f'w_pho_{mod_idx_not_matched[i]+1}'] = comparing['pho'][pho_idx_not_matched[i]]['word']
```

```
def compare_mod_pho(mod_list, pho_list):
    matched = {}
    matched['mod'] = {}
    matched['pho'] = {}

    for idx, mod in enumerate(mod_list):
        count = 1
        if idx not in matched['mod']:
            matched['mod'][idx] = {}
        matched['mod'][idx]['word'] = mod
        matched['mod'][idx]['matched'] = False

    pho_matched = {}
    for idx, pho in enumerate(pho_list):
        count = 1
        if idx not in matched['pho']:
            matched['pho'][idx] = {}
        matched['pho'][idx]['word'] = pho
        matched['pho'][idx]['matched'] = False

    for idx_pho, pho in enumerate(pho_list):
        found = False
        for idx_mod, mod in enumerate(mod_list):
            if pho == mod and not matched['mod'][idx_mod]['matched']:
                matched['mod'][idx_mod]['matched'] = True
                matched['mod'][idx_mod]['with_pho_idx'] = idx_pho
                matched['pho'][idx_pho]['matched'] = True
                matched['pho'][idx_pho]['with_mod_idx'] = idx_mod
                break

    return matched
```

Figure 9 Scripts in Python programming language

These two images are provided in the supplementary file too.

More generally speaking, algorithms need clearly ordered data structure and transcripted child spoken language records are – by their nature – not an example of a phenomenon easy to put in rigid boxes or cells.

Language is a continuous phenomenon: every attempt to turn it into discrete units helps rigour and make comparisons possible but – at the same time – force the researcher to make trade-off choices between preserving the originality of what has been said (and then transcripted) and what can be analysed by computational techniques.

A technical problem that soon arised was that a given phonetic unit could be differently pronounced depending on the other phonetic units between it: the "ɹ" (alveolar approximant according to IPA chart, "uvulaire-fricative" in French is different when we utter "rat" (*ɹa*) from when we utter "serrure" (*seˌyˌɹ*), to give another example, the "c" in cactus differs from the "c" in "Collioure".

The same holds for homophones and allophones: the data structure we have derived from the original corpus and the algorithms we applied on it to look whether *pho* is equivalent to *mod* both in SPVR and Normalised Levenshtein Distance do not take into account – for instance – if the child can properly pronounce homophones such as "mère" et "mer" ou "je vais *jouer*" et "j'ai *joué*" in the same way.

As the algorithm works on IPA characters, it only relies on the degree of differences and similarities between the two distinct strings of graphemes, without taking in account anything else and without keeping memory of previously occurred correct forms of the same word.

So, it can happen that a child properly pronounces a very common word such as "mer" (the sea) in its correct form and a less common homophone word such as "maire" (mayor) in a variated form. This is probably due to what we would call a "semantic interference" – as usually ( i.e most of the times) different sounds relate to different entities – rather than an indirect influence of the written form to the oral one (as in the previous example the words are homophones but not homographs). A similar consideration needs to be made for verb conjugation: French language has the peculiarity to have many more agreements in the written form rather than the oral. For instance "jouer" in its infinitive form is pronounced in the same way as in all the past tense plural (and gender) forms and in the second plural form of the present simple. The inverse holds for the same verb (and almost all the first group of

French verbs ending with the suffix "_er") that it is pronounced in the same way in the "imparfait" and "conditional" conjugation and it is typed in the same forms (je jouais, tu jouais, il jouait and so on). Other similarly structured Romance languages such as Italian have a more transparent grammar: every verb in every conjugated form is different from each other bith in its oral and written form.

Other important examples of homophones in French are gender and plural written morphological differences in the suffix: to write the feminine form, most of the time you add an "e" that is not pronounced in the pral form. The same holds for adding a "s" or a "x" in the plural forms: these forms are pronounced exactly in the same way of the singular form, with the notable example of the typical phenomenon of "*liaison*" (that should not be written but it is pronounced in certain circumstances, giving a more harmonical structure passing from one word ending with a consonant and the next beginning with a vowel, e.g "l'eau", "les eaux").

It is hard to program set of algorithms able to recognise these differences: context-dependent sensitiveness is a key obstacle to the developing of a fully-fledged model for these articulatory features.

As the algorithm developed to calculate whether *pho* differs from *mod* (in the form of a Normalised Levenshtein distance too) does not take into account how the position occupied by a phoneme influences the way it is pronounced, it has been provided a reference to the occurrence in the "CHI-pho-mod": by doing so, the reader is able to evaluate the specific context at stake.

Similar technical questions on the transcription of spoken language corpora and on the transformation of transcripted corpora in machine-readable (especially speech-to-text) formats ready for automatic recognition and computing are higly debated topics (cfr. Adda-Decker M.; 2006) for a review and some examples specific to French language.

A possible answer (and future direction of this thesis) would be to take in account the syllabic level (onset-coda or even onset-nucleus-coda): this would probably imply to make a position-sensitive coding for every phoneme *via* a matrix vector.

Another possible improvement would be to take in account child directed speech (IDS at the earlier ages and CDS later). In this thesis has not been taken in account simply due to lack of time: modelling this part would have required an amount of computing resources, data filtering and additional interpretative difficulties on results that would have probably made

this work much longer. Undoubtedly parents input plays a fundamental role in first language acquisition: *motherese* represents an instinctively way through which parents tend to modify and adapt their language to make it more suitable to their child, hoping that he/she will be able to extract and learn as much information as possible from that sequence of information.

Previous research has demonstrated how

"the frequency distributions of utterances produced by children and their caregivers are generally extremely similar[82]"

So, a main question arises: *"*with what knowledge, if any, does the child begin?"[83]

CoLaJE has been conceived to take in account child directed speech too: every parents's occurrences – either in verbal or non-verbal forms – are transcribed and coded in specific *tiers*, MOT, BRO and FAT (and even OBS). It could have been possible to develop a model to see whether and how CDS evolves over time, in other words if parents - phonologically and lexically speaking – do fine-tune their language as they become progressively aware that their child's language is evolving at different paces.

Almost every graph of this thesis would have been potentially ready to be created even for adult language, thus providing a way to highlight and evaluate the mutual interactions between them.

Phonetic proportion graphs, as well as the Multiresolution streamgraph are potentially interesting tools for analyzing this interactional dynamic: by displaying monthly records' differences and similarities over time it could be possible to see eventually occurring changes in the phonetical structure of their production and, to give an example, test whether bilabials are less frequent when the child is five instead of when he/she is 2 year old.

Then, from an *inter*-children comparison perspective, it could be interesting to superpose graphs coming from different children and test whether the quantity and quality of parents'input correlates with already existing index available in this thesis such as SPVR,

---

[82] Ambridge B.; Kidd E.; Rowland C. F.; Theakston A. (2015). "The ubiquity of frequency effects in first language acquisition". Journal of Child language. 42. P239-273. P 242

[83] Vihman M. & Kunnari S., 2006

NLD (Normalized Levenshtein Distance) or with other index such as type/token ratio, mean length of utterance, number of total words per hour[84].

From such kind of assesment it would probably be possible to see whether CDS is a predictor of learning rate and/or learning outcome at a given age or not. Many intertwined variables are at play in this outcome and the contingent nature of first language acquisition will interfere in outlining this hypothetical correlation. Yet, in the case a strong correlation was found in all the seven CoLaJE children, it should then be fair to deepen this research question and look if a certain kind of CDS (in terms of quantity and quality) would trigger in a significant way better learning outcomes.

These ideas will probably be tested once the thesis will be finished.

[84] Morgenstern A; Parisse C. "The Paris corpus" Journal of French Language Studies 2012

# Chapter 4 - Phonetic and phonological aspects

## 4.1 What is a phonological theory

The main aim of this thesis is not to propose and defend a theory of acquisition, trying to explain quantitative results derived from CoLaJE *corpora* through a particular lens. The choice of Clements's "Theory of phonological traits" is simply aimed to account to results and give them a phonologically-informed order. Thus, results will be compared to current theories on acquisition and, more generally, to basic references such as consonant acquisition order baselines (see annexes), but they will not be thought of as a way to confirm and/or refute existing and competing theories.

The main objective is to highlight all the possible interesting insights from quantitative results in the most objectively possible way. The risk is falling into a poorly state of "descriptive adequacy", in which prudence would bring me to simply describe data and results without trying to explain them in a rigorous way.

Despite so, it is important to summarize what it is considered to be fundamental in current debates on first language acquisition in order to raise awareness on past and current questions that are still seeking answers.

According to Fikkert, a theory of language acquisition:

"must first give a characterisation of the developmental stages of language acquisition [and] must provide a characterisation of the errors children make when acquiring their first language. However, a theory of language acquisition must further explain why certain types of logically possible errors do not occur (Brown, 1973)[85]

In this thesis, some particular learning path will be proposed, and the consequent possible/impossible errors (errors will be rather called as "variations" to the norm), for instance why "tracteur" can be pronounced in its variated form "kracteur" but not as

[85] Fikkert P., 1994, p16

"bracteur" or "fracteur" because of – it is supposed – the effect of the markedness avoidance principle stated by Clements ("k" is a voiceless consonant as "t", it differs from the latter for the place of articulation: it is not a coronal, it is a dorsal).

She continues by writing that:

"a theory of language acquisition must not only make explicit exactly how development takes place, but also specify what the triggers are for the transition from one stage to the next. In other words, what is further needed is a LEARNING THEORY which explains the patterns of development [..][86]"

This thesis will not address this problem, it will simply propose a theoretical stance and try to check whether statistical inferences and computer graphical models could provide improved empirical ways of mining longitudinal *corpora* in search of underlying structures.

« '*épistemologiser ' l'acquisition du langage* » (Sauvage, 2015) in English would sound like « 'to epistemologise' first language acquisition » is the aim of the essay that I have adopted as a reference to give a framework to my thesis.

The author chose a foreword that I would like to recall, because I think it will be essential for understanding what it will follows:

« [..] *Il nous faut regarder la façon dont nous concevons l'ordre, regarder la façon dont nous concevons le désordre, et nous regarder nous-mêmes regardant le monde, c'est-à-dire nous inclure dans notre vision du monde* [87]».

The core question is to reconsider language acquisition through the lenses of complexity theory and non linear dynamics' modelling: it is widely acknowledged by the scientific community that from birth to approximatively the age of 6 year-old every child will be able to learn his/her own native language(s), but there is a lack of demonstrated empirical knowledge on what is going on during those years.

After having introduced the debate between nativists and empiricists, the essay focuses on French phonetic acquisition in order to propose arguments in favor of a renovated framework to study language acquisition, highlighting on some particular phonemes.

---

[86] Fikkert, 1994, p16

[87] E. Morin, *La complexité humaine.* 1994, p 301.

Below there is an introduction about what have been recognized to be the main stages of language acquisition (Moreau & Richelle, 1981 ; Fletcher & McWhinney, 1996): it would be useful to frame in a better way the future development of my thesis.

## 4.2 The pre-linguistic period

During the 5th month of gestation, cochlea begins to function and the baby starts to hear his/her firsts sounds. He begins to catch the prosody of the language of his parents and, consequently, he starts to specialize his neuroperceptual system to certain sound patterns.

According to Sauvage « in neurophysiological terms, perception precedes production as well as comprehension will always precedes production [88]». The fact that the baby develops some key anatomical parts for speech production from the age of six months is an argument for this claim:

« [..] glottal and pharyngal's volumes grow as well as trachea's orientation, allowing the articulation of voiced sounds (Kent, 1981 in Sauvage, 2015) »

At the same time, infant's breathing changes from nasal to oral bringing him to the babbling phase. This period is extremely important to future development and many hypotheses are on the table. For instance – according to Westerman and Miranda (Westerman & Miranda, 2004):

« it is often hypothesized that the first speech-like articulations and the babbling phase between 5 and 10 months of age allow infants to develop a link between articulatory settings and the resulting auditory consequences. This link forms the basis for the development of the phonetic inventory and the adaptation to the ambient language by exposure to other speakers [89]»

---

[88] J. Sauvage; "L'acquisition du langage. Un système complexe ». L'Harmattan, Louvain-la-neuve, 2015. P 60

[89] G. Westermann ; E. Miranda (2004) « A new model of sensorimotor coupling in the development of speech » . Brain and language. Elsevier.

There is an open debate on babbling in which two hypotheses are at the opposite sides:

- The **discontinuity hypothesis** (Jakobson, 1941/68) according to which babbling phase has no role in the later language development. It is a claim who is rarely supported by researchers due to the lack of empirical evidence to confirm it. In few words, supporters of this claim argue that there is no real order that we could infer from the first twelve months: infants would randomly vocalize in order to experiment many possible sounds and the relative articulatory possibilities, any relation between the exposure to caregivers' sounds and these vocalizations is not considered. Then, after a certain age, children would be able to learn orderly and progressively all the sound patterns and rules of their phonological *milieu*.

- The **continuity hypothesis**, according to which the babbling phase would function as a preparatory stage to later language development. This claim is widely supported in the scientific community because a growing amount of evidence would seem to be coherent to it (Goldstein et al., 2008). For instance, evidence shows that early babbling is approximatively identical in every culture due to the identicity of the anatomy of the vocal tract (Kuhl P. & Meltzoff A., 1996). In other word, we should say that infants produce « universal sounds » that are in some way experience-independent and then, approximatively after three months, when babies begin to imitate adults' sounds and experience-dependent factors starts to play a greater role than before, researchers would begin to be able to infer some babbling's features that seem to be language-specific. For instance, infants that grows in French speaking families would babble in a rising intonation during CV sounds compared to the babbling of infants raised in English speaking families, in which no particoular rising intonation could be found.

In more general terms, we could say that according to the *continuity hypothesis* a sort of retroactive feedback between infants and caregivers would play a key role in the selection of some reduplicated canonical babbling sounds (*e.g* "ma ma " and "ba ba" instead of other bilabial sounds such as "mo mo") in a way that some type of babbling would be "canalized" by caregivers' signification on it, reinforcing some sounds in spite of others, thus allowing a process of a progressive focalization on the association between sounds and external reality.

Evidence shows how sensorimotor coupling is mainly learned rather than innate, as well as fundamental to the continuity of development of speech found a direct demonstration in the observation that deaf infants do not babble normally and thus often they do not develop comprehensible speech (Oller & Eilers, 1988; Wallace, Menn & Yoshinaga- Itano, 1998)

So, just to give a quick glimpse on the first twelve months of life, we could say that from isolated sounds (mainly cooing and cries demanding help) to babbling and then, from 6 month-old onward, canonical babbling enriched by adult-like stress and intonation, there is a lot to explore.

This process, according to Sauvage, is

« a parallel training in which phonological knowledge and articulatory ability develop together [90]».

As he points out later, pointing gesture (Tomasello, 2001; 2003) « as an action to express joint attention that normally emerges between 9 and 13 months, will serve as a direct contribution to the first symbolic operations [91]» and I would underline that this cognitive ability emerges approximatively at the same time of the first « proto words » used by infants to gain parents' attention.

## 4.3 The linguistic period

Between 9 months-old and 1;8 years-old infants pronounce their first words, with an average of 12 months-old of age. We have to remember that every baseline in first language acquisition is only a general guideline because children display among themselves a large variability.

Since this period

[90] Sauvage, 2015, p61

[91] Sauvage, 2015, p61

« the lag between production and comprehension grows exponentially as children are progressively able to understand complex sentences while they produce just one word at a time, even if this single word is often a holophrastic one[92] ».

Lexical inventory is initially made up of simple association like a word and his related adjective. Little by little lexical inventory grows in terms of number and in terms of related combinations until the age of 24 months-old, when an « *explosion grammaticale* » (Slobin, 1971) takes place allowing the child to learn his/her language's specific syntax.

To give an idea of this exponential growth, results derived from parental questionnaires and quantitative data from longitudinal corpora showed how – on average -from 18 to 24 months-old active vocabulary expand from 50 to 300 words, then 900 words at the age of 36 months-old (Fenson L., 1996; Kern S., 2019).

A number of researches in this field show how the amount of words that a child can understand is far higher than the number he can say (Bates, 1993). The same author has demonstrated the difference in the use of words during this period: around the age of 18 months-old children use mainly context-dependent words while, after 18 months, their mastery of reference allow them to use words that are context-independent (for a similar observation see the stackgraph "evolution of POS tags in chapter 10, especially the consideration that has been made regarding the development of pronouns).

Phonological development is strictly linked to the lexical and the syntactical ones.

Since Jakobson (1939) modern linguists have tried to give an account to the order of acquisition of phonemes, as well as to account for the interindividual variation of them and the relative dynamics of this cognitive development.

The main difficulty seems to be in the language itself: it is hard to find universal pattern of phonetic acquisition, while it is clear to a growing amount of linguists that every language has its own peculiar schemes of appropriation.

Aim of this chapter would be to « redefine the object of study in a dynamic perspective. This means to start to ask ourselves not simply

' how children learn to speak ? '

[92] Sauvage, 2015, p62

but rather to ask ourselves

'how children do really speak? »[93]

Although my project is mainly about phonetic acquisition, I think it is good to contextualise this important aspect of first language acquisition with another dynamics of acquisition that develop in a parallel and complementary way to it: the process of symbolization (Sauvage, 2015, p97).

Learning a language implies learning a referential system, that is a conventional way to convey meaning through sounds. Learning a language is even the mean by which children introduce already existing external representations of the world in their brains, developing a skill that will allow them to semiotize future experiences through words.

As Bronckart stated, language implies « *se répresenter ses répresentations* » (Sauvage, 2015, p97) that, in other words, is one of the key concept of Dennett's last essay on consciousness: « noticing that you're noticing » (Dennett D.; 'From bacteria to Bach, and back' ; 2018).

This ability can be seen from an evolutionary and developmental perspective and it is considered to be an emergent cognitive phenomenon that humans share to few other species (Plotnik & de Waal, 2006).

Being able to represent herself as a distinct cognitive unit compared to the others forms of life (including, of course, conspecifics) is a metacognitive question hard to observe and test: attempt to study consciousness through neurosciences (Hameroff, 2006), behaviour (Gallup, 1970), philosophy (Chalmers, 2010) or by thought-experiments (Hofstadter, 1979) is an on-going and never-ending scientific effort.

The first stage a child encounters in his/her development toward a fully conscious state is well exemplified in the "*peek-a-boo*" game: during the first year of age, a child does not still feel herself as an entity separated from his/her parents and the outside world, consequently she does not realize that something could happen even if she is not there to see this thing happen.

Little by little, when the mother will hide her eyes with her hands, the child will gradually be less scared by the fact that the mother will not be back smiling her as she disappeared behind

[93] Sauvage, 2015, p64

her hands: excitement will then substitute fear as the child begins to learn from experience that entities of the outside world (including mummy) exist independently from her sight.

An historical example of these studies has been given by Piaget's theory of the six stages of "object permanence" (Piaget J., 1976), while for recent up-to-date studies on perceptual models and constraints in infants and the age in which they developed see the works of Bremner G. (Bremner G. et al, 2015)[94].

Because of its targeting a core learning process, "peekaboo" is not by chance a cross-cultural game loved by all babies. Every one of us has learnt that things have a physical existence even when they are out of sight: yet is difficult for adults to grasp this obvious concept, but the first and most important effort adults have to make when studying child language and psychology is that anything is already given for babies, they need to deduce from experience every concept that is already structured in adults's brains.

According to the Swiss author, children spent a lot of time laughing at "peek-a-boo" and other similar games because they need to test as much as possible the "object permanence" (Piaget J., 1976): then, once the "sensorimotor stage" will be completely reached, children will have integrated tactile, visual and motor representation of the outside world in a coherent way and - after the age of two – peek-a-boo will progressively become less and less surprising until it will disappear.

It would be fair to bet that linguists will be able to explain the traditional core questions of their domain such as the arbitrariness of sign, the groundedness of symbols and phonetic/phonology interface once consciousness will be exhaustively described.

This because until it will be impossible to know whether representations or computations (or computations of representations) are in our brain, it will then be impossible to shed light to any other theory accounting for first language acquisition because it will be inevitable to assume extremely important questions that would seem to be at the root of language.

The mirror mark test (Gallup, 1970) and all its different subsequent versions (Rochat et al., 2012 for a state of the art) proved what is needed to pass this test and claim that passing this test means to be "[able] to generate mental models of the self "(Rochat et al., 2012).

---

[94] Bremner G.; Slater A.; Johnson S. (2015) "Perception of object persistence: the origins of object permanence in infancy". *Child Development Perspectives*. **9** (1): 7–13

The progressive building of a phonological consciousness by the child goes in parallel to the progressive rise of self-consciousness : when an infant start to realize that his/her own body is something apart from the one of his/her mother, when he/she starts to recognize him/herself in a mirror is – approximatively – the same period in which infants start to listen to his/her own sounds, beginning to notice that his/her way to speak is quite different to the one of his/her parents and, by this way, a reinforced process of learning will enable infants to momentarily avoid sounds they know they are still not be able to properly pronounce, as well as looking for synonyms, reformulations and self-corrections (see for avoidment strategies and self-repairing strategies)

As the authors of the study stated in their conclusion:

"Based on the population tested in the present study, it appears that for typically developing children, early mirror self recognition is linked to social awareness. We view such link as the landmark of human sociality that forms around a propensity toward self-consciousness and a unique concern for reputation (Rochat, 2009). Lacking the core propensity toward self-consciousness and the unique concern for reputation could be a major obstacle in human social-cognitive development, the kind of obstacle encountered by autistic children in their development[95]"

Thus children from about 22 months of age successfully pass the mirror mark test showing that they are capable of perceiving their selves as a "self" among others.

But how this relates to linguistic ability?

Typical abilities that require self-consciousness (knowing that you know that the other knows that you know, and so on) such as joking, lying and feeling ashamed emerge around 22 months and then develop in quantity and quality.

What is common among joking, lying and feeling ashamed is that the child has expectations on what is going to happen next based on what has happened before (what s/he has experienced so far): joking implies to say something that is out-of-norm (knowing that this given norm is socially shared), lying implies knowing what a child is expected to do in a given situation, knowing the others' expectation about that and dissimulate his/her

[95] Rochat P. (2012). "Social awareness and early self-recognition. Consciousness and cognition 21. Elsevier. 1491 – 1497. P 1496

compliance to the norm, while feeling ashamed requires knowing what the others expect to be done in a given situation and realising that this norm has not been properly adopted.

What a child finds funny or shameful gives researchers rich information about how s/he represents the world in his/her mind and the stage of development in which s/he currently is: if a two-years-old child laughs at the view of a ball not falling as it should have done because of the gravity effect is because s/he knows what to expect and the violation of this expectation seen in a tv cartoon looks funny to him/her. For similar reasons a two years-old child usually laughs when hearing a sound sequence which is not a word that is presented just after a real word: that particoular combination of sounds is unusual, it does makes sense because non-sense syllables are not really used by adults: this is funny because it is out of what is perceived to be as ordinary experience.

It would be interesting to cross these findings on mirror mark test and social awareness and other studies concerning first language acquisition (French in this case) targeting specific linguistic structures that would seem to imply this cognitive hallmark: do they evolve together? Does the pronoun “je” could be considered as a sign and/or a precursor of social awareness[96]? Do other pronouns such as second singular form ”tu” and third personal form ”il/elle” would be a sign of an increased social awareness[97]? Does pragmatic inference is linked to metacognition and how this is reflected in language?

For sure it is possible to observe that exponential increase of tokens and types (and type/token ratio too) starts around 22 months (see the graphs draw from “The Paris Corpus” article).

This short overview on consciousness helped to give an introduction to the following related topics: children’s ability to become aware of their own language (and its related consequences) and the building of a phonological consciousness.

It is easily observable how reformulations, hesitations, and avoidment strategies are detected from around the same age (Sauvage, 2015), this would bring us to think that children of this

---

[96] Morgenstern A. (2006) ”Un ”Je” en construction: genèse de l’auto-désignation chez le jeune enfant”. OPHRYS Editions. Paris

[97] Orvig A. et al. ”Dialogical beginnings of anaphora: the use of third person pronouns before the age of three”. Journal of Pragmatics 42 (7). 1842-1865. See results on chapter 7 of this thesis, automatic parsing

age begin to clearly be able to listen their own sounds and know that the other knows that they know.

« Furthermore, as there is always a constant negotiation of the arbitrariness of the sign and its relative signification, a child will become a 'linguistic agent' and a 'social element' by his/her way of progressively interpreting rather than passively absorbing adult's symbolic systems [98]».

[98] Sauvage, p98

## 4.4 An example of phonological development

« Plosive» is a large category describing a manner of articulation that – in the aim of this thesis - has been used to comprehend six corresponding consonants classified in the IPA alphabet (see the latest version of the IPA chart in the Annex) as plosive (or more commonly « stops) /p/, /b/, bilabials, /t/,/d/ dental-alveolar /k/,/g/ velar plosive.

According to Sauvage, « the most intriguing phonological phenomenon in first language acquisition is variation and the status we should give to it » (Sauvage, 2015, p99).

Phonological variation could be viewed as « a physical realization of a mental state that would be considered as a step in the process of the building of the representation of the phonological system of a given language » (Ibidem, p99).

Phonetic acquisition is a sort of « learning by doing » process because children build up a structured representation of the phonemes of the language to which they are exposed to. Children have to use it despite the fact that they don't know it completely and how to use it appropriately.

A sketch of this complexity could be given by this graph:

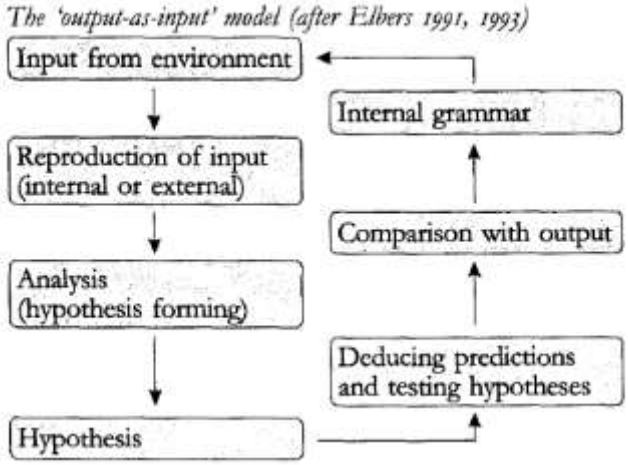



Figure 10 « output as input model » [99]

[99] Fikkert P, 1994, p 26 "the output-as-input" model

“if the child's own system does not contain velar plosives, but adult input forms do, there are in principle two ways in which the child can deal with these input forms. One is to simply avoid such forms in his or her own production forms. In other words, the child only selects forms that fit into his or her grammatical system at a particular stage. The second strategy is one of repair. In this case, the child selects adult input forms with velar plosives, but the velar plosives are either 'deleted' or 'replaced' by some other place of articulation [..] » (Fikkert, 1994, p13).

To have an idea about the complexity of phonological variation in French, let’s give an example from a direct experience: in the same recording session (Sauvage, pp103-104) a child that is learning how to pronounce plosive-liquid phonemes (onset-rhyme units as [tr][gr][kr] and [dr]) performed what could be interpreted as a paradox.

To give a premise, according to certain interpretations of theories that link together motor development and language acquisition

« development of language should be viewed in the context of the body in which the developing language is embedded. In infancy, there are significant changes in the way in which the body moves in and interacts with the environment; and these may in turn impact the development of skills and experiences that play a role in the emergence of communication and language[100]».

Anatomical changes can be viewed as a basic driver of change for phonetic variation: for these reasons it would be plausible to interpret children tendency to reduce [gr] to [dr] to the fact that the second form is easier to pronounce due to the fact that you prononce [d] at the top of your mouth and [g] at the bottom (Sauvage, 2015). Anyway, other linguists argument the other way round, claiming that it is easier to prononce two consonants in the same area of the vocal tract (at the bottom).

It is not the case to verify whether one of the two options is correct because the puzzle in this case it is still there as the child reduces the French **«** grand » to « dran » and - during the same conversation, despite adult’s correct pronunciation of the same word – he reduces the French word « druide » to « gruide », doing exactly the opposite.

[100] J. M. Iverson ; «Developing language in a developing body : the relationship between motor development and language development ». J Child Lang. 2010; 37(2): 229–261. P 230

So, even if he knows how to utter it in the correct way, because he already pronounced it in a correct way, he deviates from the norm for reasons that are still unclear to the scientific community.

This is an example to what will be a core question: do these changes are purely random?

Do they follow some developmental patterns that we still do not clearly see?

How do we have to deal regarding exceptions to adult norms?

“mismatches between the adult input form and the child's production form will be argued to be the result of mapping the adult target word onto the child's template. The child's template at each stage of the development determines the relation between the input and the output” (Fikkert, 1994, p13).

To give another example to better focus on this case: in the same recording session the adult – pointing to a big familiar object - ask to the children:

« c’est [dra] ça ? » instead of correctly saiyng « c’est grand ça ? », in a deliberative way, in order to see the possible reaction of the child, that in turn answer to him « on dit pas [dra], on dit [dra] !! [101]».

This exclamation reveals the lag between perception and production that is typical to the child of this age: their ability to perceive what an adult say is better than their ability to articulate it in the same manner**;** going into details, we can deduce – at least for this single case – that the child is not completely conscious of what he is saying : he can distinguish the difference between [gr] and [dr] when he listen to it, but it seems that he cannot distinguish the same difference while pronouncing it.

As Fikkert points out:

“the input forms violate the child's phonological system which s/he is building up. Therefore, repair strategies appear. They alter the input representation in such a way that they no longer violate the child's grammar[102]”.

[101] Sauvage, 2015, p125. See also the tabel concerning the evolution of the pronunciation of the word “regarde”

[102] Fikkert, 1994, p 14

In current literature on first language acquisition there is an effort to explain this mechanism through Bayesian statistics[103], this thesis is not the place to develop this idea but it should be said that the « violation of the input form » and the « repair strategies » could be considered as analogous mechanisms of the « upper bound » that has been explained in the short overview of Friston's FEP.

Another example of this lag between perception and production is that at 20 months-old a child can understand on average 50 words while he is able to pronounce just around five (Sauvage, 2015). This lag lasts for the following ages, some claim until the adulthood[104].

## 4.5 The non linear nature of phonetic acquisition : an hypothesis

Every infant has his/her own learning path: as in every complex system with many interacting parts in which the whole is more than the sum of its parts, the occuring dynamics are for some aspects similar due to common constraints (*e.g* the anatomy and physiology of the development of the vocal tract) but they can largely differ from many other aspects.

Impairments taken apart, every child will learn his/her native language (or the language to which he/she has been exposed mostly, in cases of bilingual environments) during the first six years of his/her life. But every child will display a unique path characterized by different learning rate, different errors, different hesitations and reformulations' patterns, different repair strategies and so on.

Different environments (in terms of sociolinguistic context) and different time of exposure to parents' language will affect different language development paths and, consequently, different learning outcomes, but – as Lebrun directly observed in 1982 – twins, despite the

[103] See for a state of the art review Pearl L.; Goldwater S. (2016) "Statistical learning, inductive bias, and Bayesian Inference in language acquisition". The Oxford Handbook of Developmental Linguistics

[104]Brysbaert M, Stevens M, Mandera P and Keuleers E (2016). "How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age". *Front. Psychol.* 7:1116

fact that they grow together in the same sociolinguistic environment and they consequently are exposed to the same amount of time to parents' direct speech, they do not show identical capabilities in terms of perception and production.

So, it is a matter of proportion between the amount of what changes (variance) and what remains the same (invariance) over time; to put it in a different manner, children are very different between each other and learn in a very different and « unique » way, as well as the quality and the amount of the information to which they are exposed to is always different and difficult to be objectively evaluated but - despite all these intertwined variables that mix together in an unpredictable way - researchers can often observe similar learning paths (e.g on consonants and opposite traits in French phonetic acquisition; Yamaguchi, 2012**)** as well as they often observe (seemingly ?) random quirks, regressions and paradoxes, as I showed two paragraphs earlier: thus, the main challenge is to improve our understanding on how this variable proportion between differences and similarities develops over time.

A further reflection on the « uniqueness » of every learning path and the puzzle about the variable proportion between differences and similarities could bring to a sense of disorientation: if every path is different from every other, how any scientifically rigorous method could be helpful to draw a generalization or just even some partial deductions?

I will explore this question in the next chapter.

I think that a simple graph in cartesian axis would give an introductory idea[105]

[105] Sauvage, 2015, p102

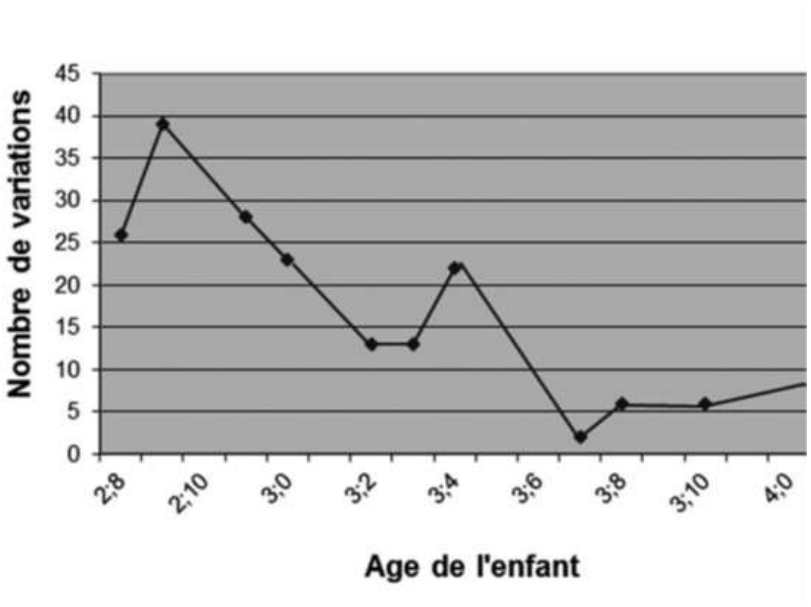



Figure 11 : Phonological variations by age

This graph derives from a number of recorded observations of phonological variations on a specific child during 18 months.

What can we learn from the non linearity of this development?

Is this simply randomness or is there an order that we are not still able to tell?

Why are there pitches? Why is not there a steady decline of the number of phonological variations over time as some intuitive logic could suggests?

And above all, two questions:

how should we interpret a regression ?

Is it proper to define it « regression » or should we rather define it a « variation » in the perspective of a succession of improving steps?

In general terms, we can almost always see a decline of the number of variations during every time span beyond one year, but what is interesting is what is going on inside these temporal windows: when we are dealing with non linearity we are always facing the problem of predictability.

But here is one of the main questions, that I think depends mostly on the level of analysis that we want to adopt: as I have written in the introductory chapter on complexity, when we are in

front of a system with many interacting parts where the same cause gives always rise to different outputs (see the concept of bifurcation in the logistic map) , it is almost impossible to predict – on the basis of the interactions at a given time – what will going on at *n* time steps later while – using differential equations – it could be possible to predict what will going on at *n+1*, at the next time step.

The fact is that in phonological variations it seems to be exactly the opposite: we know that at the age of six every child will be able to speak without any variations but we have any idea on what would be the next variation in the short term (*n+1*).

To put in other words, in the long term we know that a child will learn a given phonological norm but, if we take in account a short term ( a month, for instance) « every acquisition of a phonological unit seems to be a temporary one[106]».

So, the core aim of this Phd project is to try to shed some light on the following puzzle that we can observe if we focus on short term phonological variations:

« It has been demonstrated that any onset does not randomly vary in any possible other onset and, to go in further details, it has been observed that the process of neutralization of this kind of variation was based on a parallel process: the building of the representation of a phonological system mainly driven by adults' intervention[107]»

This observation would bring us to focus on the relation between the level of accuracy of the pronunciation of a phonetic unit and its correspondent phonological semantic reference.

A hypothesis could be that different steps in the process of the building of a complete phonological representation would create articulatory difficulties at the phonetic level and, consequently, once the child has built up the whole representational structure of phonological units, then he/she would begin to be able to neutralize every variation, starting to speak « as an adult ».

---

[106] Sauvage, p103

[107] Sauvage, p103, this is the original extrait in french « « Il a pu ainsi être montré que n'importe quelle attaque ne variait pas en n'importe quelle autre, et que le processus même de neutralisation de ce type de variations faisait appel à une répresentation du système phonologique résultant de l'action d'autrui »

The point is that this hypothesis is hard to verify because phonological variations are quite unpredictable. Researches adopting optimality theory have been conducted in order to model these kind of phonological variations, and researchers have tried to deeply focus on case-study to see the development of phonological variations : for an example, dos Santos on the Lyon corpus (dos Santos, 2007).

To give an idea of this quite abstract concept, I would like to show a direct observation in which the same child, in the same short sentence, shows – in the following order – a hesitation, a variation, a correct prononciation and finally another variation, always on the same target word « trouve » that starts with a plosive-liquid phonetic unit [tr][108].

« Nous pouvons alors établir ces variations de base à propos des attaques /gR/, /dR/,

/kR/, /tR/ comme suit:

/gR/ → /dR/

/dR/ → /gR/

/tR/ → /kR/

/kR/ → /tR/

« 52 Q : ben attends, on essaie de/ de l'touver, si on le trouve pas ze/ ben c'est pas grave hein, ça c'est un gros euh c'est bien… si on le krouve pas, alors c'est pas grave…donc, [..] [109]»

This example shows how much is hard to deal with variations: how could we explain this sort of « false improvement » of the child ? Does he knows what he is saying or not? Does he knows that, despite his incorrect pronunciation, the adult would probably understand what he is referring to?

[108] Sauvage, 2015, p105

[109] Sauvage, 2015, p105

In other words, does he know that the adult knows that he knows that he is not still able to speak properly?

Can we establish the degree of phonological consciousness a child can have at a given time? If so, by which means?

And, above all, can we model this process of neutralization of variations in a formal (logico-mathematical?) way?

I would like to give another example of the (apparently) randomness of variations that I will use in the forthcoming chapters. Once again, it seems that a lag between perception and articulation is at work:

109 A : on va jouer aux playmobils
110 J : On joue a ca ? C'est quoi ca ?
111 A : (chuchote) **un kracteur**
112 J : un quoi ?
113 A : **(k) racteur** ?
114 J : j'entends pas ce que tu me dis
115 A : **(k) racteur**
116 J : j'entends pas, faut que tu me parles normalement ! ca s'appelle comm/
117 A : **(k) racteur** ! (toujours en chuchotis)
118 J : parle moi normalement comme ca avec ta grosse voix (je fais une grosse voix). Alors, ca c'est quoi ?
119 A : une balancoire !
120 J : et ca c'est un ?
121 A : **un kracteur**, je t'ai dit
122 J : d'accord j'ai bien entendu maintenant, tu vois c'est bon, il faut que tu me le lises (*sic*). Bon alors on joue tu fais quoi, toi ?
123 A : moi je fais… je monte le le le le le monsieur il est dans **le tracteur** tu vois ?
124 J : d'accord, et pis moi, je le mets dans la remorque, d'accord ?
125 A : nan s'assit pas
126 J : mais si, i s'assit !
127 A : voila ! c'est **le kakteur**, lui, oh i va ecraser les p'tits bouts
128 J : i va ecraser les p'tits bouts ?[110]

[110] Sauvage J., 2015, p109

How could we find the edge of the skein from this heterogenous data?

Is there an underlying logic that could link every variation to any another, allowing us to see a path that is developing itself over time?

Does a retroactive feedback between perception and articulation is at play? If so, how could we infer this dynamic circle of causality from data?

Crossing different kind of data, longitudinal and cross-sectional, while approaching them with different theoretical framework, as well as different processing methods seems to be a plausible hypothesis to solve the puzzle.

The general theoretical framework that would allow us to put together all these data on child language acquisition could be the one who takes the perspective in which any variation would be considered as a temporary achieved structure at a given time in the development that - in turn – will structures the next possible variation, working as a costraint, and so on.

So, a dynamic causal circle that – over time – would allow us to see how different variations could be hypothetically linked together on the basis of a subsequent articulatory and phonological co-organization.

Do children are conscious of what they pronounce? And, if so, to what extent do are they?

At which age children acquire a fully developed consciousness allowing them to compare what they hear and what they pronounce and consequently evaluate differences and similarities?

The first way through which children aquire their language is the parental input and, more generally, everything coming from the environment. Is is called “positive evidence” (Fikkert, 1994)

« intervention de l’adulte » (Sauvage, 2015) and “direct negative evidence” (Fikkert, 1994)

Which is the role of this external intervention, what can trigger?

Or, to put it in other words: does a child realize that what s/he is saying is not pronounced in the same way as parents do?

Obviously this degree of consciousness largely depends on age: around 18 months old children start to become aware of theirselves generally speaking (*e.g* a recent re-edition of the “mirror mark test” in Rochat P. et al. 2012)

According to Sauvage (2015), when a child reformulates the words s/he has listened from their parents with other semantically similar words that are easier for him to be articulated or when a child begins to hesitate before pronouncing a given word, it could probably mean that s/he is becoming conscious of what s/he is saying and the related gap between perception and production is losing its importance.

It is possible to observe two kind of reformulation:

"on observera l'autoreformulation étayée, lorsque, par exemple, l'enfant opère une reprise à la suite d'une intervention de l'adulte[111] »

[111] Sauvage, 2015. P126

# Chapter 5 - Recording, sampling and population

## 5.1 *In vivo* vs *in vitro* data

There are two kinds of data available for inquiries in first language acquisition: *in vitro* and *in vivo:* the former comes from well-defined experiments conducted in a lab, it is a goal-oriented and elicited way of inducing a child to perceive, say, spell or read something while the latter consists mainly in recording children in a natural setting without eliciting any kind of responses from them[112]. In the former, researchers set parameters for a task that need to be solved, in the latter researchers simply observe and record children in a broad way, without focusing on particular behaviors.

Besides this distinction there are some exceptions that are tolerated: in CoLaJE it is possible to use a query to look for specific occurrences, for instance if we type "Comment on dit x?" (trad. "how do you say x?) (http://ct3xq.ortolang.fr/ct3xq/check-interro ) the result will consist of several examples of elicitations from the parents. The same holds for Fikkert corpus: parents can ask their children to repeat what they have said or to ask for a particular word, sometimes by inducing them indirectly. A distinction based on the level of control of child's utterances is proposed:

"Typically, when the elicitor is the parent, the data are assumed to be naturalistic, but where the elicitor is an investigator, the data are considered to be experimental, since in the latter case observation is more controlled" (Fikkert, 1994, p 24).

The experiments aimed at testing the transitional probability between syllables sequences (Saffran J., 1996) briefly explained in the introductory part of this thesis are an example of *in vitro* data directly elicited by the American researcher through the implementation of a carefully designed experimental setting. The aim of the test is precise: look how children are

capable to segment fluent speech into words by using the information about transitional probability (supposed to be the only available cue for this task) to discover word boundaries. The experimental setting and the sequence of stages are described in the following way by the author:

"In this procedure, infants are exposed to auditory material that serves as a potential learning experience. They are subsequently presented with two types of test stimuli: (i) items that were contained within the familiarization material and (ii) items that are highly similar but (by some critical criterion) were not contained within the familiarization material. During a series of test trials that immediately follows familiarization, infants control the duration of each test trial by their sustained visual fixation on a blinking light. If infants have extracted the crucial information about the familiarization items, they may show differential durations of fixation (listening) during the two types of test trials. We used this procedure to determine whether infants can acquire the statistical properties of sound sequences from brief exposures" (Saffran, 1996, p1927)

I personally find the experimental setting objectively smart and costraints are put in place in a way that the output obtained (duration of fixation) should exactly adress to the hypothesis initially targeted by the author. Despite that, the equivalence between durations of fixation and listening is questionable, as well as the use of a speech synthesizer in a monotone female voice. These two factors could be considered arbitrary and would thus probably interfere in unpredictable ways on final results. If we look at them in Table 1 (Ibidem, p1927), p-values are below 0.05 as required but - as we explained in a previous paragraph – it would probably be more rigorous to set this threshold to 0.005 in order to filter only highly reliable scientific results (Wagenmakers J., 2019[113])

Concerns on *in vitro* results are a lot: compared to adults, children do not (and depending on their age, cannot) accomplish a given task because of their lack of goal-directed attention:

---

[113] Benjamin D.; Wagenmakers E-J. Et al. (2018). "Redefine statistical significance". Nature Human Behavior. Vol2. No 1. P 6-10. "[..] we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with P< 0.05 results in a high rate of false positives even in the absence of other experimental, procedural and reporting problems. For fields where the threshold for defining statistical significance for new discoveries is P< 0.05, we propose a change to P< 0.005 [..].

“Experimental studies are not very suitable for very young children for a variety of reasons. A young child can be easily distracted, can quickly lose interest in the tasks, and has a tendency to tire easily. On the other hand, the disadvantages of naturalistic observation are also apparent. The research is not easy to replicate. In 'raw' data the variables are uncontrolled and unidentified. It is time consuming and the production of certain data, and of certain types of errors, is left to chance and circumstances” (Fikkert, 1994, p24).

There are pros and cons on studying language with one or the other kind of data, usually the advantage of *in vitro* data results in disadvantages of *in vivo* data and viceversa: for instance, reproducibility is strong in *in vitro* data and almost absent in *in vivo,* but reliability of data is stronger in *in vivo* and highly questionable in *in vitro.*

Some authors have a neat position on the matter:

“I had decided that you could only study language acquisition at home, *in vivo*, not in the lab, *in vitro*. The issues of context sensitivity and the format of the mother-child interaction had already led me to desert the handsomely equipped but contrived video laboratory...in favor of the clutter of life at home. We went to the children rather than them come to us “. (Bruner, 1983 in Roy D., 2006)

## 5.2 Sampling child language: a short overview

There are two different ways of collecting data to study language acquisition, what they have in common is that every record is expected to contribute to the drawing of a developmental trajectory from sparse data, in a similar way as a regression line cross its different points to trace the average path that summarizes all of them.

“Longitudinal collection captures the continuous language development of one child, and the premise is that this individual development might be generalized to the global language development of children who speak this particular language. Cross-sectional collection

captures stages of language development in children of different ages, and the premise is that these different stages might represent a continuous temporal development[114]".

A reliable sampling is the base of any rigorous scientific inquiry: there can be no space for any kind of generalization until the sample does not reach a statistically representative part of the population of the phenomena under study.

It is hard to establish "How much is enough" (Stahl & Tomasello, 2004), to put in other word: which is the threshold to reach and possibly overcome to be sure to have a reliable sample in which chance only plays a negligible role in further generalizations made upon it.

"[..]perhaps surprisingly, there has been very little discussion in the field of the quantitative aspects of child language sampling, that is, how much to sample and at what intervals and for how long and for how many children" (Stahl & Tomasello, 2004)

As Vihman pointed out, linguists still do not know the proportion of differences and similarities regarding phonetic and phonological development across individuals learning the same language:

"the difficulty of obtaining data that are sufficiently rich to yield insights (*i.e* data derived from intensive longitudinal studies) while at the same time extensive enough to provide some confidence in the generalizability of the findings (*i.e*., data based on the study of relatively large numbers of children)[115]".

In fact, time constraints, lack of long-lasting funding opportunities and external reasons (*e.g* a family leaves a city and the child being under study cannot be recorded anymore) often impede researchers to obtain an ideal corpus.

An inadequate sampling can potentially lead to estimation errors: not being aware of the "distance" between the temporal density of its sampling schema and the relative frequency of the linguistic structure fatally undermines any research.

---

[114] Naomi Yamaguchi. "What is a representative language sample for word and sound acquisition?". Canadian Journal of Linguistics / Revue canadienne de linguistique, University of Toronto Press, 2018, 63(04), pp.667-685. P2

[115] Vihman, p247

For rarer phenomena, it is possible to underestimate them or simply ignore their existence while for common phenomena, it would be possible to overestimate their relative quantity due to the fact that a sparse sample could catch just them, thus provoking a partial image of development at a given time. A linguistic structure that increases its frequency over time can potentially be detected later than its real emergence, providing in this way a low-quality evaluation of child development.

There have been cases in which a denser sample led to the revisiting of previous results based on sparse sample (Yamaguchi N, 2012): in chapter 6, I will express some doubts about this possibility in the parsing analysis applied to Adrien.

On the study on phonemes there is no doubt that CoLaJE temporal density of sampling is enough, while for POS tags – especially infrequent ones – the same temporal density it is probably not sufficient and has potentially brought me to misleading results. This worry will be discussed later.

It is important to point out that sampling is one of the fundamental steps of a reliable and replicable research: in a given domain sampling should be detailed and standardized with a lot of accuracy allowing other researchers to replicate the same conditions of any given study and giving them the opportunity to confirm or refute it. If this requirement is not satisfied, results from different protocols will not be comparable and the advancement of science will be hampered:

"Known or unknown differences between the replication and original study may moderate the size of an observed effect, the original result could have been a false positive, or the replication could produce a false negative. False positives and false negatives provide misleading information about effects, and failure to identify the necessary and sufficient conditions to reproduce a finding indicates an incomplete theoretical understanding. Direct replication provides the opportunity to assess and improve reproducibility[116].

To reliably compare different results deriving from generalizations drawn from different naturalistic longitudinal observation it would be preferable to have a similar empirical base, that is a similar sampling density (the number of minutes recorded per month for a given time period, e.g one hour per month) applied over a similar age span ( e.g. from 1;0;0 to 5;0;0, as it

[116] Open Science Collaboration. (2015). "Estimating the reproducibility of psychological science". Science **349**

is approximately the case in CoLaJE). This “protocol” should be applied to every child being recorded: by doing so, any conclusion drawn from a set of *corpora* would give us a sound and replicable way, despite there is not a general consensus on how many children for a given language should be recorded to draw rigorous and certain conclusion on language acquisition. To give an example: how many children do we need to establish the baseline describing the order of acquisition of consonants for French?

In this thesis six French children have been analysed regarding their specific consonants developmental trajectory: in Chapter 11 graphs showing phonemes development over time in two different forms (histograms and Multiresolution Streamgrapghs) will show how difficult is to draw any conclusions except confirming already established literature on consonant acquisition order (McLeod & Crowe, 2018).

More generally speaking, a sampling density should be set for tackling these issues (Stahl & Tomasello (2004):

(a) the percentage of the real phenomenon actually captured,

(b) the probability of capturing at least one target in any given sample,

(c) the confidence we can have in estimating the frequency of occurrence of a target from a given sample,

(d) the estimated age of emergence of a target structure

Which kind of speech sample could address all these issues?

An overview of the characteristics of corpus available in many different languages in CHILDES shows that often corpus consist in a variable number of children (let’s say 5) recorded one hour every one/two weeks for one year/one year and a half.

Is this sampling schema reliable enough? The question is not rightly formulated: the research question shapes the sampling density and modality, this will be explained through the “capture rate” formula in the paragraph 5.3.

## 5.3 An “ideal” corpus.

An “ideal corpus” should cover all child verbal occurrences, allowing the researchers to avoid any bias due to the low density of the sampling. This “ideal corpus” would be totally replicable and would reach a complete objectiveness as recordings would simply be the “film of his life”, recording nearly everything a child says without any modifications.

An example is given by Deb Roy, linguist at MIT Media Laboratory in Boston, who directed a pilot research on first language acquisition called “Human Speechome project[117]”:

“The recent surge in availability of digital sensing and recording technologies enables ultra-dense observation: the capacity to record virtually *everything* a child sees and hears in his/her home, 24 hours per day for several years of continuous observation [..]” (Roy D., 2006).

To overcome obvious privacy reasons, the Canadian researcher chose to record his son in his home. The additional strenght of this project is that it eliminates the Labov’s paradox (the observer effect):

“We have designed an ultra-dense observational system based on a digital network of videocameras, microphones, and data capture hardware. The system has been carefully designed to respect infant and caregiver privacy and to avoid participant involvement in the recording process in order to minimize observer effects[118]”.

By doing so, social interactions between the child and his/her caregivers are not unwittingly influenced by someone who is recording them: it is not clear how much a child is influenced by the observer, it could be fair to say that this probably depends to varying factors such as child’s personality, age, whether it is familiar with this person or not.

In Roy’s Speechome Project is clear that the child will be completely unaware of hidden cameras and microphones, but it would be possible to ask ourselves to what extent this kind of “Orwellian sensation” of being continuously recorded every second for three years and

[117] Roy Deb et al. (2006). “The Human Speechome Project”. Proceedings of the 28th Annual Conference of the Cognitive Science Society.

[118] Ibidem, p1

then knowing that recordings will be studied for several years to come would probably modify in some ways spontaneity between children and parents.

As it is quite hard to get a fund from MIT or NSF, it may be better to wonder how to avoid using such a massive recording and looking for more feasible projects. In the case the aim is to study the phonological/lexical level, ideally representative would be a sample

"long enough to reflect as faithfully as possible the child's productions but short enough to be transcribed in a reasonable amount of time" (Yamaguchi N., 2018)

In fact, linguistic development can be difficult to predict and is far from being a linear process in which different variables grow together in a directly proportional manner according to age.

Here a list of expected results from a study of two different *corpora* (called "Prems" and "PSPT") having sampling temporal densities similar to the "CHILDES standard". These expectations have been partly unvalidated by quantitative results, sometimes in a surprising way:

"1. We predict more word types in a long session than in a 30 minutes session, since the children are engaged in more and potentially more diverse activities.

2. We predict more word tokens in a long session than in a 30 minutes session, since the children have the possibility to produce more utterances.

3. We predict no difference in the number of target sound types between long and 30 minutes session, since thousands of instances of sounds may occur in 30 minutes, so every phoneme of the language has chances to be produced. The same applies for produced sound types, since the children have the chance to produce many instances of every sound they make.

4. We expect more produced sound tokens in a long session than in a 30 minutes session, since the children have the possibility to produce more utterances" (Yamaguchi N, 2018, p6).

Having more time, children are expected to produce more utterances, so the word tokens should be more in longer sessions: this seems to be true (at least for the analysed *corpora*) if the child is younger than 1;10; while after this period *p-value* does not indicate significativity.

For word types, despite time is twice longer, there are no statistically significant differences between the two different sessions in all ages considered.

For sound types results are counterintuitive: the amount of different sounds produced by children is higher in short session rather than long ones[119]. While the number of target sound types resulted to be almost identical between short and long session, confirming in this way point 3. Regarding sound tokens, they are higher in long session and – differently from word tokens – they steadily increase over the ages while approximately keeping the same amount of difference in proportion between short and long session.

However, it is hard to draw conclusions from these results because the number of children taken in account is not high and it could be possible that the specificities of the language under study could bias final results. Additionally, SD bars overlap in almost every graphs, pointing out that a huge variability should prevent from any definitive conclusion.

Another important remark regarding the third prediction is the following

"[..] produced sound types do not obligatory correspond to phonemes of the target language, but to phones that the children produced" (Yamaguchi N., 2018).

In the data mining results of this thesis, this difference has not been taken into account in CHAID analysis because it relied on the difference between "pho" and "mod" tiers, while in results coming from histograms ("proportion phonétique) and Multistream graphs, a simplified and adapted list of phonetic units specific to French language has served as a reference to the analysis of everything a child said: this resulted in a filtering procedure where phones that were not in the list have been considered as a special case. In any case, transcriptions made for CoLaJE seem to not highlight these kinds of phones.

To conclude, the author is not claiming that a high density sampling is not better than a low-density one, she is underlining how the difference between short and long sessions is – in

[119] But, as Yamaguchi underlines, this depends on the phonetic transcription choice: a coarse-grained transcription with no diacritics and other phonic details can give very different results compared to a fine-grained transcription in which the slightest sound does distinguish a phoneme to another.

some cases - counterintuitive in the sense that there is not always a direct proportionality between duration and words/sounds' types/tokens quantities and, moreover, that the number of different situations/activities that happen during recordings are times more important than the simple duration of the session because they do trigger different reactions from the children involved in:

"[..] if dense corpora are used in the perspective of recording multiple activities and situations, the chances to record rare events, such as rare phonemes, rare combinations of phonemes, or rare words are multiplied, which could help obtain a fuller picture of child language development[120].

## 5.4 Considerations on CoLaJE sampling techniques

The question of the representativeness of a sample is relative to the scope of the research: a sparse sample can give good results if the target structure is highly frequent, but if the target is a rare phenomenon, the same sampling technique would become insufficient.

The same holds for the number of children taken into account: to establish the age of the emergence of a linguistic item, common linguistic outputs do require a relative small number of participants to be considered representative while focusing on linguistic specificities (*e.g* a third person plural form conjugation of a given verb in a conditional form) do require more children to be sure to draw from this sample a reliable generalization.

According to Yamaguchi:

"The level of linguistic investigation is decisive in the sampling of data: if one needs more corpora in order to observe morphological or syntactic events, a phonological or lexical

[120] Ibidem, pp 13-14

investigation could be performed on a smaller data sample[121]". (see below for quantitative results)

To avoid large random effects due to an insufficient sampling there is a strong need of samples that are as large as possible, but this goes in contrast to research constraints such as financial resources, time of transcription (around 30/45h to transcribe one hour of recording) and families' duties.

Referring to CHILDES average sampling of recordings, Deb Roy pointed out that:

"Most researchers rely on speech recordings that cover less than 1.5% of a child's complete linguistic experience[122]".

This sentence could sound dramatic if you realise this percentage during the third year of your thesis (as I did) but, if we relax a bit and master a little of inferential statistics, this sentence turns out to be quite ordinary.

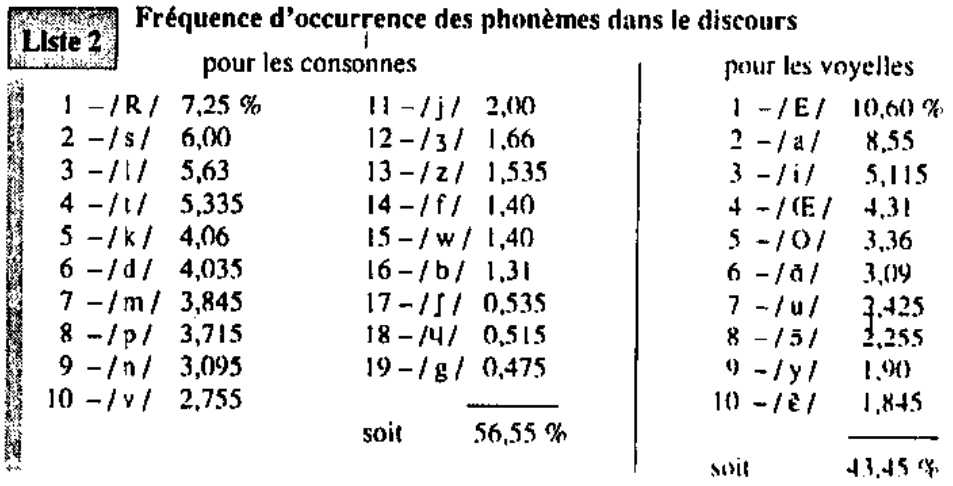
The point to be raised is the following: does CoLaJE *corpora* provide a reliable sampling of the total amount of what children hear and speak during development?

Supposing that a child is awake 10 hours per day, he then hears and speaks inputs from parents and environment for around 300 hours per month, CoLaJE samples one hour per month, then 1 hour out of 300 hours means less than 1% (approximately 0.33%) of the total amount.

A reflection on whether CoLaJE sampling would be enough for studying phonemes' acquisition is provided. In the following tables there are two examples of the frequencies of occurrence of French phonemes:

[121] Naomi Yamaguchi. (2018). "What is a representative language sample for word and sound acquisition?". Canadian Journal of Linguistics / Revue canadienne de linguistique, University of Toronto Press, 63 (04), pp.667-685. Page 3

[122] Roy D. 2006, p2



**Liste 2** **Fréquence d'occurrence des phonèmes dans le discours**

| pour les consonnes | | | | pour les voyelles | |
|---|---|---|---|---|---|
| 1 | –/R/ | 7,25 % | 11 –/j/ 2,00 | 1 –/E/ | 10,60 % |
| 2 | –/s/ | 6,00 | 12 –/ʒ/ 1,66 | 2 –/a/ | 8,55 |
| 3 | –/l/ | 5,63 | 13 –/z/ 1,535 | 3 –/i/ | 5,115 |
| 4 | –/t/ | 5,335 | 14 –/f/ 1,40 | 4 –/Œ/ | 4,31 |
| 5 | –/k/ | 4,06 | 15 –/w/ 1,40 | 5 –/O/ | 3,36 |
| 6 | –/d/ | 4,035 | 16 –/b/ 1,31 | 6 –/ɑ̃/ | 3,09 |
| 7 | –/m/ | 3,845 | 17 –/ʃ/ 0,535 | 7 –/u/ | 2,425 |
| 8 | –/p/ | 3,715 | 18 –/ɥ/ 0,515 | 8 –/ɔ/ | 2,255 |
| 9 | –/n/ | 3,095 | 19 –/g/ 0,475 | 9 –/y/ | 1,90 |
| 10 | –/v/ | 2,755 | | 10 –/ɛ̃/ | 1,845 |
| | | | soit 56,55 % | soit | 43,45 % |

Figure 12 : Consonants and vowels' relative frequency of occurrence in French [123]

Limit of this table is that these frequencies are calculated on adult language, that differs from child language. But – as we do not have better statistics – we could rely on these numbers that should logically be not so far from reality. After all, adult language (or child-directed speech) could be considered as a target structure that the child aims to imitate

Alternatively, a more precise and up-to-date statistical overview on French vowels and consonants is provided by professor Adda-Decker in the two following graphs:

[123] Taken from F. Wioland, 1991, p30

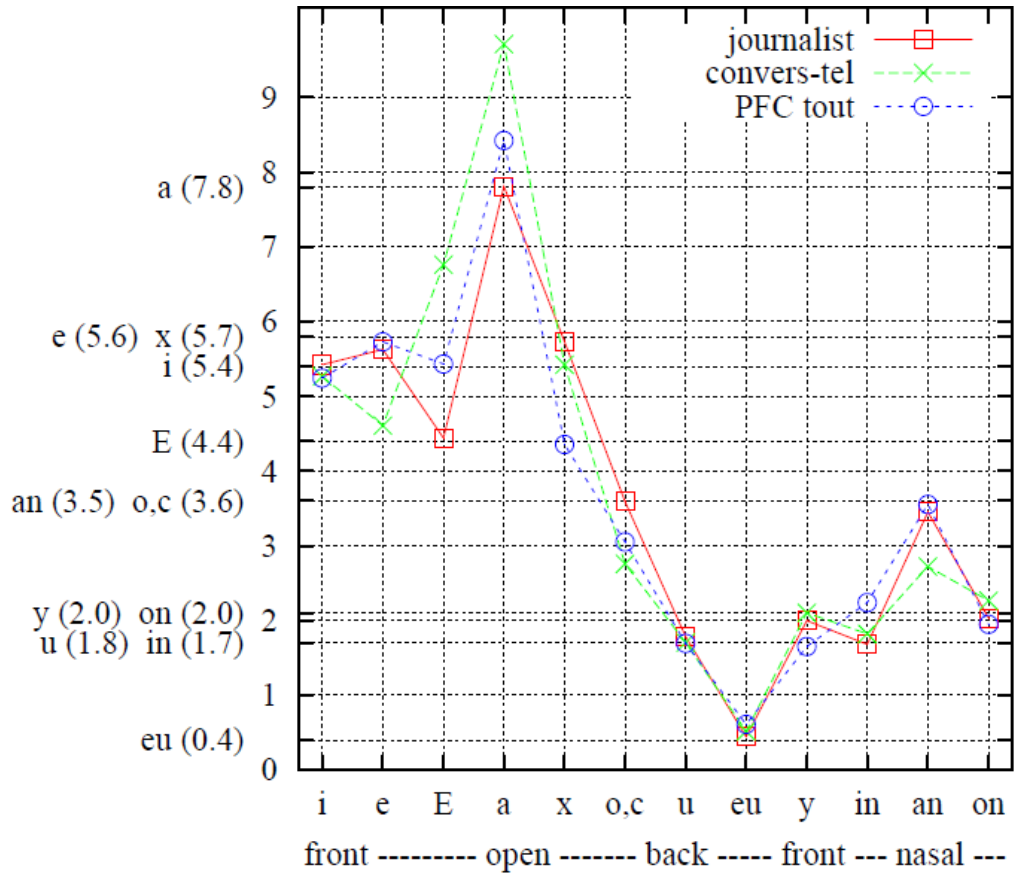



Figure 13 : Vowels' relative frequency of occurrence in French[124]

[124] Adda-Decker M. (200&). "De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux". Proceedings of JEP2006 - XXVIes Journées d'Étude sur la Parole, 12-16 juin 2006, Dinard (France). P883

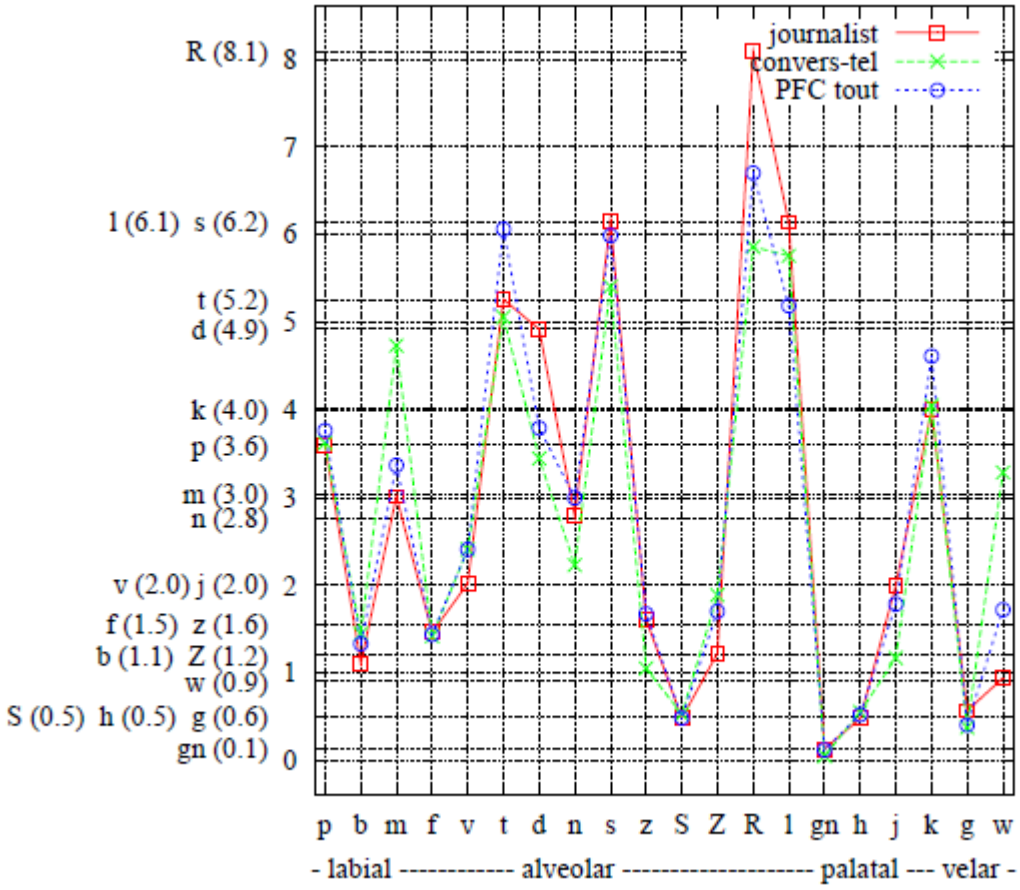



Figure 14 : Consonants' relative frequency of occurrence in French[125]

As it can be noted, values presented in these two graphs and the first table are quite similar between them, especially in the consonant/vowel rank that is almost the same.

Here below is an adapted formula (the original is on a weekly base. Stahl & Tomasello, 2004) based on a Poisson distribution (a statistical distribution used even in one of the next chapter to cluster grammatical categories) that would describe the capture rate of phonemes in CoLaJE:

"The Poisson distribution is a discrete distribution used to model the number of events occurring in some unit of time (or space), and it is mainly used if the occurrence of events is rare. It assumes that each event occurs independently of the others and at random" (Ibidem, p107)

[125] Ibidem, p883

$$\text{Lambda} = \frac{sample\ density\ (\frac{hours}{month})}{hours\ talking\ per\ month\ (i.e\ 300)} \text{ X number of target/ month}$$

Figure 15 the "capture rate" formula (Stahl & Tomasello, 2004)

The first table showed is based on a 200'000 phonemes sample that is thought to be representative by the author (Wioland, 1991, p30).

According to the first table, 8000 is the absolute frequence of « d » obtained by dividing the sample 200'000 by 0.04 (percentage of occurrence of « d »).

Mean word length is around 5 phonemes, 2000 is the mean number of words per hour of recording taken approximately from the article «The Paris corpus » (Morgenstern & Parisse, 2012). In this article, in Figure 3 we find the "Number of words per hour of recording according to age": so by multiplying 5 * 2000 we obtain 10'000 phonemes per each hour of recording record. Obviously this number varies dramatically from 1;0 (age around which CoLaJE recordings begin) to 4;2 (average age in which recording ends) as it is possible to see in the graph.

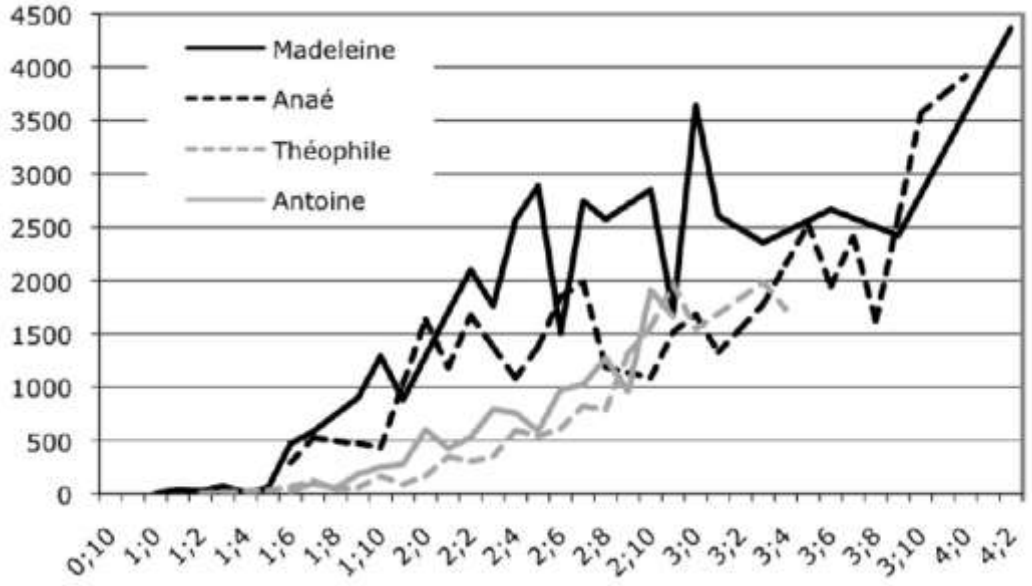



Figure 16: "Number of words per hour of recording according to age" (Morgenstern & Parisse, 2012)

We suppose that – in this specific case under investigation - 10'000 is as statistically representative as 200'000, so we calculate the 4/100 (0.4) part of 10'000, that is 400

Do these results are consistent with the data obtained from Python (See histograms at Chapter 11)

Finally, we can put these values in the formula explained before:

$$\chi = \frac{1}{300} * 400 = 13.3 \tag{7}$$

this means that – according to the data taken in account – it should be expected to have approximately thirteen “d” per hour of recording: obviously this value will vary according to the age, we would expect that in a median age – let’s say 2;6 – the value should be more or less this one. If we look at the histograms and at the Multiresolution Streamgraphs created for each of the six CoLaJE children https://marine27.github.io/TER/site_aquisition_du_langage/stackgraph.html , we could say that this value has been underestimated: “d” appears more than expected.

# Chapter 6 - Data cleaning, filtering and descriptive statistics

## 6.1 General considerations on data format and intepreting issues

First, we need a format ready to be recognised by algorithms like .csv and/or .xls, luckily CoLaJE project provided us a way to convert different formats between them at this link http://vheborto-ct3.inist.fr/teiconvert/index-en.html

To get a general overview, we thought that it would have been fair to know how a given child's pronunciation evolves over time, *a priori* from the specificity of variations he/she articulated (that of course is far from being a trivial question, but I will deal with morphology later).

We chose to employ the term «variation » instead of « mistake » because this seemingly slight meaning detail implies instead an important epistemological difference: children do not speak in an erroneous way, they rather speak differently from us. We must keep in mind that the interpretative prism of adult language is a major source of misunderstanding of child language, especially when dealing with varied forms of a target structure that we, as adults, are supposed to pronounce by using an established norm.

A child speaks differently because he/she has not still exposed to a sufficient amount of adult language and because his/her phonatory structure is still developing, and thus it does not enable him to produce certain sounds an adult can produce. So, it is both a matter of anatomy and physiology, of structure and dynamics, of experience independent and experience-dependent factors intertwined together.

Naming a difference «variation » instead of « mistake » reveals another approach toward child language: it has to be understood instead of being correct. That means that it is more important that we seek to understand the conditions that make a child pronounce a consonant instead of another instead of simply noticing that he is still not saying the same thing an adult would have pronounced and correct him by repeating him the « normed » form.

By following this perspective, every variation will be considered as a result issued by a temporary achieved structure that is evolving toward a more stable form directed by adult inputs. This process seems to have a need to be chaotic and to explore as much variations as it can during its internal self-organization.

Here is the same example taken before, but seen from another perspective: a « variation » will be pronounced in many cases without being aware of the difference it has to a target (supposing that children consciously see adult language as a learning target).

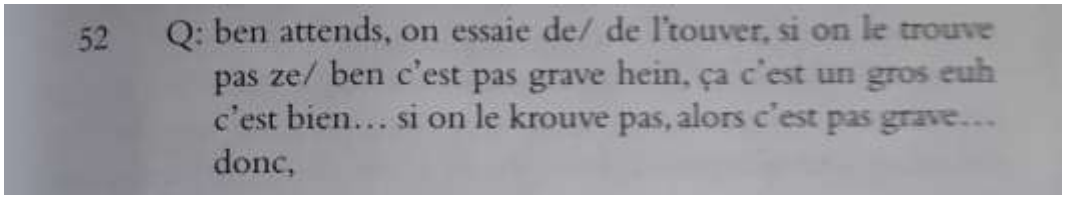

52 Q: ben attends, on essaie de/ de l'touver, si on le trouve pas ze/ ben c'est pas grave hein, ça c'est un gros euh c'est bien… si on le krouve pas, alors c'est pas grave… donc,

Figure 17 An extract [126]

In this short sentence, in a few seconds Albane pronounces a verb « trouver » (« to find » in English ») three times: the first is a varied form where he miss the « r » (a quite typical deletion that children do at his age when dealing with consonantal clusters), then he pronounces the verb in the correct conjugated form (third singular person) and finally he pronounces a different variated form by replacing « t » with « k ».

Do this child is aware of what he is saying?

Do « t » sounds like « k » in his internal phonological structure because of is common voiceless feature?

As far as we know, it is impossible to have certainties about his degree of consciousness on what he perceives both from adults and from himself. A lag between perception and production is often observed, so children usually can perceive more sounds than the number they can properly articulate.

[126] Sauvage, 2015, p 105

Related to the example above is the phenomenon of « regression » that is described by Sauvage as the absence of definitive linguistic structure during acquisition, meaning – in his words – that "tout acquis n'est jamais définitif [127]".

This statement relates to the acquisition period, from birth until approximately six year-old, but I think it could apply also to L2 adult learners (at least to my case!).

It is difficult to model this phenomenon because it is counter intuitive: once a given linguistic structure is properly articulated and in the right syllabic or syntactic position, we would be brought to think that it has been learnt once for all.

Beyond this simplistic view, regressions, as well as multiple variations of a same target (as « touver » and « krouver » in place of « trouver » in the above example) hide exactly what would seem more interesting and informative about the acquisition process : a non linear dynamics where different factors contribute in different proportions at different ages in different ways through different paths, thus giving an unpredictable overall process that – despite all – will ends in a comparable result.

While in statistics a regression describes how a variable can predict the behaviour of another variable, it is a statistical technique aimed at providing a way through which variables influence each other. There are many types of regression (linear, nonlinear, non parametric..)

## 6.2 List of softwares used in this thesis

In this chapter descriptive statistics have been calculated by using SPSS ver. 25, STATA ver. 15 and Microsoft Excel 2010

For CHAID we used SPSS ver. 25 and STATA ver. 15

For EM clustering we used "STATISTICA" Statsoft ver. 10, for ANOVA results we used R ver 3.5.3

[127] Sauvage, 2015.

To transform .csv data exported from CoLaJE's website into the formats showed in tables provided in "Supplementary files" we used several routines in Python language

While in the final Chapter "Data Mining" we used Python ver. 3.8 (the latest on current date)

## 6.3 Data export and first descriptive statistics

As a first and exploratory analysis to get knowledge from a huge amount of data that other researchers have collected, we chose to export them in *.xls* format in order to get from them an overview made up of descriptive statistics, some simple visual representations and a closer look to how phonetic variation rate (from now PVR) evolves along CoLaJE monthly records

We chose to analyse four later age *corpora* of Adrien. The choice was on this child because its transcription are the most completed ones, especially because in every record is provided a division between pho and mod that in other CoLaJE children is missing. We decided to start from 3 years old because in later ages transcriptions are more readable (as child language is closer to the adult one, the transcriber have to make simpler interpretations).

The four records are: number 22 (3_01_13, numbers stand for 3 years, one month, thirteen days), number 24 (3_03_12), 27 (3_08_05) and 34 (4_03_26).

After having exported the datasets, we divide it in two ways: rows represent, alternately, « pho » and « mod » values while every column represents a word in its order of appearance (first, second, third etc.) until 20, which was the longest sentence found. Subsequently we create the same number of columns for variations, using 0 to represent a correct form and 1 to represent a variation.

Here is an example of the Excel spreadsheet :

Figure 18 : A raw data structure in a spreadsheet (see supplementary file for a better image resolution)

This raw data structure allows us to calculate some descriptive statistics such as mean, variance and Standard Deviation (SD). The following tables are taken from the last record here analysed:

| Summary | tot_var | n_words | perc_err |
|---|---|---|---|
| | 2794 | 5772 | 48,41% |
| **Mean** | 2,99 | 6,18 | 0,48 |
| **Variance** | 13,45 | 19,39 | |
| **S.D** | 3,66 | 4,40 | |
| **C.V.** | 1,22 | | |
| **Min** | 0 | 1 | |
| **Max** | 20 | 20 | |
| **Count** | 673 | | |
| | | | |
| **perc_err (weighted)** | | | |
| **N** | Valid | 4029 | |
| | Missing | 0 | |
| **Mean** | | **27,97%** | |
| **Median** | | 25,00% | |
| **Mode** | | 0,00% | |
| **Std. Deviation** | | 24,47% | |
| **Minimum** | | 0,00% | |
| **Maximum** | | 100,00% | |
| **Percentiles** | 25 | 11,00% | |
| | 50 | 25,00% | |
| | 75 | 38,00% | |

Table 1 Descriptive statistics

By comparing the four spreadsheets, we can observe that the word count globally increases over time : 386, 336, 591 , 673 that could be interpreted as an element of regular development, but a more fine-grained analyses should take into account an index of lexical richness such as type/token ratio (see CLAN list of commands) that gives the amount of different words in relation to the total amount of words in a given corpus.

We can observe how phonetic variation rate decreases as was expected to do, despite the fact that there is no direct proportionality between time and variation rate: 0.62, 0.56, 0.49, 0.48 (the same holds for the word count too): for instance, the last interval is the wider one (7

months compared to the previous 2 or 3, but it results in a minimal variation that could be simply due to chance).

Standard deviation shows a more complex behaviour, 1.7 , 1.4, 1.4, 3.6 respectively. This sharp increase may be due to the emergence of complex sentences at 4 years old, as it is showed in this graph describing the number of word types (a similar value to the type token ratio described above) per hour of recording according to age.

These results are coherent to similar studies conducted on the same set of *corpora*, as shown in the graph:

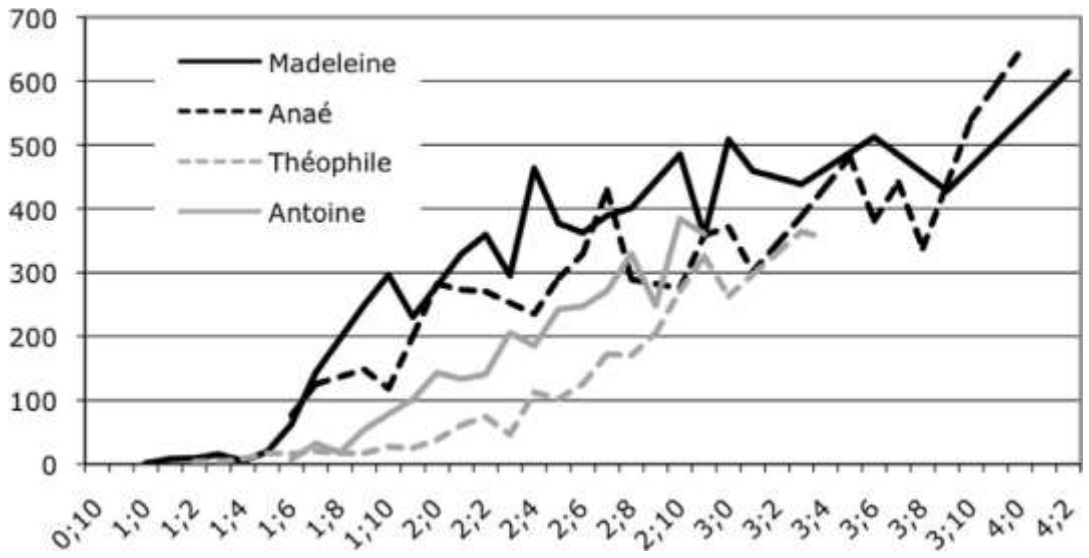



Figure 18 Number of word types per hour of recording according to age (Morgenstern & Parisse, 2012)

It is clear how the two children who have been recorded until four show an increase in lexical variability at four. This could imply that, counterintuitively, standard deviation increase over time would be in this case a result of an improvement instead of a loss of ordered data. This because as lexical variability increases, the range of possibility within which variations can occur increases too in a directly related way: if my word repertory is around 300, the number of possible different variations I can pronounce will not be so much higher, while if my word has raised to 500, despite an increase in my general competence of my mother tongue, the range of possible variations has approximately doubled.

To conclude, a graph that summarizes what has been written until now on the four Adrien *corpora* analysed.

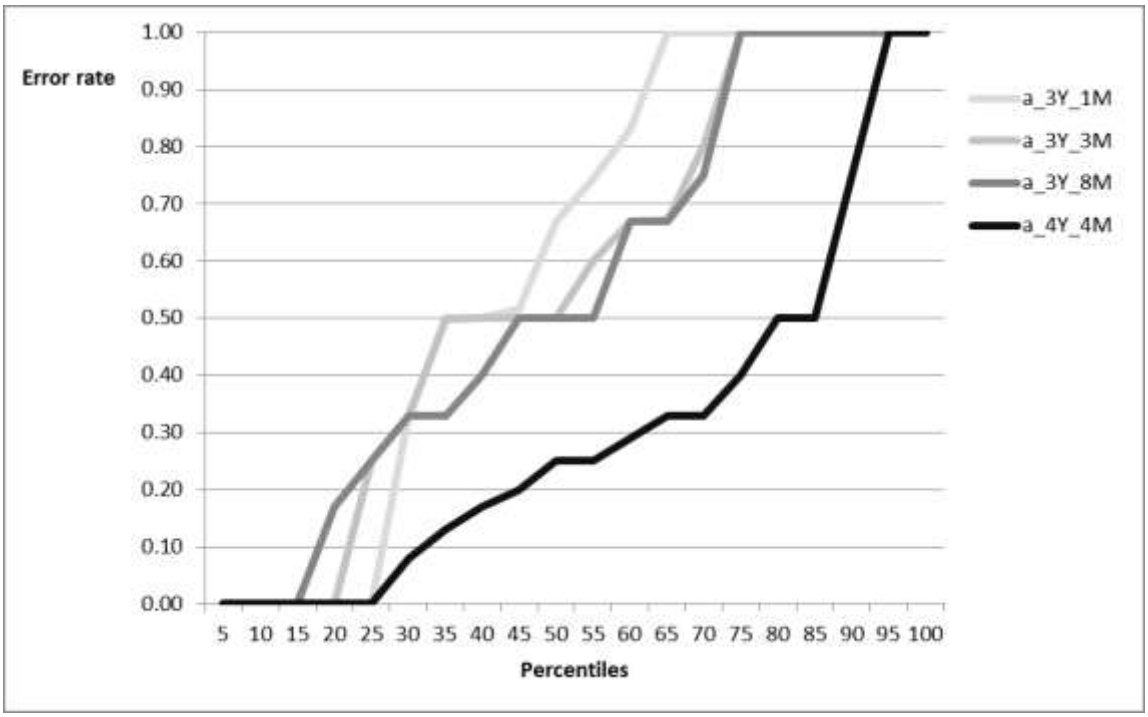



Figure 19 Variation rate. Comparison between ages

We can clearly see how, if we imagine to draw a median line on the value 50 on the *x* axis, we can easily observe an improvement between the consecutive plots shaded by a different grey. The lighter one (corpus no. 22) has its median at 0.70, while in the black one, when Adrien is four years old, variation rate crosses the median line in a value that is half the previous mentioned. A similar evolution can be spotted at the bottom of the *x* axis.

These results are similar to an analogous previous study done by Sauvage[128] .

The number of variations globally decreases over time, but what happens in the middle are counterintuitive phenomena such as regressions and stationary periods that need a special focus because they could reveal a dynamic of an ongoing process of building of successive temporary achieved phonological consciousness. To get into a more detailed view, let's see

[128] Sauvage, 2015, p102

how variation rate relates to sentences' length. It is hard to make predictions or to have hypotheses on it, because many factors are at work in determine how a child can properly pronounce a sentence made up of one or more words : the effort it takes, the grammatical complexity of the sentence, the morphological complexity of the words involved, whether this sentence is an answer to an adult question or it derives from a child monologue and so on.

Supplementary file. Number 27, 3 years 8 months 5 days

In this spreadsheet (see attached table for further comparisons between ages and for other descriptive statistics values) calculated from the raw Adrien's data, we can observe the frequencies of words pronounced in successive orders along the sentences : the first, the second and so on until the tenth (actually, there are few sentences that span until the twentieth, but as these are less than 10, we judged impossible to infer whatever from them).

**err_1**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| **Valid** | 0 | 308 | 52,1 | 52,1 | 52,1 |
| | 1 | 283 | 47,9 | 47,9 | 100,0 |
| | Total | 591 | 100,0 | 100,0 | |

**err_2**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| **Valid** | 0 | 215 | 36,4 | 49,2 | 49,2 |
| | 1 | 222 | 37,6 | 50,8 | 100,0 |
| | Total | 437 | 73,9 | 100,0 | |
| **Missing** | System | 154 | 26,1 | | |
| **Total** | | 591 | 100,0 | | |

**err_3**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| **Valid** | 0 | 170 | 28,8 | 51,8 | 51,8 |
| | 1 | 158 | 26,7 | 48,2 | 100,0 |
| | Total | 328 | 55,5 | 100,0 | |
| **Missing** | System | 263 | 44,5 | | |
| **Total** | | 591 | 100,0 | | |

**err_4**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| **Valid** | 0 | 119 | 20,1 | 49,6 | 49,6 |
| | 1 | 121 | 20,5 | 50,4 | 100,0 |
| | Total | 240 | 40,6 | 100,0 | |
| **Missing** | System | 351 | 59,4 | | |
| **Total** | | 591 | 100,0 | | |

**err_5**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| **Valid** | 0 | 78 | 13,2 | 51,3 | 51,3 |
| | 1 | 74 | 12,5 | 48,7 | 100,0 |
| | Total | 152 | 25,7 | 100,0 | |
| **Missing** | System | 439 | 74,3 | | |
| **Total** | | 591 | 100,0 | | |

**err_6**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| **Valid** | 0 | 42 | 7,1 | 41,6 | 41,6 |
| | 1 | 59 | 10,0 | 58,4 | 100,0 |
| | Total | 101 | 17,1 | 100,0 | |
| **Missing** | System | 490 | 82,9 | | |
| **Total** | | 591 | 100,0 | | |

**err_7**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| **Valid** | 0 | 27 | 4,6 | 46,6 | 46,6 |
| | 1 | 31 | 5,2 | 53,4 | 100,0 |
| | Total | 58 | 9,8 | 100,0 | |
| **Missing** | System | 533 | 90,2 | | |
| **Total** | | 591 | 100,0 | | |

**err_8**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| **Valid** | 0 | 18 | 3,0 | 62,1 | 62,1 |
| | 1 | 11 | 1,9 | 37,9 | 100,0 |
| | Total | 29 | 4,9 | 100,0 | |
| **Missing** | System | 562 | 95,1 | | |
| **Total** | | 591 | 100,0 | | |

| err_9 | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
| **Valid** | 0 | 10 | 1,7 | 47,6 | 47,6 |
| | 1 | 11 | 1,9 | 52,4 | 100,0 |
| | Total | 21 | 3,6 | 100,0 | |
| **Missing** | System | 570 | 96,4 | | |
| **Total** | | 591 | 100,0 | | |

| err_10 | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
| **Valid** | 0 | 7 | 1,2 | 43,8 | 43,8 |
| | 1 | 9 | 1,5 | 56,3 | 100,0 |
| | Total | 16 | 2,7 | 100,0 | |
| **Missing** | System | 575 | 97,3 | | |
| **Total** | | 591 | 100,0 | | |

Table 2 : Descriptive statistics based on variation rate calculated from the raw Adrien's data set

First, we want to be sure that our calculations were in line with similar calculations applied to other CoLaJE children, thus we compare the results with each other :

Adrien 386 (coordinates C 731 spreadsheet 5) and following. As we can observe in the graph below, 386 is near to the values of the other two boys (Théophile and Antoine). .

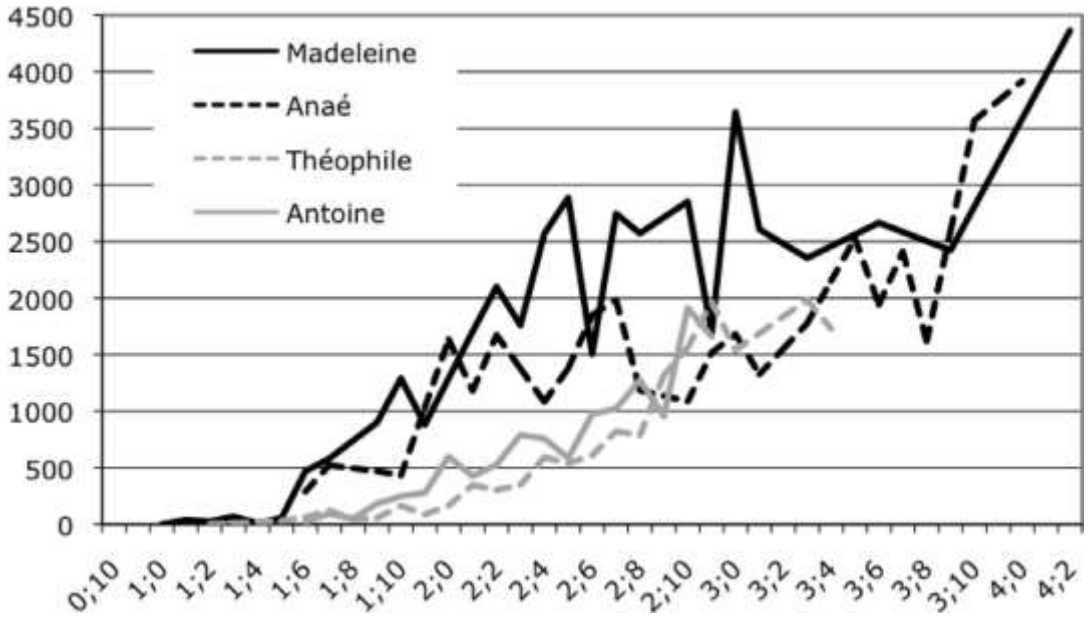



Figure 20 : Number of word per hour of recording according to age (Morgenstern & Parisse, 2012)

From the previous extract from Excel spreadsheet, we can observe that it is difficult to find a relation between the length of a sentence and the variation rate associated with its constituent words : there is only a slight tendency toward an increase of the 1 value (1 is variation 0 is the correct form), but a sequence of ascents and descents impede every possible generalization from these values.

While if we look at the « younger » values (I'm referring to stat_22 and stat_24 in the spreadsheet attached), it is clear how there is any sequence of ascents and descents, but there is a rather steady increase in 1 value, that would bring us to say that – at the least until three years and three months age, the length of a sentence is a predictor of its variation rate (as one increases, the other increases too).

While in the last record analysed (stat_34), we found a more stable relationship between sentences' length and words variation rate : this would bring us to the conclusion that at four years old this child has no particular problem in uttering longer than usual sentences, thus length does not influence variation rate.

But, after all, correlation is not causation : there are many other factors that we did not take into account such as morphology, morphosyntactically dependent suffixes, child-directed speech (corrections from the adult). We will provide a syntactically-informed parsing in the next chapter (Universal Dependencies)

We found these data consistent with state-of-the-art literature such as this graph provided by Morgenstern and Parisse, former heads of the project CoLaJE (please note that despite Adrien is not represented in this graph, as well as in the previous one, these children have been videorecorded and transcripted by using the same protocol, thus results are supposed to be comparable).

# Chapter 7 - The CHI-squared Automatic Interaction Detection : an application on SPVR

## 7.1 An overview

As I discussed in the previous paragraph, inferential statistics is a fundamental step toward a significative[129] study on first language acquisition: as it is impossible to draw conclusions based on everything aa childn infant says for obvious reasons (time and money), it becomes necessary to sample in a smart way (see Tomasello, 2004) to catch every phenomenon we want to focus on at least once in every record we make, as well as to be able to infer from this limited sample as much information as we can in a reliable way.

This inductive process is at the core of statistical hypothesis testing and represents the basis of the coherence of every comparison between children. Beyond this approach, our idea is essentially to find out how the response variable is related to potential predictors. We therefore move towards a decision tree technique.

CHAID[130] is a decision tree technique conceived to overcome in a non-parametric way (*i.e.*, there are no formal theoretical assumptions to meet) the problem of multiple comparisons.

In particular, the CHAID algorithm consists of three stages: merging, splitting and stopping (Magidson, 1993; Ratner, 2017):

[129]We recall that, the meaning of « statistically significant » rely on a threshold probability decided on the basis of several factors, often related to the specific domain.

[130] Kass, G.V. (1980) "An Exploratory Technique for Investigating Large Quantities of Categorical Data". App. Statist 29(2):119-127

- The first step (merging) consists in pooling similar classes of the predicting variables. The criterion for similarity is the lack of significant differences between every two classes of a predictor.
- In the second step (splitting), the best predictor and four competitors for dividing the root node and subnodes are chosen.

The CHAID tree grown until either only one object remains in the subnodes or until user-specified restrictions are met (stopping rules). The restrictions can be specified in terms of minimum segment size, significance level used in merging and splitting and depth limit (*i.e.* the number of levels of the CHAID tree). The significance level of splitting defines whether the statistical association between the target variable and the predictors is sufficient to perform a split or not.

In the classic approach, CHAID is based on $\chi^2$ (chi squared) test, while if the dependent variable is continuous, the F (Fisher) test is used.

By testing how a supposed dependent variable (phonetic variation) is dependent to an independent variable (time + utterance's length) the algorithm iteratively forms subsequent smaller sub-groups.

Limit of this method is that it does not take into account phonetic and morphological differences between phonetic units (e.g a bilabial from an plosive-liquid, see Annex 6 for more details on articulatory difficulties): in order get over this problem we choose to evaluate its validity by interpreting its results through the lenses of a "consonant acquisition chart" (4, se annexes) and some considerations on first language acquisition specific to French[131].

[131]Parisse et al., 2012

## 7.2 The method at work

Based upon an iterative procedure, for any dependent variable CHAID evaluates every possible supposed predictive variable by relying on the more appropriate significativity test. The most significantly independent variables are used to split sample in subgroups. Consequently, every subgroup is analyzed to individuate a further predictive variable that could further split a given subgroup.

Two main reasons make a subgroup not further splittable: there are no more significant predictive variables, so-called « stop rules » set by the user are met. These rules are defined by the minimum number of units a subgroup can be made of and the minimum number of subgroups that can be formed by subsequent splitting procedure (as it is set by the user too).

"CHAID proceeds in steps: first the best partition for each predictor is found. Then the predictors are compared and the best one chosen. The data are subdivided according to this chosen predictor. Each of these subgroups are re-analysed independently, to produce further subdivisions for analysis. The type of each predictor determines the permissible groupings of its categories, so as to build the contingency table with the highest significance level according to the chi-squared test.[..] This implies that there are enough observations to ensure the validity of this test" (Kass G., 1980, p2)

We may call a final group a « segment » because of its mutually exclusive nature

Difference between most significant predictors and most explanatory predictors, CHAID improves the second ones.

How the algorithm recognizes and categorizes words in the corpus ?

« CHAID partitions the data into mutually exclusive, exhaustive, subsets that best describes the dependent variable, it operates on a nominal scaled dependent variable and maximizes the significance of a chi-squared statistic at each partition, which need to be a bisection[132] »

Moreover, when the target variable is continuous, such as the case of Sentence Phonetic Variation Rate (from now SPVR) and predictors are categoricals or continuos, such as for

[132] Kass, G.V. (1980) "An Exploratory Technique for Investigating Large Quantities of Categorical Data". App. Statist 29(2):119-127. P120

Part Of Speech Tags (from now, POS Tags), the statistical test of reference is Fisher (from now F) and not Chi Squared. The significance value for splitting nodes and merging categories is set to 0.05 (also known as p-value). For multiple comparisons, significance values for merging and splitting criteria are adjusted using the Bonferroni method.

CHAID is a flexible method: by default the decision tree diagram is ordered by the value of Fisher (F-statistics) in a top-down way, but -when the researcher has a preconceived description of the data, it is possible to set a given variable as the first splitting variable instead of Fisher: is what we have done in the "CHAID forced-time" (see supplementary file) when we set time as a first factor.

This consisted in creating four time periods of nine months each: by doing so, we allowed CHAID to focus its analyses on variation rate (SPVR) in specific time periods raising in this way the accuracy of its results and, at the same time, we obtained a more easy-to-read decision tree. We did so because we suppose that time (*i.e* ages) is the most important factor that influences phonetic variation.

## 7.3- CHAID applied on Adrien

First we did the analyses on pho & mod and then on POS tags. We first analysed 27 recordings of a single child named "Adrien". 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 3 are summarized the main statistics for 27 recordings: 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.

| Time | Mean | Length | S.D. |
|---|---|---|---|
| 1.97 | 1.59 | 17 | 1.064 |
| 2.04 | 1.02 | 63 | .126 |
| 2.12 | 1.11 | 183 | .362 |
| 2.17 | 1.59 | 41 | 1.224 |
| 2.23 | 1.37 | 251 | .836 |
| 2.33 | 1.67 | 250 | 1.118 |
| 2.41 | 1.72 | 316 | 1.210 |
| 2.48 | 1.78 | 376 | 1.193 |
| 2.64 | 1.98 | 212 | 1.579 |
| 2.71 | 1.71 | 319 | 1.049 |
| 2.80 | 1.72 | 169 | 1.023 |
| 2.89 | 2.00 | 283 | 1.302 |
| 2.96 | 2.07 | 184 | 1.430 |
| 3.04 | 2.37 | 465 | 1.353 |
| 3.12 | 2.61 | 324 | 1.764 |
| 3.20 | 2.92 | 433 | 1.893 |
| 3.29 | 2.44 | 240 | 1.389 |
| 3.38 | 2.46 | 196 | 1.729 |
| 3.46 | 2.96 | 330 | 1.708 |
| 3.69 | 3.30 | 517 | 2.346 |
| 3.79 | 3.15 | 310 | 2.321 |
| 3.88 | 3.76 | 324 | 2.565 |
| 3.97 | 4.70 | 396 | 3.094 |
| 4.04 | 4.39 | 584 | 3.190 |
| 4.12 | 3.59 | 334 | 2.602 |
| 4.21 | 3.73 | 473 | 2.656 |
| 4.33 | 6.11 | 624 | 4.548 |
| Total | 3.03 | 8214 | 2.649 |

*Time= age (sessions); Mean = average number of words per sentence; Length = number of sentences in a given session; S.D.= standard deviation of the number of words per sentence*

Table 3 : Corpus statistics (full database)

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 29 (max number of words in a single sentence.) From table 2 we can see how nonlinearity affects language acquisition: globally, SPVR decreases over time but counterintuitive phenomena such as regressions (Sauvage, 2015) are frequent: it could happen that a child mispronounces something that he had previously correctly pronounced. The same holds for SPVR over sentence’s length: we expect (and observe) that rate increase as the length increases, but there are some exceptions to the norm that could require a specific account.

| **Time (year)** | | **SPVR** | **Levenshtein_distance** | **feature_edit_distance** | **weighted_feature_edit_distance** |
|---|---|---|---|---|---|
| 1.97 | Mean | 94,12 | 3,18 | 0,49 | 3,24 |
| | N | 17 | 17 | 17 | 17 |
| 2.04 | Mean | 83,33 | 1,71 | 0,56 | 4,23 |
| | N | 63 | 63 | 63 | 63 |
| 2.12 | Mean | 57,10 | 1,40 | 0,52 | 3,87 |
| | N | 183 | 183 | 183 | 183 |
| 2.17 | Mean | 85,37 | 2,39 | 1,36 | 10,86 |
| | N | 41 | 41 | 41 | 41 |
| 2.23 | Mean | 79,68 | 2,90 | 1,02 | 7,60 |
| | N | 251 | 251 | 251 | 251 |
| 2.33 | Mean | 56,24 | 1,71 | 0,64 | 4,99 |
| | N | 250 | 250 | 250 | 250 |
| 2.41 | Mean | 43,15 | 1,50 | 0,44 | 3,21 |
| | N | 316 | 316 | 316 | 316 |
| 2.48 | Mean | 49,87 | 2,19 | 0,78 | 5,91 |
| | N | 376 | 376 | 376 | 376 |
| 2.64 | Mean | 54,85 | 2,65 | 1,06 | 8,16 |
| | N | 212 | 212 | 212 | 212 |
| 2.71 | Mean | 64,22 | 2,20 | 0,79 | 6,04 |
| | N | 319 | 319 | 319 | 319 |

| | | | | | |
|---|---|---|---|---|---|
| 2.80 | Mean | 53,78 | 2,07 | 0,84 | 6,50 |
| | N | 169 | 169 | 169 | 169 |
| 2.89 | Mean | 50,12 | 2,08 | 0,70 | 5,26 |
| | N | 283 | 283 | 283 | 283 |
| 2.96 | Mean | 54,17 | 1,99 | 0,61 | 4,47 |
| | N | 184 | 184 | 184 | 184 |
| 3.04 | Mean | 52,57 | 2,22 | 0,70 | 5,27 |
| | N | 465 | 465 | 465 | 465 |
| 3.12 | Mean | 48,04 | 2,60 | 0,91 | 6,94 |
| | N | 324 | 324 | 324 | 324 |
| 3.20 | Mean | 52,75 | 3,15 | 1,02 | 7,63 |
| | N | 433 | 433 | 433 | 433 |
| 3.29 | Mean | 45,64 | 1,95 | 0,62 | 4,77 |
| | N | 240 | 240 | 240 | 240 |
| 3.38 | Mean | 40,00 | 1,71 | 0,66 | 4,95 |
| | N | 196 | 196 | 196 | 196 |
| 3.46 | Mean | 52,48 | 2,74 | 0,87 | 6,48 |
| | N | 330 | 330 | 330 | 330 |
| 3.69 | Mean | 48,24 | 2,62 | 0,84 | 6,23 |
| | N | 517 | 517 | 517 | 517 |
| 3.79 | Mean | 44,77 | 1,97 | 0,75 | 5,65 |
| | N | 310 | 310 | 310 | 310 |
| 3.88 | Mean | 37,53 | 2,15 | 0,83 | 6,29 |
| | N | 324 | 324 | 324 | 324 |
| 3.97 | Mean | 41,01 | 2,96 | 0,97 | 7,16 |
| | N | 396 | 396 | 396 | 396 |
| 4.04 | Mean | 39,65 | 2,58 | 0,85 | 6,36 |
| | N | 584 | 584 | 584 | 584 |
| 4.12 | Mean | 30,97 | 1,68 | 0,45 | 3,25 |
| | N | 334 | 334 | 334 | 334 |
| 4.21 | Mean | 32,06 | 1,75 | 0,53 | 3,90 |
| | N | 473 | 473 | 473 | 473 |
| 4.33 | Mean | 23,99 | 1,88 | 0,61 | 4,47 |
| | N | 624 | 624 | 624 | 624 |
| Total | Mean | 46,56 | 2,24 | 0,75 | 5,66 |
| | N | 8214 | 8214 | 8214 | 8214 |

Table 4 : Main statistics (indexes) of language development over time (full database)

## 7.4. Cleaning data toward CHAID

As far as we know, CHAID has never been used for language analysis in general and first language acquisition in particular

As CHAID is a statistical technique conceived for domain-general purposes especially in the field of economics and demography, its classification and clusterization schemas need to be cautiously applied in other fields far from the previous ones, as it is the case of linguistics.

To give an example that initially biased CHAID procedure, we could spend a few words on the coincidence in the number of « pho » and « mod » lines (tiers in the CoLaJE project jargon). In fact, CHAID – in order to be able to apply its algorithms – need a perfect correspondence in terms of numbers between « pho » and « mod », otherwise it will give results that we may define « out of phase ».

Transcribing oral language is in general a difficult task : speaking is not the same verbal process as writing, there are a number of phenomena such as repetitions, hesitations, reformulations, pauses etc that appear while chatting and disappear while we put black on white our thoughts. Of course, when dealing with child language, these difficulties are multiplicated because toddlers do not follow a norm, they are more creative than adults in uttering whatever it comes to their minds and their attention is flawed by many novelties and inputs coming at the same time when they are recorded.

So, as « pho » is what the child says and « mod » what he should have said according to the adult norm, these two lines are sometimes made up of a different number of words due to different choices in transforming utterances to written texts, especially when words are partially pronounced or expressed in a varied form, or when children use in a « creative way » grammar rules such as apostrophes and French *liaisons.*

Here some examples taken from datasets we exported

| | | | | |
|---|---|---|---|---|
| 1826 | ADRIEN | 2_07_18 | nɔ̃ mamɑ̃ te ɕoɕi mamɑ̃ | nɔ̃ mamɑ̃ ʃosyʁ mamɑ̃ |

Table 5 - Example taken from the dataset

The first text is « pho », the following « mod », the sentence in english would sound like « no, mummy, shoes, mummy » with something similar as a pronoun in the middle if the sentence, that is quite hard – even by watching the video http://modyco.inist.fr/tools/trjsread/trjsread.html?t=/data3/colaje/adrien/ADRIEN-16-2_07_18/ADRIEN-16-2_07_18.tei_corpo.xml to establish to whom he is referring to, whether to himself, to his mother or to the observer.

So, in this case the transcriber decided to transcribe "ɛoɛi" in "ʃosyʁ" (in standard orthographic would be "chaussures", in english "shoes") because she can directly see this object while filming: if the child would have said the same term referring to it in an abstract way, where the object was not there, it would have been more difficult to interpret it being sure that was exactly what the child had in mind. Phonetically speaking, the two sounds (the variated form and the correct one) share some similarity at the onset, but the end of the word seems, at least to me, quite different.

Here "pho" is composed by 5 words while "mod" is composed by only four. The mysterious element « te » supposed to be a filler, had been judged by the transcriber to be syntactically incoherent to what she thoought to be the correct form an adult would have uttered in the same situation.

What can CHAID automatically detect from these two lines ?

That « pho » equals « mod » in the first word : and we could qualitatively deduct that this child can pronounce this specific nasal vowel « ɔ̃ » correctly (see IPA chart at the bottom of the thesis for further details). It holds the same for the second word, and if we want to be a little optimistic, we could draw from these two correct words that Adrien has a good mastery of the nasal trait (see Yamaguchi on chapter 4 and tables in Annexes). But when it comes to the third word, « te », the software will compare it to the third word in mod, thus giving incorrectly an incorrect outcome, then the fourth word will be compared to the fourth on mod « mamɑ̃ », thus resulting in an another negative response, and finally the fifth « pho » word would be compared to a blank case, that will result in an « error » too.

Considering this difficulty in matching on one hand the authenticity of the trancripts and the transcriber's choices and, on the other hand, the recognition and computational related software constraints, we were stuck for some days in a sort of trade-off choices.

Regarding the final result : which bias will influence more our outcomes ?

The fact of keeping sentences that will count something that is not related to the pho-mod relationship or removing *tout court* something that the child had said, after all, half correctly ?

In other words : it is better to have a whole corpus but flawed with « out of phase » data or is it better to have a smaller corpus with all the data meeting the software's needs ?

It depends on the amount of this discrepancies between pho and mod :

| Typology | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| PHO=MOD | 7812 | 95,1 | 95,1 | 95,1 |
| PHO<MOD | 172 | 2,1 | 2,1 | 97,2 |
| PHO>MOD | 230 | 2,8 | 2,8 | 100,0 |
| Total | 8214 | 100,0 | 100,0 | |

Table 6 : Corpus by tipology

As we can read from the above table, sentences where the equivalence between « pho » and « mod » is not met are less than 5% of the total amount of transcripted lines, so we thought that it was fairer to avoid biases of any kind on the decision tree outcomes and we thus put aside these pho> mod and pho<mod.

See the supplementary file « Adrien_results_20_4_2020 »

We used CHAID to get a general insight on how SPVR[133] changes over time and which kind of phonetic units are correctly articulated and which are not. From the results obtained[134], we can clearly see how time is the main regressor because it splits most part of the *corpus*, then the length of sentences 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

[133] Briglia A.: Mucciardi M. Sauvage J. (2020). Identifying the speech code through statistics. A data driven approach". Proceedings SIS. Book of short papers. For the concept of SPVR in detail

[134] All statistical analyses were performed using R, Excel and SPSS. In the CHAID model, cases are weighted by TNW. Furthermore, due to lack of space, main statistics and tree diagram are provided in a supplementary file.

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.

| Node | PVR_20 (Mean) | N | Primary Independent Variable | p-value | Split values |
|---|---|---|---|---|---|
| 15 | **0.971** | 68 | w_mod_1r | 0.000 | ãkɔʁ; sɛlsi; spidœʁma; isi; vjɛ̃; pʁefɛʁ; boku; bɔ̃ʒuʁ; vwatyʁ; vɛʁt; kaʁgo; osito; ɛskaʁgo; pjɛʁo; by; tʁwa; katʁ; sɛ̃k; sis; sɛt; õz; 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; ãkɔʁ; lwi; sɛlsi; spidœʁma; akʁoʃe; otuʁ; sali; tɔ̃be; uvʁ; dɛʁjɛʁ; pɔʁt; isi; sɥisi; alɔʁ; ã; 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; komã; dø; ba; duz; tʁɛz; katɔʁz; dã; 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; mamã; sa; lə; akemi; dɔn, |
| 27 | **0.025** | 119 | w_mod_2r | 0.000 | nɔ̃; le; papa; lə; isi; bys; ʒoli, |

Table 7: Tree table for CHAID model (main results - first and last three PVR_20 values)

while in the node 11 (PVR_20 mean 0.267 – not shown) words are shorter and contains 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 calculate interactions between occurrences

The problems is that CHAID does not tell us where exactly the variation is, because it is blind to phonemes’ specificities

Most frequent words are better pronounced than less frequent (give numbers, CLAN), consonants and age (order of acquisition)

This method could be viewed as a simple and quick way to get an overview of different words (and thus phonemes) learned differently at different times. But we do not think it can provide a detailed account of language acquisition

## 7.5 CHAID applied to the whole Adrien corpus

From now onward I am referring to the decision tree graph named “CHAID_SPVR (total sample)” that is one of the graphs available on the spreadsheet named “adrien_results_20_04_20” (attached).

So, as we can read from “Node 0” the total amount of tiers (lines) found in twenty-nine transcripted records from Adrien is n=8214 and mean equals 46. The mean itself, at this point of analysis, is unuseful: it will become useful when comparisons between ages will be made.

As we can see from the following table (available in the same document, at the same page, at the bottom):

| Node | N | Percent | Mean |
|---|---|---|---|
| 6 | 430 | 5,2% | 99,1512 |
| 46 | 65 | 0,8% | 96,1538 |
| 35 | 197 | 2,4% | 93,9292 |
| 47 | 140 | 1,7% | 84,8958 |
| 14 | 251 | 3,1% | 83,4795 |
| 27 | 160 | 1,9% | 80,3126 |
| 15 | 185 | 2,3% | 75,7023 |
| 36 | 62 | 0,8% | 74,1935 |
| 42 | 187 | 2,3% | 69,4743 |
| 28 | 183 | 2,2% | 68,8809 |
| 43 | 99 | 1,2% | 63,2813 |
| 48 | 121 | 1,5% | 60,7617 |
| 21 | 130 | 1,6% | 58,8462 |
| 37 | 291 | 3,5% | 58,4275 |
| 44 | 97 | 1,2% | 53,5841 |
| 34 | 77 | 0,9% | 52,9114 |
| 29 | 251 | 3,1% | 52,6816 |
| 38 | 387 | 4,7% | 52,1663 |
| 16 | 87 | 1,1% | 50,7576 |
| 39 | 384 | 4,7% | 45,6743 |
| 24 | 109 | 1,3% | 45,0785 |
| 45 | 86 | 1,0% | 42,0750 |

| | | | |
|---|---|---|---|
| 30 | 200 | 2,4% | 41,5833 |
| 49 | 199 | 2,4% | 38,9479 |
| 40 | 171 | 2,1% | 36,9277 |
| 23 | 186 | 2,3% | 36,3553 |
| 25 | 222 | 2,7% | 35,7430 |
| 18 | 384 | 4,7% | 33,0961 |
| 17 | 190 | 2,3% | 27,1479 |
| 41 | 123 | 1,5% | 26,0689 |
| 50 | 101 | 1,2% | 25,8806 |
| 32 | 820 | 10,0% | 25,2260 |
| 19 | 295 | 3,6% | 24,6372 |
| 22 | 62 | 0,8% | 22,7495 |
| 26 | 217 | 2,6% | 22,6525 |
| 31 | 369 | 4,5% | 20,5281 |
| 20 | 159 | 1,9% | 19,5755 |
| 13 | 256 | 3,1% | 16,8039 |
| 33 | 201 | 2,4% | 15,0632 |
| 9 | 80 | 1,0% | 7,9261 |

Table 8 : CHAID gain summary for nodes and relative means (variation rate)

Nodes can vary in their number of constituent nodes (what we could call “weight” of a node): the minimum is node number 22, it counts 62 words and its relative value is 0.8% of the total sample, while the biggest is node number 23, it counts 820 words and its relative value is 10% of the total sample.

To check the exact functioning of CHAID in details, we can see – at the bottom of the same spreadsheet how the iterative procedure works by combining the logical operators IF, OR and THEN: for any dependent variable CHAID evaluates every possible supposed predictive variable by relying on the more appropriate significativity test.

Even if the above table has been conceived to give an account of the gain each node contributes to give, we set it in order to it can show us an descendent order in terms of the mean (that is variation rate), allowing us to start the analysis.

-

In **node 6** almost everything the infant says is a variated form, while - at the bottom - **node 9** has a variation rate (mean) of 7.9, meaning that Adrien utters the 92% of what he said correctly

Let's see in details which kind of words and phonemes are concerned and at which age: generally speaking, simpler words are more properly pronounced than more complex ones, an easy observation.

If we look at annex 2, 3 and 5, especially number 2, we could see whether CHAID segments match with current literature on language acquisition (estimated chronological order of phonemes acquisition) : in node there are mainly bilabials like "p", "b", "m", "n" learnt before all the other consonants. While in node 6 there are full of "s","z", "r", "l" and other sounds are articulated by combining more articulatory elements (see Yamaguchi N., annex 6 and 8).

These findings show that – despite CHAID is blind to any morphological/articulatory detail – it can sort children sentences in a roughly correct way that approximately goes beyond 50% in every segment.

As time is the major predictor of SPVR, we chose to improve this splitting process (and thus improving its subsequent evaluation too) by costraining CHAID processing with time: by doing so, the iterative calculations will give priority to the variable "time" in creating subsequent nodes.

This allows us to evaluate sentences variation rate in relation to time: so, compared to the previous attempt, our hypothesis was that we would have had more targeted segments, and this has been (see the same Excel document, spreadsheet named "CHAID_SPVR (ts) time forced"

Node 1 represents all the sentences uttered before 2 and three months of age, then – in this part of the total sample, the most predictive variable is w_mod_1r, that means the first word.

So, we could consider all the subsequent nodes belonging to "Node 1" as derived from two subsequent conditional probabilities (time and w_mod_1r).

An important remark: first we tried to run CHAID decision tree in which, by default, chi squared was splitting an independent categorical variable, but as results were deceiving in

terms of clusterings' precision, we then opted to set CHAID by using Fisher (F) as a continuous independent variable.

If we compare node 11 and node 12, it seems us that we can see in a more clearer way how the accuracy of the splitting variable has been improved: segment 12 matches really well with annex 2, and even segment 11 (node 11) do the same.

Future directions: What we are seeing in an approximative way (at a glance) is that if we interpret these clusters through the lenses of three tables (see annexes): the one in which are listed the % of the relative frequencies of phonemes (P. Léon), the articulatory effort table from Sauvage and the table "average estimates of consonant production" we can (partially) account to the nodes partitions automatically done by CHAID.

What we need to do is to program a set of algorithms that could automatically compare CHAID's segments and the aforementioned tables: doing so manually is a time-consuming activity

Main limit is always that CHAID does not consider morphological differences and, it counts minimal variations such as a complete one: for example, if a child says "tacteur" instead of "tracteur" the score will be 0 and if the child says "tato" instead of "tracteur" the score will be always 0. For this reason it would be better to pursue our research by using Levenshtein Distance

Here below an example of a CHAID decision tree. See supplementary file (Adrien_20_4) for a better visualization.

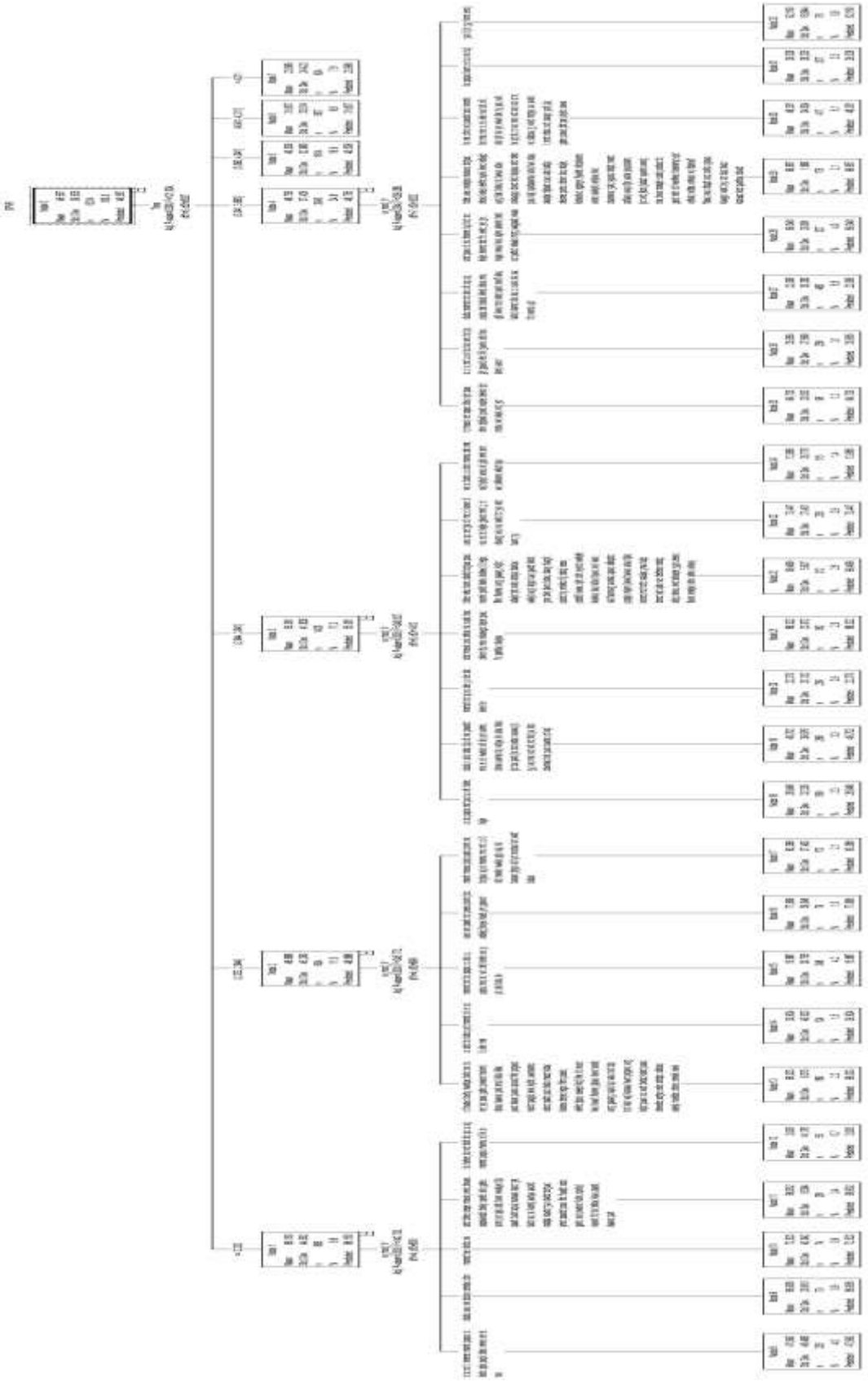


Figure 21 A decision tree splitting all Adrien's *corpus*

# Chapter 8 - CHAID on POS tags

## 8.1 Parsing with Universal dependencies POS (part-of-speech) tags

0.68+-1.96*rad(0.68*0.32)/805 due volte

We recall the "capture rate" formula given in Chapter 5 to open this chapter. This because

"One important dimension that always needs attention is the amount of sampling required for obtaining an accurate picture of the phenomenon of interest" (Stahl & Tomasello, 2004, p118)

We were sure that the score of the capture rate is good for phonemes, but we have some doubts that is good enough for POS tags, especially for not frequent ones and especially for earlier ages (where data are sparser). CoLaJE sampling is a monthly record, so we cannot avoid this constraint and, in any case, CHAID and EM give the number of occurrences ( and sentences) in which a given POS tag occurred. So we can have an estimate of the occurrences of POS tags over time.

We remind that in CoLaJE *corpus*, as well as in most of CHILDES related *corpora,* only around 1% of the total amount of what a child hears and speaks is sampled : for consonants and vowels (and consonantal sequences like Plosive-Liquid) 1% is a reliable sample because, as we explained before (see Chapter 5) we could be sure that at least one target of every French phoneme will be captured by the one hour sample.

For POS tags we have to reframe the above considerations because we are dealing with words (aka "sequences of phonemes") that are – by definition – rarer than their constituents.

At a glance, considering what we get from the "capture rate" formula in chapter 5, it seems to be not sufficient: in fact, ADJ or DET or AUX do not seem to occur more than 100 times a day, the threshold we should overcome to get a reliable capture rate according to this graph (Stahl &Tomasello, 2004)

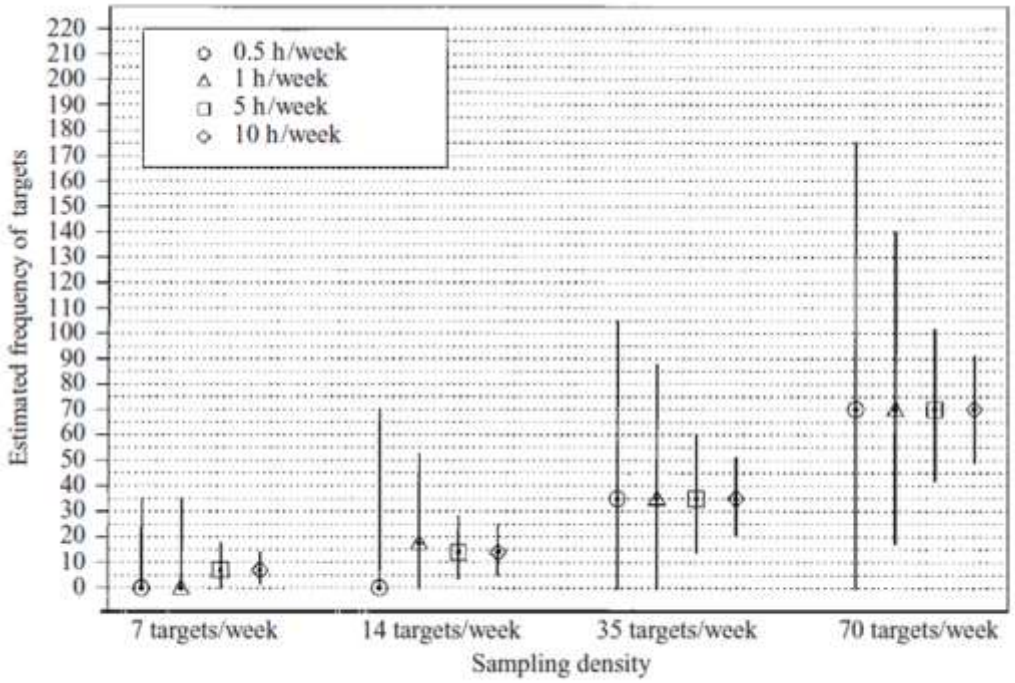



Fig. 4. Estimated weekly frequency of target (median and 95% confidence intervals) as a function of rate of occurrence and sample density.

Figure 22 (Stahl & Tomasello, 2004, p110)

Despite this potential limit, after an analysis at the phonetic unit level and single word level, we thought that an analysis at the "upper" level, let's say how words are combined together to form sentences, was needed to make our work more complete.

If it is true that language can be divided in different parts (and their related fields), we have to keep in mind that these parts are complimentary to each other:

For example, the following sentence:

"*Voulez-vous sortir ce soir?"*

draws its meaning as a question not only by the fact that the verb precedes the subject differently from affirmative sentences, neither for the presence of the question mark, but it is more likely understood by other speakers by the fact that it has an ascendant prosody[135].

[135] But in the case "est-ce que tu veux sortir ce soir?" the ascendant prosody will not hold. So, it should be made a different case between different ways of asking a question.

Yet, the boundaries between these levels are due to different disciplines rather than for objective reasons: we are used to see language as a phenomenon composed by different parts connected to each other (e.g phonetics, phonology, syntax, semantics and so on) and it makes sense, but we could even see it as a whole phenomenon that encompass all these different parts that are there only for the sake of a supposed simplicity of our analysis.

To analyse in fact means essentially to divide a big problem that cannot be understood as a whole into what we think are its elementary components: the problem is that there are many cognitive and epistemological limits, as well as school of thought, that make this logical procedure different in time and place.

To reduce as much as we can this limit and to account in what we think is the most objective way the acquisition of syntax from CoLaJE *corpora*, we chose Universal Dependencies as a reference.

How to parse a language is a process made up of many different choices: first of all, languages differs in typology (SVO, SOV, VOS) so syntax depends mostly on the language-specific convention. Then, there are many subtle differences even between languages part of the same typological area (such as English and French) , and even between languages that share a common root such as french and italian.

Here we provide a table in which POS tags specific to French language have been listed according to the Stanford University leading project "Universal Dependencies"[136]

[136] https://universaldependencies.org/

| Open class words | Closed class words | Other |
|---|---|---|
| ADJ | ADP | PUNCT |
| ADV | AUX | SYM |
| INTJ | CCONJ | X |
| NOUN | DET | |
| PROPN | NUM | |
| VERB | PART | |
| | PRON | |
| | SCONJ | |

Table 9 List POS tags specific to French

The above taxonomy is divided in three parts:

- Open class
- Closed class
- Others

Another limit is that - as far as we know - we cannot find out the rate of occurrence of POS tags as we did with single phonemes thanks to the work of Pierre Léon and colleaugues (see Chapter 5 and annex 2 at the bottom)

Maybe by using in a smart way the query provided by CoLaJE (see "interrogation descripteurs" window in the website http://modyco.inist.fr:8984/restxq/interro/ ) and by making these results complimentary to other results we could find a way to provide a rough estimate and thus calculate the confidence interval through the "capture rate" formula.

But, after all, CoLaJE – as far as we know – is one of the best existing French corpus in terms of sample density, quality of transcription and age span considered, so nothing could have been done better regarding the choice of the corpus.

We chose to apply the automatic parser based on UD POS tags on the CHI line, then we chose to verify this parsing on the "mod" line that – leaving aside rare exceptions where a given sentence has been splitted differently (see pho<> mod in paragraph..) – should be the

same sentence transcribed in two different ways (standard orthographic form CHI and IPA symbols “mod”).

In some *corpora* “mod” is missing for many reasons related to the ends of CoLaJE project (for some children more importance has been given to prosody and nonverbal cues instead of phonology) and because it may happen that from 4 years old onward, the child pronounces almost always in an adult way, so the transcriber decided to simply write a “mod” line whenever a variation appears. In this case, we decided to verify the automatic parsing made on the CHI line on a “pho” line and by controlling manually random samples from the whole corpus to check whether it was fitting or not.

Here an example from Madeleine, corpus number 30, age 4_07_04, exported from *.cha* to *.xls*:

| | | | |
|---|---|---|---|
| pho | | | katʁ ɑ̃ e dmi |
| **add** | | | à OBS. |
| **OBS** | 76,66 | 78,283 | ohlala@i . |
| **MOT** | 78,283 | 80,641 | ah et+puis il faut que tu présentes ta nouvelle chambre à Martine . |
| **CHI** | 80,641 | 84,331 | bon euh voilà une maison . |
| **pho** | | | bɔ̃ ə vwala yn mɛzɔ̃: |
| **act** | | | CHI se lève et présente sa chambre. |
| **CHI** | 84,331 | 89,251 | euh pousse [///] ici y+a des jeux +... |
| **pho** | | | ə pus isi la de ʒø: |
| **add** | | | à MAR. |
| **CHI** | 89,251 | 91,303 | +, ici y+a un tableau . |
| **pho** | | | isi la ɛ̃ tablo: |
| **CHI** | 91,303 | 93,013 | là y+a mon bureau . |
| **pho** | | | la la mɔ̃ byʁo: |
| **CHI** | 93,013 | 95,065 | là y+a mon lit bien+sûr . |
| **pho** | | | la la mɔ̃ li: bjɛ̃syʁ: |
| **CHI** | 95,065 | 97,47 | le lit d(e) mes poupées . |
| **pho** | | | lə li d me pupe: |
| **act** | | | CHI fait le tour de la chambre. |
| **MOT** | 97,47 | 99,365 | il est pas très rangé ton lit dis moi . |
| **CHI** | 99,365 | 100,007 | 0 [=! petit rire] # et euh [=! soupire] . |

Table 10 Example from Madeleine’s corpus

Here variations are almost absent, except from the reduction of “il y a” to “la”, in this case the parsing would be difficult because it will automatically detect “Il y a “ on the CHI, parsing it into a Personal pronoun (PRON), a determiner (PRON) and a verb (VERB)

and on the pho line, it will turn “la” simply into a single article (ART)

thus creating a mismatch.

In cases like this, we decide to rely only on the CHI line, renouncing to a double check: luckily, this kind of cases are rare.

Parsing is a delicate phase that has to be done in the best possible way, otherwise it will bias all the statistics and decision tree based on it afterward: so, in order to assure a correct parsing, we decided to verify it again with another parsing system called “Treebank project[137]”. Obviously, from different tags derives different parsing schemes: what we manually did is to qualitatively compare the outcomes in order to see why a given word was tagged in a different way and how, and check if this choice was coherent to the other choices taken for the other syntactically related words.

We found that – in most cases – sentences were parsed in the same way by the two automatic systems: this is probably due to the fact that they share much of their schema and by the fact that – syntactically speaking – sentences from children are for certain aspects shorter and simpler than adults one (for example, they contain less conditional clauses)

To give a representative example taken from the document “Madeleine with POS 2020_05_26”, index 6487 (Madeleine is 3;0;28):

- CHI j'ai un truc pour la soigner comme ça
- Mod j ɛ œ̃ tʁyk puʁ la swanje kɔm sa
- Pho ʒ e œ̃ tʁyk puʁ la swaŋe kɔm sa

[137] http://ftb.linguist.univ-paris-diderot.fr/ URL consulted on 13/8/2020

In this case, the sentence consists in nine grammatical elements (POS to be tagged)

| j | NC | X | ai | V | VERB | un | DET | DET | truc | NC | NOUN | Pour | P | ADP | la | DET | DET | soigner | NC | VERB | comme | P | ADP | ça | PRO | PRON |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

The first tag is automatically done by the TreeBank system while the second is done by Universal Dependencies system. We choose to apply both system to understand their inner working and to evaluate their accuracy. It is easy to observe that the second one seems better: the verb “soigner”, not so easy to recognize for an automatic system because it comes just after a pronoun that has no preceding subject (it may be inferred from the visual context, but the algorithm does not know what is going on, it only reads text), it is considered as a noun from TreeBank and correctly as a verb from UD.

Let’s look at another example:

Madeleine at 3;6;8, row 7635

- CHI c'est là où y a des autocollants
- Mod se la u ja de zotokolɑ̃
- Pho sə la u ia dɛ otokolɑ̃

In this case, the sentence consists in eight grammatical elements.

| c | DET | PRON | est | V | VERB | là | ADV | ADV | où | PREL | ADV | Y | CLO | ADV | A | V | VERB | des | DET | DET | autocollants | NC | ADJ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

In this case it is possible to observe other different cases of diversity in POS tags assignment: “c”, “où”, “y”, “autocollants”. Again, we think that Universal Dependencies better

accomplish to the task compared to TreeBank, though it is not perfect: the last word is surely a name, but probably the algorithm has given too much importance to the composite nature of this word and not as much importance to the syntactic context that would clearly suggest the choice of a common noun instead of an adjective.

A specific consideration not concerning POS tagger should be made to the occurrences of French *liaison,* as it has occurred in the last example. The algorithm set to recognize the equivalence between "pho" and "mod" has not been thought to be sensitive to such a particularity specific to French.

So, if Madeleine says

- "dɛ otokolɑ̃"
  instead of pronouncing
- "de zotokolɑ̃"

it results - according to the functioning of the algorithm set – that the child has mistakenly pronounced the noun while, strictly speaking, her pronunciation is correct. Yet, all depends on whether the *liaison* falls on the preceding word ending with a consonant or the following word beginning with a vowel. In this case, the preceding word is an article that is not correctly pronounced *a priori* to the *liaison*. The final score is two variation on two words.

For a review of the phenomenon of the *liaison* and the study of its lexical status in French children, see the work by Chevrot & Fayot[138].

Another example taken from Madeleine at 1;7;15, row 427

- CHI les petits poussins
- Mod lɛ pəti pusɛ̃
- Pho e ti puscœ̃

[138] Chevrot J-P.; Fayol M. "Acquisition of French Liaison and Related Child Errors". Research on Child Language Acquisition, vol. 2, M. Almgren, A. Barreña, M.J. Ezeizabarrena, I. Idiazabal, and B. MacWhinney (eds), Cascadilla Press, pp.760-774, 2001

There are three elements to parse. Here the results:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| les | DET | DET | Petits | ADJ | ADJ | poussins | NC | NOUN |

This sentence is simpler and less ambiguous compared to the previous ones, so the two automatic parsing systems agree on the POS tags assignment.

Another consideration should be made: the sentence is parsed on CHI because the algorithm does not recognise IPA characters. This implies that we totally rely on CoLaJE researchers' transcription: their choice regarding interpretations of partial forms such as in this sentence influence all the subsequent steps of our research. In this specific case "e" in pho tiers is intepreted as the article "les", "ti" is interpreted as the adjective "petits" in its plural and masculine form, and "pusœ̃" is interpreted as "pusɛ̃" in its plural and masculine form.

The syntactic form of this sentence makes it easy the interpretation: the context clearly narrows the possibility of the first two words

But in some cases transcription details can hamper automatic parsing. In Madeleine 2;6,10 row 4421

- CHI pour i l puisse jouer avec moi faut lui mettre ça
- Mod puʁ i pɥis ʒwe avɛk mwa fo lɥi mɛtʁ sa
- Pho pu i pɥis ʒue avɛ mwa fo lɥi mɛt sa

There are ten elements to be parsed, but for reasons unknown the pronoun "ils" has a space between the first vowel "i" and the following consonant "l": it is not clear if this is due to a simple transcription error or if the person who transcribed wanted to highlight a vocalic lengthening of "i", maybe due to the fact that is a quite complex sentence for a child as it contains a subjunctive form and, although Madeleine is considered to be a linguistically gifted girl – during this recording she was still two year and an half. A "que" is in fact missing between the adposition "pour" and the pronoun "il".

The two systems parsed the sentence as follows:

| pour | P | ADP | i | DET | X | l | NC | X | puisse | VS | AUX | jouer | VINF | VERB | avec | P | ADP | moi | PRO | PRON | faut | V | VERB | lui | CLO | PRON | mettre | VINF | VERB | ça | PRO | PRON |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Universal Dependencies does not recognise the “i” and “ l” while Tree Bank seem to interpret them in a quite arbitrary way as an article followed by a noun, despite in French “i” is not an article.

To sum up, parsing could be sometimes messy for the reasons explained and many other reasons that it is not the case to write in this thesis. In light of these considerations, we finally decided to build up our decision trees on the basis of Universal Dependencies POS tagger.

See supplementary file “Adrien_results_20_4_20” CHAID_SPVR_TAGS

Aim is to see how a given POS tag could influence the splitting procedure over the ages, then verify if these branches contain sentences and/or words that would confirm current theories regarding development of grammar and morphosyntax.

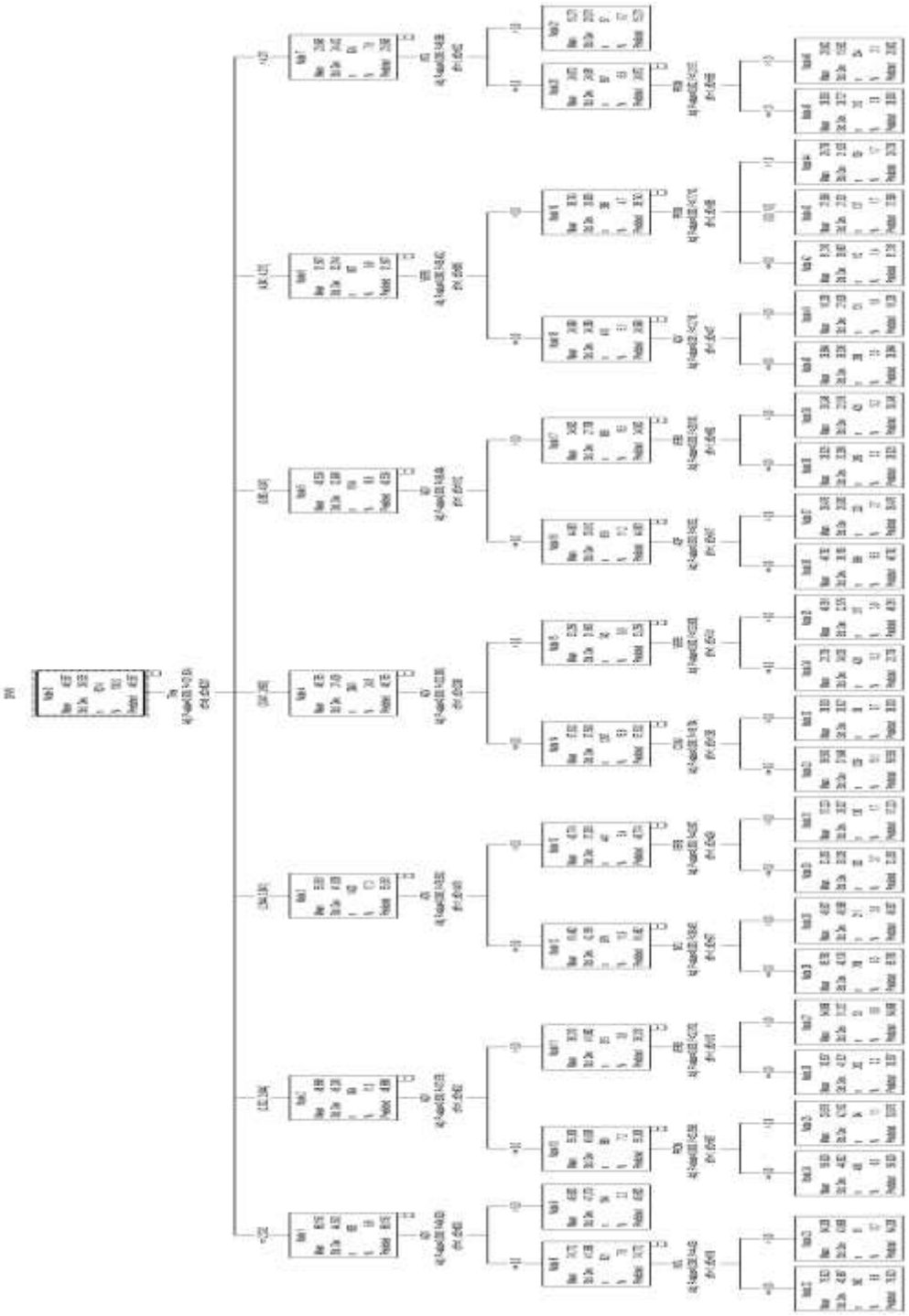


Figure 24 : Decision tree (*see supplementary file for a better resolution)*

| Node | Mean | Std. Deviation | N | Percent | Predicted Mean | Parent Node | Primary Variable | Sig.a | F | df1 | df2 | Split Values |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 46,5566 | 39,53482 | 8214 | 100,0% | 46,5566 | | | | | | | |
| 1 | 68,1504 | 44,50235 | 805 | 9,8% | 68,1504 | 0 | Time | 0,000 | 121,624 | 6 | 8207 | <= 2.332 |
| 2 | 48,6885 | 45,28974 | 904 | 11,0% | 48,6885 | 0 | Time | 0,000 | 121,624 | 6 | 8207 | (2.332, 2.644] |
| 3 | 55,0505 | 41,82774 | 1420 | 17,3% | 55,0505 | 0 | Time | 0,000 | 121,624 | 6 | 8207 | (2.644, 3.041] |
| 4 | 48,7546 | 37,42604 | 2040 | 24,8% | 48,7546 | 0 | Time | 0,000 | 121,624 | 6 | 8207 | (3.041, 3.693] |
| 5 | 40,5388 | 32,99019 | 1614 | 19,6% | 40,5388 | 0 | Time | 0,000 | 121,624 | 6 | 8207 | (3.693, 4.041] |
| 6 | 31,6072 | 33,01391 | 807 | 9,8% | 31,6072 | 0 | Time | 0,000 | 121,624 | 6 | 8207 | (4.041, 4.211] |
| 7 | 23,9946 | 24,42161 | 624 | 7,6% | 23,9946 | 0 | Time | 0,000 | 121,624 | 6 | 8207 | > 4.211 |
| 8 | 74,7719 | 41,39774 | 621 | 7,6% | 74,7719 | 1 | ADV | 0,000 | 64,929 | 1 | 803 | <= .0 |
| 9 | 45,8031 | 47,36978 | 184 | 2,2% | 45,8031 | 1 | ADV | 0,000 | 64,929 | 1 | 803 | > .0 |
| 10 | 55,3085 | 45,63760 | 589 | 7,2% | 55,3085 | 2 | ADV | 0,000 | 37,578 | 1 | 902 | <= .0 |
| 11 | 36,3102 | 41,98269 | 315 | 3,8% | 36,3102 | 2 | ADV | 0,000 | 37,578 | 1 | 902 | > .0 |
| 12 | 61,4815 | 42,15575 | 979 | 11,9% | 61,4815 | 3 | ADV | 0,000 | 78,592 | 1 | 1418 | <= .0 |
| 13 | 40,7740 | 37,36312 | 441 | 5,4% | 40,7740 | 3 | ADV | 0,000 | 78,592 | 1 | 1418 | > .0 |
| 14 | 57,6316 | 37,59272 | 1297 | 15,8% | 57,6316 | 4 | ADV | 0,000 | 222,060 | 1 | 2038 | <= .0 |
| 15 | 33,2586 | 31,66334 | 743 | 9,0% | 33,2586 | 4 | ADV | 0,000 | 222,060 | 1 | 2038 | > .0 |
| 16 | 44,8010 | 35,91226 | 919 | 11,2% | 44,8010 | 5 | ADV | 0,000 | 36,404 | 1 | 1612 | <= .0 |
| 17 | 34,9030 | 27,70801 | 695 | 8,5% | 34,9030 | 5 | ADV | 0,000 | 36,404 | 1 | 1612 | > .0 |
| 18 | 24,9990 | 34,38882 | 419 | 5,1% | 24,9990 | 6 | VERB | 0,000 | 36,452 | 1 | 805 | <= .0 |
| 19 | 38,7434 | 29,90491 | 388 | 4,7% | 38,7434 | 6 | VERB | 0,000 | 36,452 | 1 | 805 | > .0 |
| 20 | 24,8716 | 24,66142 | 567 | 6,9% | 24,8716 | 7 | INTJ | 0,005 | 8,096 | 1 | 622 | <= .0 |
| 21 | 15,2707 | 20,07096 | 57 | 0,7% | 15,2707 | 7 | INTJ | 0,005 | 8,096 | 1 | 622 | > .0 |
| 22 | 75,9226 | 40,99665 | 560 | 6,8% | 75,9226 | 8 | INTJ | 0,036 | 4,429 | 1 | 619 | <= .0 |
| 23 | 64,2077 | 43,86808 | 61 | 0,7% | 64,2077 | 8 | INTJ | 0,036 | 4,429 | 1 | 619 | > .0 |
| 24 | 59,9293 | 44,83201 | 495 | 6,0% | 59,9293 | 10 | PRON | 0,000 | 33,558 | 1 | 587 | <= .0 |
| 25 | 30,9751 | 42,19231 | 94 | 1,1% | 30,9751 | 10 | PRON | 0,000 | 33,558 | 1 | 587 | > .0 |
| 26 | 30,5070 | 41,52128 | 262 | 3,2% | 30,5070 | 11 | VERB | 0,000 | 32,763 | 1 | 313 | <= .0 |
| 27 | 64,9977 | 31,32653 | 53 | 0,6% | 64,9977 | 11 | VERB | 0,000 | 32,763 | 1 | 313 | > .0 |
| 28 | 65,7796 | 40,13606 | 768 | 9,3% | 65,7796 | 12 | INTJ | 0,000 | 38,461 | 1 | 977 | <= .0 |
| 29 | 45,8373 | 45,59761 | 211 | 2,6% | 45,8373 | 12 | INTJ | 0,000 | 38,461 | 1 | 977 | > .0 |
| 30 | 33,2826 | 39,32782 | 303 | 3,7% | 33,2826 | 13 | VERB | 0,000 | 42,607 | 1 | 439 | <= .0 |
| 31 | 57,2226 | 26,02683 | 138 | 1,7% | 57,2226 | 13 | VERB | 0,000 | 42,607 | 1 | 439 | > .0 |
| 32 | 58,5504 | 37,64755 | 1239 | 15,1% | 58,5504 | 14 | CONJ | 0,000 | 16,754 | 1 | 1295 | <= .0 |
| 33 | 38,0028 | 30,62670 | 58 | 0,7% | 38,0028 | 14 | CONJ | 0,000 | 16,754 | 1 | 1295 | > .0 |
| 34 | 23,7392 | 34,03525 | 426 | 5,2% | 23,7392 | 15 | VERB | 0,000 | 102,608 | 1 | 741 | <= .0 |
| 35 | 46,0512 | 22,57930 | 317 | 3,9% | 46,0512 | 15 | VERB | 0,000 | 102,608 | 1 | 741 | > .0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 36 | 46,7918 | 38,10594 | 699 | 8,5% | 46,7918 | 16 | ADP | 0,003 | 9,052 | 1 | 917 | <= .0 |
| 37 | 38,4755 | 26,92993 | 220 | 2,7% | 38,4755 | 16 | ADP | 0,003 | 9,052 | 1 | 917 | > .0 |
| 38 | 28,0248 | 33,26598 | 269 | 3,3% | 28,0248 | 17 | VERB | 0,000 | 28,100 | 1 | 693 | <= .0 |
| 39 | 39,2463 | 22,51634 | 426 | 5,2% | 39,2463 | 17 | VERB | 0,000 | 28,100 | 1 | 693 | > .0 |
| 40 | 28,9839 | 36,30570 | 288 | 3,5% | 28,9839 | 18 | ADV | 0,001 | 12,716 | 1 | 417 | <= .0 |
| 41 | 16,2383 | 27,92844 | 131 | 1,6% | 16,2383 | 18 | ADV | 0,001 | 12,716 | 1 | 417 | > .0 |
| 42 | 51,3179 | 36,68651 | 112 | 1,4% | 51,3179 | 19 | PRON | 0,000 | 17,710 | 2 | 385 | <= .0 |
| 43 | 37,5992 | 27,52133 | 137 | 1,7% | 37,5992 | 19 | PRON | 0,000 | 17,710 | 2 | 385 | (.0, 1.0] |
| 44 | 29,7390 | 21,62464 | 139 | 1,7% | 29,7390 | 19 | PRON | 0,000 | 17,710 | 2 | 385 | > 1.0 |
| 45 | 28,0930 | 29,72688 | 313 | 3,8% | 28,0930 | 20 | PRON | 0,002 | 12,157 | 1 | 565 | <= 1.0 |
| 46 | 20,9019 | 15,56216 | 254 | 3,1% | 20,9019 | 20 | PRON | 0,002 | 12,157 | 1 | 565 | > 1.0 |

Table 11 Decision tree, node by node

Node 0 represents the total amount of what Adrien said during all the recordings. Going from left to right, mean is a value representing the variation rate (error), SD is the standard deviation, then F is Fisher. Sig.a F df1 df2 Split Values

Note that the first seven nodes are splitted by the time. Age turns out to be the main regressor, it divides the corpus according to a decreasing "mean" (although node 3 represents an exception as it is higher than the previous). From node 8 to node 17 ADV (adverbs) determines the first splitting, this is probably due to the fact that in this POS category are words such as "oui", "non", "très", beaucoup" the first two are holophrastic words while the latter two are highly frequent words. Node 6 is splitted by VERB and this is conformed to our expectation because a 4 year-old child should masters in a proper way many verbs and thus this POS tag should be more frequent than before. If we look at this age in the graphs proposed by Morgenstern & Parisse ("The Paris corpus, 2012), it is possible to observe the importance of development in this age: type/token ration and mean length of utterances reach their higher levels in the graphs.

The last node of the first row deserves a specific consideration: it is about everything Adrien said after the age of 4 years and two months and is firstly splitted by the absence or th presence of INTJ (interjections). This is curious because INTJ – according to our expectations – would not have to play a great role generally, and especially in the upper age,

where more abstract and syntactically central POS tags are expected to play a great role. So, why is there?

The answer is just below: if we look at N values, the node where no INTJ are show a N value of 567, while the node where all INTJ are show a N value of 57. This is due to the inner working of CHAID: it is of course a set of algorithm that is not grammatically-sensitive, it splits according to the rules described in the previous paragraph: if it can neatly divide two segments in a mutually exclusive way, it does it. In fact, if we check which kind of sentences are grouped in node 21, it is clear that there are mostly short sentences containing exclamations. This is confirmed by a mean value lower than the other segment (15 vs 24), and by the fact that the node with INTJ cannot be splitted anymore by CHAID iterative algorithm, while the node without INTJ, that contain richer sentences, is splitted again through selecting PRON ( pronouns) as the next most determinant variable. This is similar to what happen to the other close segments and is in line to what we found by using the EM clustering method (see next chapter).

Let's look closer on focus on a specific branching schema, here is an extract of the algorithm at work:

*/* Node 22 */.*

*IF (Time NOT MISSING AND (Time <= 2.331506849315069)) AND (ADV IS MISSING OR (ADV <= 0)) AND (INTJ IS MISSING OR (INTJ <= 0))*

*THEN*

*Node = 22*

*Prediction = 75.9226249999999*

*/* Node 23 */.*

*IF (Time NOT MISSING AND (Time <= 2.331506849315069)) AND (ADV IS MISSING OR (ADV <= 0)) AND (INTJ NOT MISSING AND (INTJ > 0))*

*THEN*

*Node = 23*

*Prediction = 64.2077049180328*

*/* Node 9 */.*

*IF (Time NOT MISSING AND (Time <= 2.331506849315069)) AND (ADV NOT MISSING AND (ADV > 0))*

*THEN*

*Node = 9*

*Prediction = 45.803097826087*

In the figure below are represented the first two nodes (earlier ages) of the decision tree obtained by applying CHAID on Adrien. It is possible to observe that mean decrease as expected and N, a value that represents the number of total sentences grouped into a given node, increases too. These two tendencies are confirmed in almost every node, besides a couple of exceptions that are in line to the graphs of Morgenstern & Parisse (2012), in which regression lines do not increase proportionally over time but instead show an up-and-down shape that increase only if considered in a global perspective (Morgenstern & Parisse, 2012).

We can see that in nodes 1 and 2 the presence of ADV determines a neat decrease in mean (variation rate) of the subsequent branches: < means that the POS tag is not present in the below node while > means that it is present: in node 9 mean is 45 (against 74) and in node 11 mean is 36 (against 55). Standard deviations are not significantly different between different branches. Both nodes with ADV are smaller than nodes without, this holds for all the other couples of branches divided by ADV and directly deriving from the first two of nodes determined by time (age classes). It is difficult to say what this tendency could reveal.

Node 8 is then divided in nodes 22 and 23 according to the presence or absence of INTJ. It is clear that when INTJ is there, sentences are less in number and are articulated in a better way, this is partly due to holophrastic words. While VERB increases variation rate in every node it contributes to create: in the image below, mean in node 27 is more than two times than mean in node 26, in the following – when the child is more than 4 years-olds, the difference is smaller but still significant: node 18 has mean 25 and node 19 has mean 38. This because verbs changes in person, temporal forms and gender, this requires more time to the child to learn all these context-dependent differences.

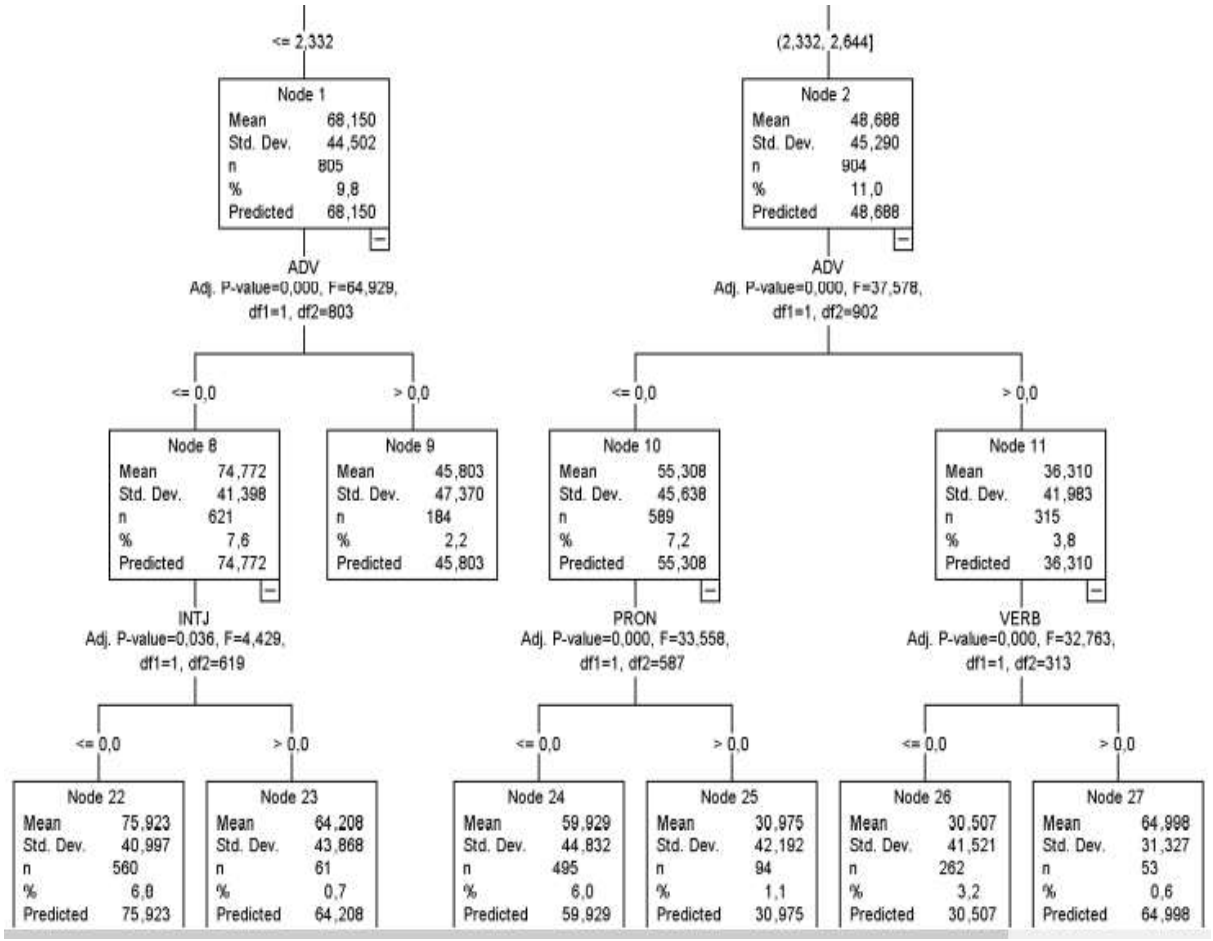



Figure 25 Decision tree in detail

For a similar reason PRON is a POS tag that becomes important in later ages, while it does not seem to play an important role in earlier ages. In Universal Dependencies standard of references are listed as French pronouns all these particles: personal pronouns, demonstrative pronouns, reflexive pronouns, interrogative/relative pronouns. In fact pronouns do not take part into the one-word holophrastic period, their use imply the passing of the mirror mark test (as explained in Chapter 2) because it involves the use of reflexity ( “ such as in “me”, “se”: “me rendre compte”, “se sentir mal”) and/or relativity (such as in “qui”, “que”; “on parlait du chat qui j’ai vu hier”).

From CHAID results, we can observe a steady evolution of the use of pronouns: it is in the increase of the N (total number of sentences) in every branch in which PRON is >1

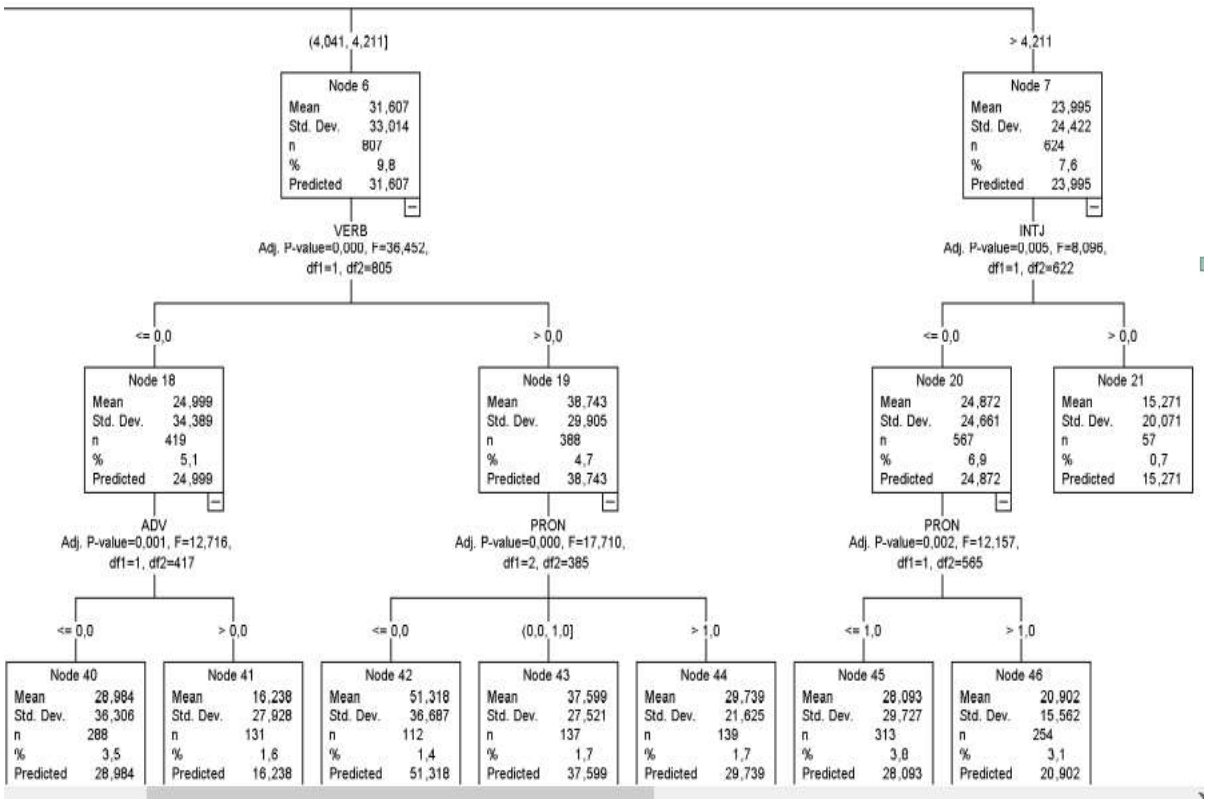



Figure 26 decision tree in detail (two)

# Chapter 9 - EM Clustering Method on Adrien parsed sentences

## 9.1 EM at works

To extend previous research on CHAID [139] we divide our database in strata considering 3 different age classes of the child (L=1.97 - 2.64; M= 2.71 - 3.39 H=3.46 - 4.33 expressed in years and months) and 3 classes of SPVR (L=33; M=>33 and 66; H>66 in percent) representing three different classes of variation rate calculated at the sentence level. This is analogous to what has been done to CHAID time-forced version (previous chapter): as time is the main regressor of child language development, we manually divided in three age classes the 8214 total sentences Adrien produced during CoLaJE recordings in order to obtain a more easy-to-read picture of his development (reading all ages together would be difficult for reason merely due to the visual rendering). In total we get 9 strata (from LL to HH).

Creating these nine strata is a strategy that we put in place to get a more readable final result: in fact, during the first attempts in using this clustering procedure, the main problem was that results were difficult to intepret primarily because of their huge quantity and secondarily because of the absence of an order in time and/or variation rate.

By framing the analysis in this way, we turn EM clustering algorithm into a potentially interesting method that could provide a reliable way to observe linguistic structures development over time.

[139] Briglia A., Mucciardi M., Sauvage J. "Identify the speech code through statistics: a data-driven Approach". Proceedings SIS 2020 (Pearson Editions). (2020)

| code | STRATA | TIME (age) | SPVR |
|---|---|---|---|
| 1 | LL | 1.97 - 2.64 | <=33% |
| 2 | LM | 1.97 - 2.64 | >33% and <=66% |
| 3 | LH | 1.97 - 2.64 | >66% |
| 4 | ML | 2.71 - 3.39 | <=33% |
| 5 | MM | 2.71 - 3.39 | >33% and <=66% |
| 6 | MH | 2.71 - 3.39 | >66% |
| 7 | HL | 3.46 - 4.33 | <=33% |
| 8 | HM | 3.46 - 4.33 | >33% and <=66% |
| 9 | HH | 3.46 - 4.33 | >66% |
| Mixture = | POISSON | | |

Table 12 : Strata details

Then, we applied Part-Of-Speech Tagger (POS Tags), a software that reads text in and assigns parts of speech to each word such as noun, verb, adjective. We used Stanza Core NLP engine[140] to tag all CHI words by using Universal Dependencies as a standard of reference for part-of-speech classification

The EM clustering is an iterative method relying on the assumption that the data is generated by a mixture of underlying probability distributions, where each component represents a separate group, or cluster. The method provides the optimal number of clusters in any empirical situation, by using a two steps iterative algorithm: the (E) or expectation step and the (M) or maximization step. These two steps are repeated until a further increase in the number of clusters would result in a negligible improvement in the log-likelihood, namely a convergence. Accordingly, the program checks how much the overall fit improves in passing from one to two clusters (formed in all possible ways, and selecting the best), then from two to three, etc. If the error function calculated for the solution with K+1 clusters is not marked (e.g at least 5 percent better) more than the simpler solution with K clusters, then the solution

[140] Zhang Y.; Zhang Y.; Bolton J.; Manning C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations, (2020)

with K clusters is considered ideal and retained [9] [10]. Considering the nature of the variables (count data), we use finite multivariate Poisson mixtures in the EM procedure.

**strata = LL**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 0.02286 | 3.34 | 1.40 |
| | N | 35 | 35 | 35 |
| **2** | Mean | 0.00746 | 1.29 | 1.11 |
| | N | 469 | 469 | 469 |
| **3** | Mean | 0.01047 | 1.90 | 1.49 |
| | N | 107 | 107 | 107 |
| **Total** | Mean | 0.00887 | 1.52 | 1.19 |
| | N | 611 | 611 | 611 |

**strata = LM**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 1.02642 | 2.49 | 2.09 |
| | N | 179 | 179 | 179 |
| **2** | Mean | 1.35000 | 3.60 | 2.00 |
| | N | 5 | 5 | 5 |
| **Total** | Mean | 1.03522 | 2.52 | 2.09 |
| | N | 184 | 184 | 184 |

**strata = LH**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 2.26148 | 2.76 | 2.20 |
| | N | 88 | 88 | 88 |
| **2** | Mean | 2.50543 | 1.47 | 1.27 |
| | N | 199 | 199 | 199 |
| **3** | Mean | 2.56135 | 1.24 | 1.05 |
| | N | 401 | 401 | 401 |
| **4** | Mean | 1.52938 | 1.65 | 1.26 |
| | N | 226 | 226 | 226 |
| **Total** | Mean | 2.26513 | 1.54 | 1.26 |
| | N | 914 | 914 | 914 |

**strata = ML**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 0.00380 | 1.45 | 1.22 |
| | N | 527 | 527 | 527 |
| **2** | Mean | 0.25863 | 3.99 | 3.32 |
| | N | 95 | 95 | 95 |
| **3** | Mean | 0.02706 | 4.53 | 2.47 |
| | N | 17 | 17 | 17 |
| **4** | Mean | 0.03165 | 1.82 | 1.52 |
| | N | 158 | 158 | 158 |
| **5** | Mean | 0.00926 | 1.67 | 1.17 |
| | N | 54 | 54 | 54 |
| **Total** | Mean | 0.03823 | 1.88 | 1.53 |
| | N | 851 | 851 | 851 |

**strata = MM**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 0.79806 | 3.25 | 2.84 |
| | N | 309 | 309 | 309 |
| **2** | Mean | 0.88307 | 4.37 | 3.52 |
| | N | 101 | 101 | 101 |
| **3** | Mean | 0.88958 | 3.95 | 3.29 |
| | N | 216 | 216 | 216 |
| **Total** | Mean | 0.84335 | 3.67 | 3.10 |
| | N | 626 | 626 | 626 |

**strata = MH**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 2.22447 | 2.31 | 2.11 |
| | N | 300 | 300 | 300 |
| **2** | Mean | 1.48203 | 3.79 | 3.04 |
| | N | 330 | 330 | 330 |
| **3** | Mean | 1.94020 | 1.42 | 1.22 |
| | N | 506 | 506 | 506 |
| **Total** | Mean | 1.88217 | 2.34 | 1.98 |
| | N | 1136 | 1136 | 1136 |

**strata = HL**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 0.16718 | 4.97 | 4.20 |
| | N | 479 | 479 | 479 |
| **2** | Mean | 0.00334 | 1.38 | 1.25 |
| | N | 673 | 673 | 673 |
| **3** | Mean | 0.18341 | 5.59 | 4.75 |
| | N | 463 | 463 | 463 |
| **4** | Mean | 0.23190 | 14.28 | 9.83 |
| | N | 147 | 147 | 147 |
| **Total** | Mean | 0.11427 | 4.54 | 3.69 |
| | N | 1762 | 1762 | 1762 |

**strata = HM**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 0.67471 | 5.77 | 4.72 |
| | N | 210 | 210 | 210 |
| **2** | Mean | 0.74856 | 6.75 | 5.77 |
| | N | 305 | 305 | 305 |
| **3** | Mean | 0.63228 | 4.32 | 3.63 |
| | N | 521 | 521 | 521 |
| **4** | Mean | 0.79298 | 3.14 | 2.72 |
| | N | 151 | 151 | 151 |
| **5** | Mean | 0.59291 | 13.55 | 9.35 |
| | N | 55 | 55 | 55 |
| **Total** | Mean | 0.68581 | 5.43 | 4.48 |
| | N | 1242 | 1242 | 1242 |

**strata = HH**

| clust_POS_p | | NLD | CHI_total_words_tokenized | CHI_total_distinct_words |
|---|---|---|---|---|
| **1** | Mean | 1.58434 | 1.21 | 1.19 |
| | N | 175 | 175 | 175 |
| **2** | Mean | 1.96043 | 1.93 | 1.63 |
| | N | 115 | 115 | 115 |
| **3** | Mean | 1.45775 | 2.55 | 2.18 |
| | N | 324 | 324 | 324 |
| **4** | Mean | 1.13025 | 4.65 | 3.55 |
| | N | 120 | 120 | 120 |
| **5** | Mean | 1.24669 | 5.55 | 4.41 |
| | N | 154 | 154 | 154 |
| **Total** | Mean | 1.46694 | 3.01 | 2.49 |
| | N | 888 | 888 | 888 |

Table 12 EM results by strata

The following table provides three general indexes describing how child language is developing in quantity, quality and accuracy: these variables are represented respectively in, Child Total Words Tokenized (CTWT), Child Total Distinct Words Tokenized (CTDWT) and Normalized Levenshtein Distance (NLD). In particular NLD [4] is a string metric for calculating the edit distance between two given words, that means the number of deletion, insertion or substitutions of a single character needed to turn one word into the other. To obtain a realistic picture of the variation rate over a child's ages, we adjust the Levenshtein Distance by normalizing it: this means that the rate will be expressed in relative values, thus obtaining a result capable of comparing shorter and longer sentences We can observe the validity of NLD by the fact that it decreases over the three slots of ages as the child improves his language. In a coherent way, CTWT, the total number of words pronounced, increases and the CTDWT, the total number of different word types (proxy of an index of lexical diversity) increases as well with a similar rate.

| STRATA | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Corpus index | LL | LM | LH | ML | MM | MH | HL | HM | HH |
| NLD* | 0.01 | 1.04 | 2.27 | 0.04 | 0.84 | 1.88 | 0.11 | 0.69 | 1.47 |
| CTWT** | 1.52 | 2.52 | 1.54 | 1.88 | 3.67 | 2.34 | 4.54 | 5.43 | 3.01 |
| CTDWT*** | 1.19 | 2.09 | 1.26 | 1.53 | 3.10 | 1.98 | 3.69 | 4.48 | 2.49 |
| # of sentences | 611 | 184 | 914 | 851 | 626 | 1136 | 1762 | 1242 | 888 |

Table 13 EM results, general indexes

| Ordered POS** | LL (3) | PSM | LM (2) | PSM | LH (4) | PSM | ML (5) | PSM | MM (3) | PSM | MH (3) | PSM | HL (4) | PSM | HM (5) | PSM | HH (5) | PSM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| POS1 | **INTJ** | 0,13 | **VERB** | 0,25 | **PRON** | 0,09 | **CCONJ** | 0,05 | **ADP** | 0,18 | **PRON** | 0,41 | **PRON** | 1,16 | **NOUN** | 0,55 | **AUX** | 0,26 |
| POS2 | **DET** | 0,09 | **PROPN** | 0,04 | **ADV** | 0,36 | **PRON** | 0,13 | **ADV** | 0,65 | **AUX** | 0,20 | **DET** | 0,32 | **DET** | 0,47 | **NOUN** | 0,31 |
| POS3 | **ADP** | 0,01 | *ADV* | 0,59 | **DET** | 0,08 | **NOUN** | 0,22 | **DET** | 0,28 | **NOUN** | 0,31 | **VERB** | 0,79 | **PRON** | 1,48 | **VERB** | 0,67 |
| POS4 | **NOUN** | 0,47 | *NOUN* | 0,75 | **VERB** | 0,18 | **AUX** | 0,05 | **SCONJ** | 0,04 | **DET** | 0,16 | **NOUN** | 0,42 | **ADJ** | 0,13 | **DET** | 0,20 |
| POS5 | **SYM** | 0,02 | *INTJ* | 0,18 | **NOUN** | 0,62 | **VERB** | 0,16 | **CCONJ** | 0,04 | **ADP** | 0,11 | **SCONJ** | 0,15 | **AUX** | 0,37 | **PRON** | 0,74 |
| POS6 | **ADV** | 0,56 | *PRON* | 0,20 | **INTJ** | 0,06 | **NUM** | 0,04 | **INTJ** | 0,17 | **ADV** | 0,38 | **ADP** | 0,23 | **VERB** | 1,02 | **NUM** | 0,09 |
| POS7 | **PROPN** | 0,02 | *DET* | 0,17 | **PROPN** | 0,05 | **SYM** | 0,02 | **NOUN** | 0,52 | **PROPN** | 0,08 | **AUX** | 0,21 | **ADP** | 0,26 | **ADJ** | 0,09 |
| POS8 | **PRON** | 0,02 | *AUX* | 0,10 | **AUX** | 0,04 | **ADV** | 0,83 | **ADJ** | 0,09 | **SCONJ** | 0,02 | **ADV** | 0,73 | **ADV** | 0,67 | **ADP** | 0,12 |
| POS9 | **VERB** | 0,02 | *NUM* | 0,07 | **ADJ** | 0,02 | **DET** | 0,09 | **NUM** | 0,04 | **VERB** | 0,44 | **ADJ** | 0,09 | **SCONJ** | 0,10 | **ADV** | 0,31 |
| POS10 | **X** | 0,02 | *CCONJ* | 0,05 | **SCONJ** | 0,00 | **PROPN** | 0,03 | **PROPN** | 0,04 | **INTJ** | 0,06 | **CCONJ** | 0,12 | **X** | 0,02 | **X** | 0,03 |
| POS11 | **CCONJ** | 0,02 | *ADP* | 0,03 | **CCONJ** | 0,01 | **ADP** | 0,03 | **AUX** | 0,28 | **NUM** | 0,03 | **SYM** | 0,02 | **CCONJ** | 0,11 | **PROPN** | 0,02 |
| POS12 | **SCONJ** | 0,01 | *X* | 0,03 | **ADP** | 0,01 | **X** | 0,03 | **VERB** | 0,62 | *X* | 0,01 | **NUM** | 0,08 | **NUM** | 0,04 | **SCONJ** | 0,04 |
| POS13 | **AUX** | 0,01 | *ADJ* | 0,02 | **NUM** | 0,02 | **INTJ** | 0,18 | **PRON** | 0,70 | *SYM* | 0,00 | **X** | 0,02 | *SYM* | 0,01 | **CCONJ** | 0,04 |
| POS14 | **NUM** | 0,10 | *SCONJ* | 0,02 | **SYM** | 0,00 | **ADJ** | 0,01 | **SYM** | 0,01 | *ADJ* | 0,10 | **PROPN** | 0,03 | *INTJ* | 0,15 | **INTJ** | 0,08 |
| POS15 | *ADJ* | 0,00 | *SYM* | 0,00 | *X* | 0,00 | *SCONJ* | 0,01 | *X* | 0,00 | *CCONJ* | 0,01 | **INTJ** | 0,16 | *PROPN* | 0,03 | *SYM* | 0,00 |

* PSM = POS Strata Mean; (#) = Clusters numbers in brackets ** POS sorted for F-test (in bold p<0.05)

Table 14 EM most influential POSt

This horizontal table summarizes the main results obtained from clustering through a detailed overview on the most influential POS tags for each strata and its related clusters. In addition,

the means of the POS are calculated in each strata (PSM). We recall that the difference between SPVR and NLD is in the different way of quantifying the variation rate: SPVR counts as a variated form every word that is not pronounced exactly as it should have been pronounced (coarse-grained), while NLD gives a percentage of the number of letters by which the pronounced word differs from the target word (fine-grained). These general indexes have been calculated to test the soundness of our dataset: this was necessary because the following analysis and computations applied (parsing and EM) would inevitably be heavily biased by any error occurred in this initial step. Let's move on to comment on the EM clustering results in detail. We can see that VERB occupies an increasing important role in development: it is almost absent in the earlier age strata (PSM = L 0.02; M 0.25; H 0.18), it develops sharply in median age strata (PSM = 0.16; 0.62; 0.44) while it is present in almost any sentence in the upper age strata (PSM = (0.79; 1.02; 0.67): it is clear also that VERB causes an increase in the error rate, as their values are higher in higher error rate strata (more than 33 percent). We can further explain the fact that VERB is higher in the LM, MM and HM strata by looking at the CTWT and CTDWT in the corresponding cells in table 1: they both have higher values as compared to the other strata: this because in these strata sentences are longer than the others and - a fortiori - they contain more verbs. If we want to know which specific verbs occur in the different clusters of a given strata, it is possible to observe the POS Cluster Mean (PCM) (values not shown) and read which kind of sentences have been placed in a specific cluster: from our results, it is possible to see how complex verbs (past and future forms, even in combination with auxiliaries) appear in later age clusters where PCM is higher than 0.5 while common verbs such as "to do", "to be", "to say", "to like" occur mainly in their present form in both low and high valued PCM in earlier strata clusters without any significant distribution detected. This difference in clustering is probably due to the fact that a two years-old child essentially expresses himself through 1-2 words per sentence, so it is hard to divide something that already represents a unit in itself. When the child is four years-old the clustering procedure divides in a much clearer way the corpus, helped by the fact that sentences are longer and grammatically richer.

Morphosyntactic coherence[141]. If we look at the single sentence, we can observe that morphosyntactic coherence is higher in HL, HM clusters compared to those in L layers, which is in line with Parisse's results, we can also observe that the parts of the speech PRON, VERB, SCONJ - which could be considered as markers of longer sentences - increase their importance (see the PSM in table 2 and 3) along the age progression.

Here below a couple of examples:

escargot tout chaud (CHI)

- EskaKgo tu So (PHO) –

didago to so (MOD)

in MH strata;

une souris verte (CHI) –

yn suKi vEKt@ (PHO) –

yn oji vat@ (MOD)

in HH strata.

In the first, morphosyntactic coherence is expressed in a coherent way in the masculine form, but the pronoun has not been pronounced while in the second sentence the pronoun is correctly there and it is morphosyntactically coherent with the feminine form centered on the noun. We would then say that EM seems capable to sort syntactically analogous sentences that are part of different error and age classes in a sufficiently precise way.

NOUN, PROPN and PRON.We can show how children develop a more abstract and adult-like way to referring to entities by pointing out the evolution of the values of PRON and the sum of the values of NOUN and PROPN: for L 0.02 vs 0.49, 0.20 vs 0.79, 0.09 vs 0.79; for M 0.13 vs 0.25, 0.70 vs 0.55, 0.41 vs 0.39; for H 1.14 vs 0.45, 1.48 vs 0.58, 0.74 vs 0.33. It is clear how children progressively learn to properly use pronouns instead of using nouns: this

[141] Parisse C., Le Normand M. T. "How children build their morphosyntax: The case of French". Journal of Child Language, Cambridge University Press (CUP), 27, pp.267-292., (2000)

is reflected and confirmed in the fact that sentences are on average longer and thus children use anaphora in order to avoid the repetition of the noun or proper noun to indicate the main subject of the sentence. These results would seem to be in line with current literature on the acquisition of pronouns in French[142].

All this would be explained also by the fact that "from first words until the age of 4, children usually tend to repeat simple words or sentences they hear from their parents" (Tomasello, 2003 : 173) , once this form is fixed, then children start to express variation based on this initial form to change the meaning of the sentence, but they do not reformulate the sentence through different words by keeping the original meaning unchanged (Martinot C., 2010 ).

This means that learning is initially "rigid" and children around three years old tend to repeat adult schemes instead of using it creatively:

"Before about 3 years of age, very few children who hear a novel verb used in one linguistic construction can then use that verb creatively in another linguistic construction " (Tomasello, 2003).

At the same time, children do continue to utter protolinguistic form after 2 years-old (Dodane, 2010), so what said so far is only partially true.

To conclude, before the age of 4 children do not show the paraphrastic competence

There are of course exceptions to these grouping tendencies but, besides that, we would suggest that these preliminary results represent a fair attempt to visualize child language development through clusters of words grouped by several criteria (age, grammatical properties, correct pronunciation). Until now, we can cautiously say that in this first stage of research the EM algorithm can provide us some mild descriptions in the classification of POS tags. In other words, the unsupervised automatic procedure seems to be able to confirm a general grammatical development over time. This because cluster memberships are made up of grammatical categories that are differently learnt at different ages. Next step will be to focus on particular POS tags development over time by scanning every cluster and looking to confirm more specific learning tendencies

---

[142] Morgenstern A., Sekali M. "What can child language tell us about prepositions?". Jordan Zlatev, Marlene Johansson Falck, Carita Lundmark and Mats Andr´en. Studies in Language and Cognition, Cambridge Scholars Publishing, pp.261-275,ffhalshs-00376186, (2009)

# Chapter 10 - Comparison between Adrien and Madeleine

On the one hand, Adrien and Madeleine datasets have been collected complying to the same protocol and they have been transcribed by following the same convention, so the two corpus should be comparable. On the other hand, the timing of monthly videorecordings is not exactly the same, some lags that do not get over one month can be found. In addition, to assure a perfect overlap, we were obliged to cut off both datasets after 3.69 (age) because, after this point in age, data in Madeleine are available only in "pho".

Then, we decided to divide both *corpora* in four class times in order to make CHAID derived decision tree more readable, this helps to find in them patterns and tendencies while looking for significative relations both *intra-corpus* and *inter corpora*

| total_days | Adrien | Madeleine | Total | year | class_time |
|---|---|---|---|---|---|
| 370 | 0 | 1 | 1 | 1,01 | |
| 406 | 0 | 19 | 19 | 1,11 | |
| 441 | 0 | 9 | 9 | 1,21 | |
| 455 | 0 | 16 | 16 | 1,25 | |
| 476 | 0 | 3 | 3 | 1,30 | |
| 507 | 0 | 19 | 19 | 1,39 | |
| 555 | 0 | 184 | 184 | 1,52 | |
| 597 | 0 | 246 | 246 | 1,64 | |
| 647 | 0 | 302 | 302 | 1,77 | |
| 682 | 0 | 433 | 433 | 1,87 | |
| 719 | 0 | 224 | 224 | 1,97 | 1 |
| 720 | 17 | 0 | 17 | 1,97 | 1 |
| 745 | 63 | 0 | 63 | 2,04 | 1 |
| 763 | 0 | 494 | 494 | 2,09 | 1 |
| 773 | 183 | 0 | 183 | 2,12 | 1 |
| 792 | 41 | 0 | 41 | 2,17 | 1 |
| 798 | 0 | 479 | 479 | 2,19 | 1 |
| 814 | 251 | 0 | 251 | 2,23 | 1 |
| 828 | 0 | 394 | 394 | 2,27 | 1 |
| 851 | 250 | 0 | 250 | 2,33 | 2 |
| 869 | 0 | 523 | 523 | 2,38 | 2 |
| 880 | 316 | 0 | 316 | 2,41 | 2 |
| 897 | 0 | 522 | 522 | 2,46 | 2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | 907 | 376 | 0 | 376 | 2,48 | 2 |
| | 926 | 0 | 387 | 387 | 2,54 | 2 |
| | 954 | 0 | 446 | 446 | 2,61 | 2 |
| | 965 | 212 | 0 | 212 | 2,64 | 2 |
| | 983 | 0 | 420 | 420 | 2,69 | 2 |
| | 990 | 319 | 0 | 319 | 2,71 | 3 |
| | 1022 | 169 | 0 | 169 | 2,80 | 3 |
| | 1025 | 0 | 548 | 548 | 2,81 | 3 |
| | 1054 | 283 | 0 | 283 | 2,89 | 3 |
| | 1060 | 0 | 165 | 165 | 2,90 | 3 |
| | 1082 | 184 | 0 | 184 | 2,96 | 3 |
| | 1090 | 0 | 40 | 40 | 2,99 | 3 |
| | 1110 | 465 | 0 | 465 | 3,04 | 3 |
| | 1123 | 0 | 406 | 406 | 3,08 | 3 |
| | 1139 | 324 | 0 | 324 | 3,12 | 4 |
| | 1169 | 433 | 0 | 433 | 3,20 | 4 |
| | 1190 | 0 | 429 | 429 | 3,26 | 4 |
| | 1200 | 240 | 0 | 240 | 3,29 | 4 |
| | 1233 | 196 | 0 | 196 | 3,38 | 4 |
| | 1264 | 330 | 0 | 330 | 3,46 | 4 |
| | 1289 | 0 | 437 | 437 | 3,53 | 4 |
| | 1348 | 517 | 0 | 517 | 3,69 | 4 |
| | 1383 | 310 | 0 | 310 | 3,79 | |
| | 1418 | 324 | 0 | 324 | 3,88 | |
| | 1449 | 396 | 0 | 396 | 3,97 | |
| | 1475 | 584 | 0 | 584 | 4,04 | |
| | 1503 | 334 | 0 | 334 | 4,12 | |
| | 1537 | 473 | 0 | 473 | 4,21 | |
| | 1579 | 624 | 0 | 624 | 4,33 | |
| **Total** | | 8214 | 7146 | 15360 | | |

Table 15 Comparison between Adrien and Madeleine

On the basis of these results, we calculated the following statistics to have an overall picture of the development of the two children. Statistics are obtained by using SPSS data analysis software. It is clear that Madeleine has an impressive development that is far better than Adrien since the very beginning: in every variable taken into account she shows greater values. In the last time span it is possible to observe how the difference between the two has decreased in favour of Adrien, and graphs below confirm this tendency in the different variables plotted.

| c_time | Child | | pho_total_words | mod_total_words | CHI_total_words_tokenized | CHI_total_distinct_words | SPV | SPVR | levenshtein_distance |
|---|---|---|---|---|---|---|---|---|---|
| **1.97 - 2.27** | Adrien | **Mean** | **1.29** | **1.27** | **1.32** | **1.11** | **0.93** | **73.5135** | **2.2414** |
| | | N | 555 | 555 | 555 | 555 | 555 | 555 | 555 |
| | | Std. Deviation | 0.756 | 0.730 | 0.854 | 0.339 | 0.809 | 43.03900 | 2.40399 |
| | Madeleine | **Mean** | **3.35** | **3.35** | **3.43** | **3.22** | **1.15** | **35.3990** | **1.4941** |
| | | N | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 |
| | | Std. Deviation | 2.575 | 2.575 | 2.413 | 2.119 | 1.307 | 34.08477 | 1.88361 |
| **2.33 - 2.69** | Adrien | **Mean** | **1.75** | **1.78** | **1.79** | **1.49** | **0.92** | **50.3254** | **1.9818** |
| | | N | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 |
| | | Std. Deviation | 1.186 | 1.265 | 1.228 | 0.756 | 1.106 | 45.41475 | 2.73562 |
| | Madeleine | **Mean** | **4.39** | **4.39** | **4.44** | **4.21** | **1.36** | **33.6807** | **1.7501** |
| | | N | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 |
| | | Std. Deviation | 3.557 | 3.557 | 3.375 | 3.075 | 1.632 | 33.25323 | 2.81313 |
| **2.71 - 3.08** | Adrien | **Mean** | **2.05** | **2.03** | **2.21** | **1.98** | **1.11** | **55.0505** | **2.1380** |
| | | N | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 |
| | | Std. Deviation | 1.313 | 1.282 | 1.496 | 1.272 | 1.023 | 41.82774 | 2.33534 |
| | Madeleine | **Mean** | **4.90** | **4.90** | **4.87** | **4.61** | **1.58** | **33.0365** | **2.0112** |
| | | N | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 |
| | | Std. Deviation | 3.762 | 3.762 | 3.480 | 3.170 | 1.652 | 30.12904 | 2.76401 |
| **3.12 - 3.69** | Adrien | **Mean** | **2.91** | **2.87** | **3.16** | **2.84** | **1.41** | **48.7546** | **2.5824** |
| | | N | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 |
| | | Std. Deviation | 1.958 | 1.930 | 2.146 | 1.871 | 1.310 | 37.42604 | 2.66654 |
| | Madeleine | **Mean** | **4.91** | **4.91** | **5.01** | **4.72** | **1.86** | **40.8909** | **2.4462** |
| | | N | 865 | 865 | 865 | 865 | 865 | 865 | 865 |
| | | Std. Deviation | 4.382 | 4.382 | 4.107 | 3.649 | 1.923 | 33.01668 | 3.43622 |

Table 16 Comparison between Adrien and Madeleine (indexes)

In the two following graphs we have plotted the number of total words counted in pho tiers and then in mod tiers. They seem to be almost identical: they do differ only by the fact that in pho there are slightly more occurrences than in mod because verbal behaviours such as repetitions, hesitations and other rare phenomena for whose is difficult to find a clear transcription have not been transcripted in exactly the same way.

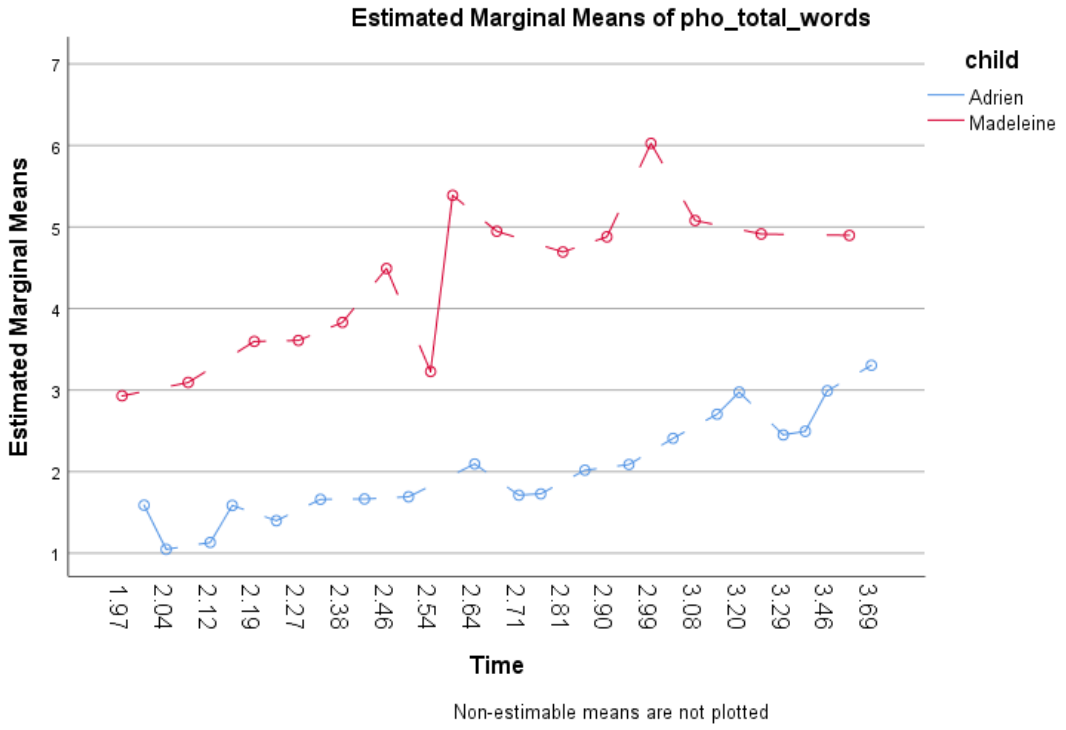



Figure 27

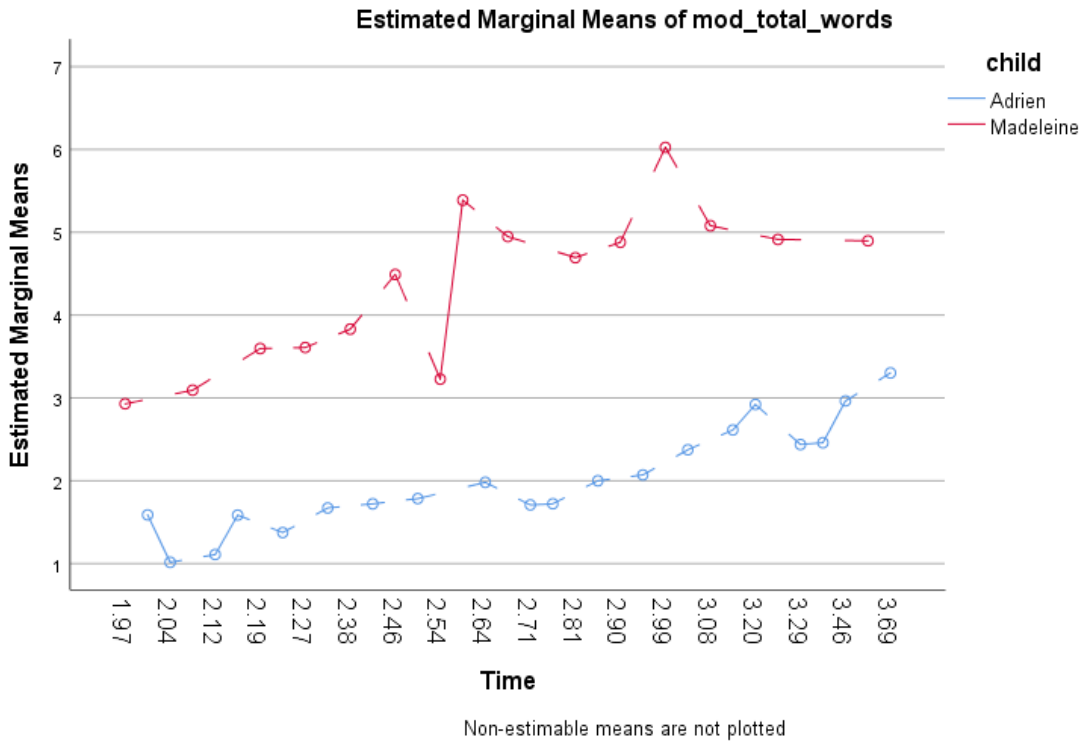



Figure 28

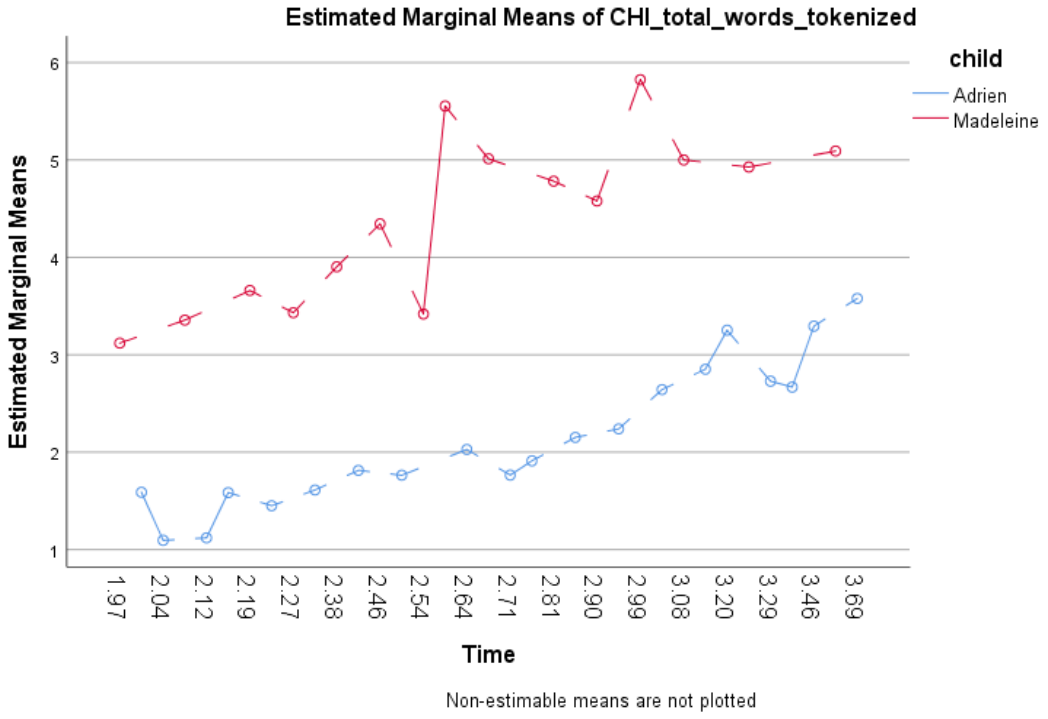



Figure 29

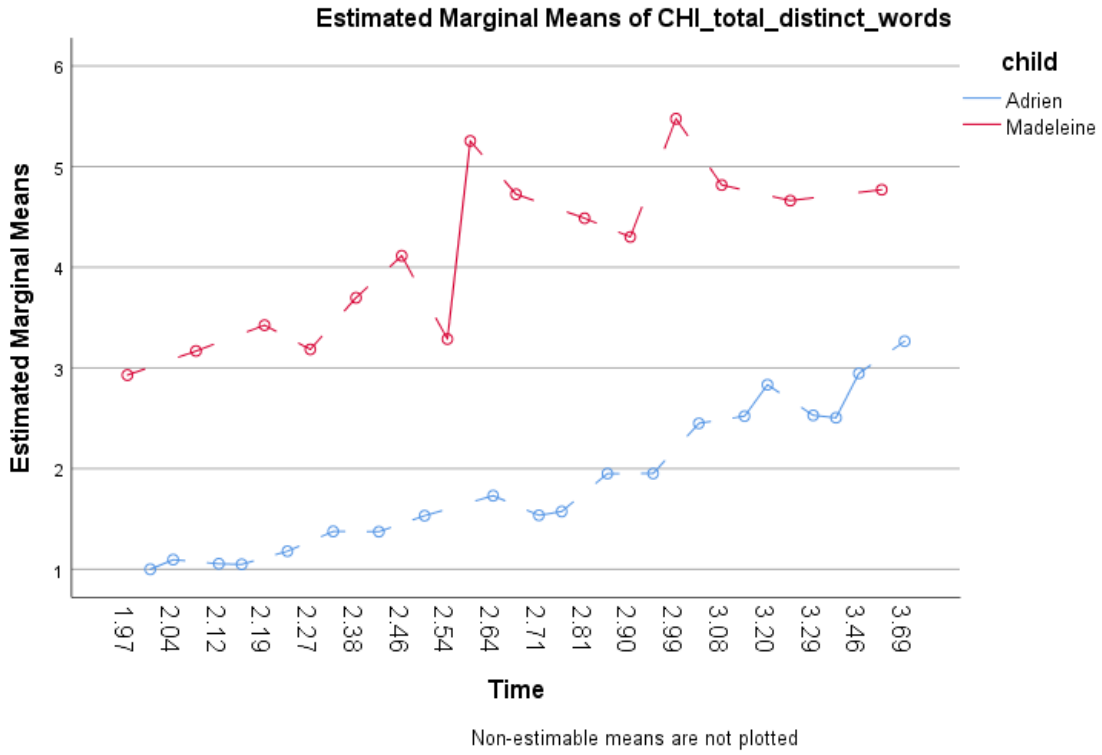



Figure 30

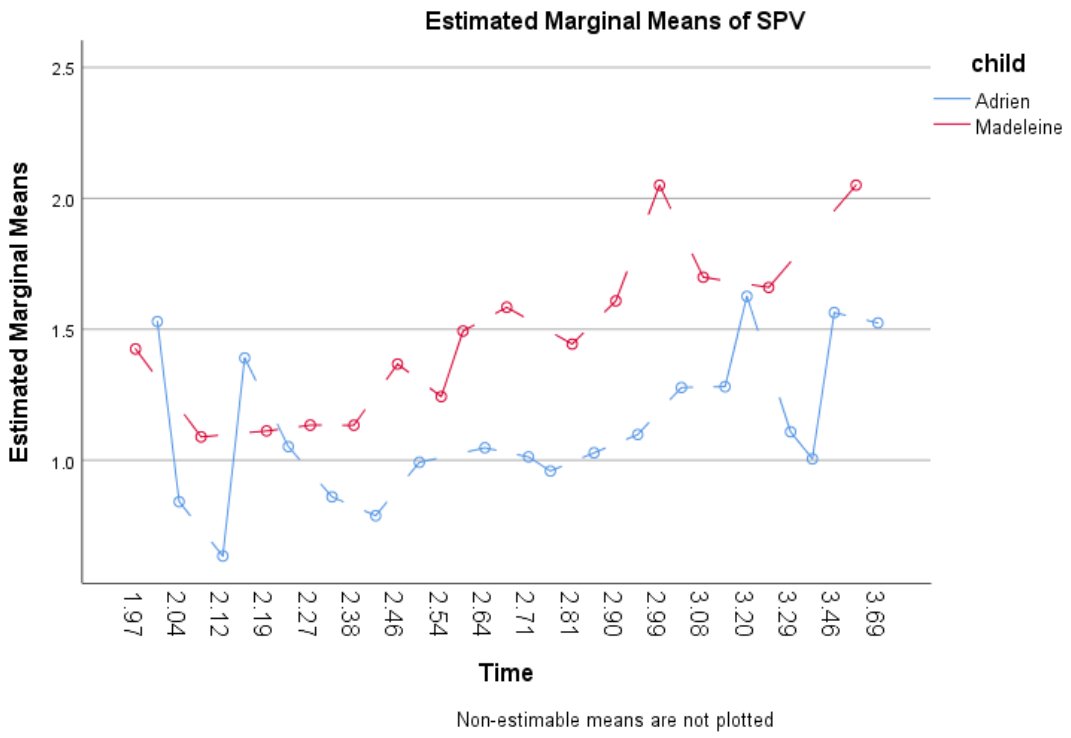



Figure 31

In the graph below we have plotted SPVR for both children: we can clearly see a global decrease of variation rate in Adrien, while for Madeleine there seems to be no improvement at all except for the very beginning. It is hard to give a direct answer to this: if we look at the two previous graphs “Estimated marginal means of CHI_total_distinct_words” and “Estimated marginal means of CHI_total_words_tokenized” we can probably explain this quirk by considering that Madeleine performs language skills that are almost twice as “better” as Adrien. If Madeleine utters richer and longer sentences she is likely to be more exposed to phonemes hard to be properly pronounced at three years old. But this is not so straightforward as it would seem to be.

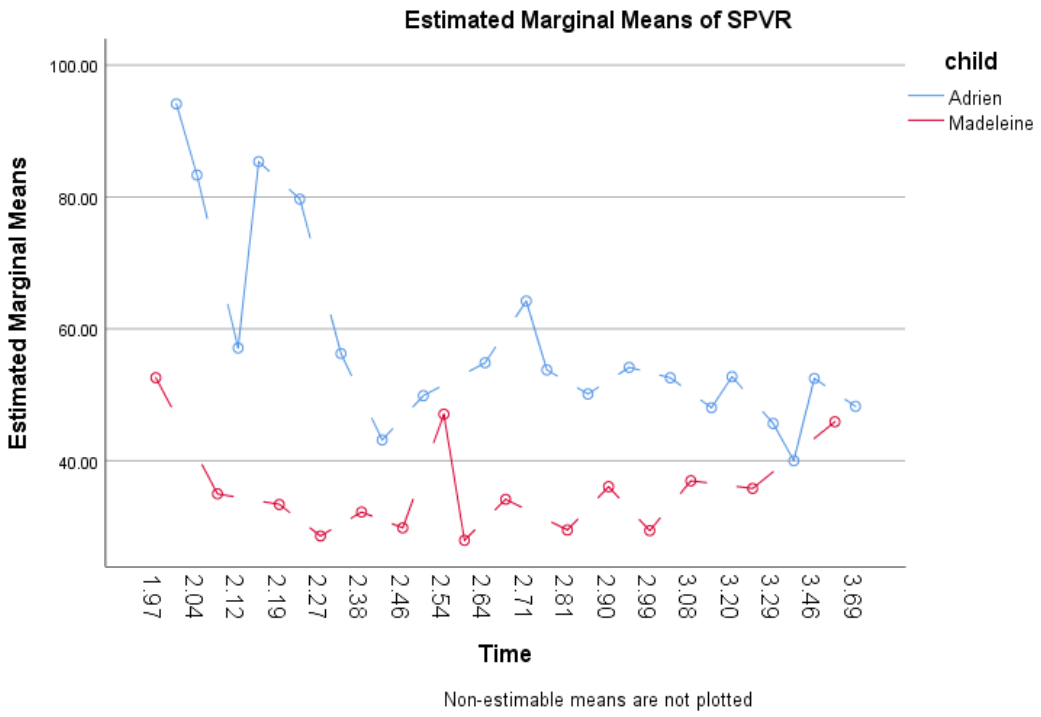



Figure 32

In this graph Levenshtein Distance is not normalized: this means that it gives an absolute number that does not take into account sentences and word length. For this reason values are highly variables (huge standard deviations) and increase over ages instead of decreasing: if sentences are longer, LD will be proportionally longer despite the child has improved his/her language skills. If the child would pronounce every word perfectly, as an adult, LD would be 0, but if a child at 2;0;0 says incorrectly a consonant in a word consisting of 5 graphemes, LD will be 1, and if the same child miss two consonants in a five-words long sentence, LD will then be 2. This could be considered in some ways, and in fact it is, but LD is considered a basic objective string metric in computational linguistics research. In any case, Normalized Levenshtein distances for different ages are provided in "EM Clustering" Chapter. In this case, it is possible to observe that LD distances follow a temporal evolution similar to the SPVR one. We provide LD on specific highly frequent words in CoLaJE *corpora* too. As these graphs are thought to be interactive, we think it is not worthwhile to paste them here in this thesis and we just provide a link to them: https://marine27.github.io/TER/site_aquisition_du_langage/distance_dl.html

We claim that highly frequent words are learnt before less frequent ones even if they have a similar perceptive/articulatory difficulty: this because it is possible that priming effects are at play in the positive retroactive feedback between child-directed speech and child's output.

It is difficult to evaluate the role frequency effects play in learning, this because

"a frequency-sensitive learning mechanism need not (and most probably does not) entail a mechanism that computes and matches the frequency of various elements in the "input" or acquires knowledge of frequency"[143].

So, we can conclude by saying that

"High-frequency forms are (ii) early acquired and (iii) prevent errors in contexts where they are the target, but also (iv) cause errors in contexts in which a competing lower-frequency form is the target" (Ibidem, p 240)

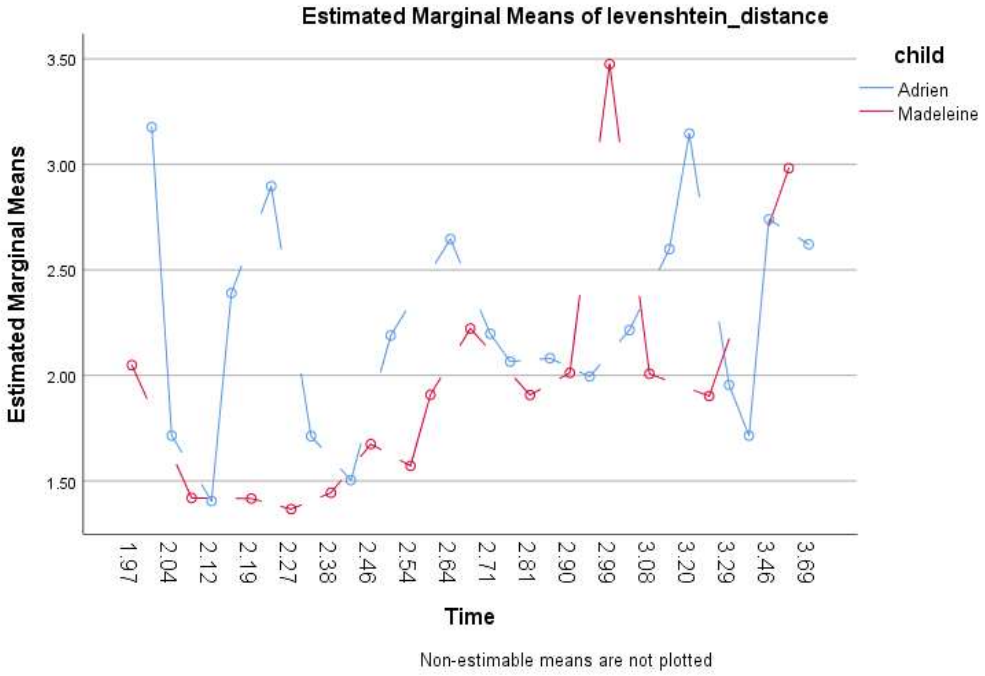



Figure 33

[143] Ambridge B.; Kidd E.; Rowland C. F; Theakston A. (2015). "The ubiquity of frequency effects in first language acquisition".Journal of Child Language. Cambridge University Press. 42. 239 – 273.

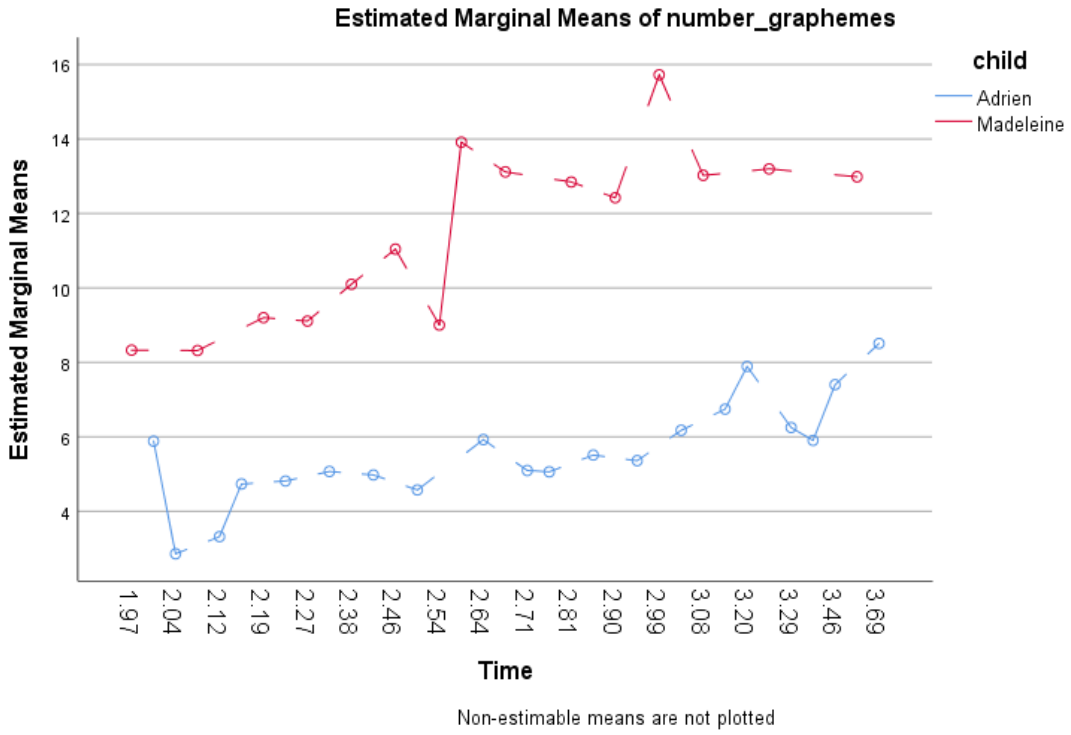



Figure 34

The graph below shows the evolution of liquid consonants such as “r” and “l”, classified as “c-liquides” according to the table showed in Annex 5 (where French phonemes are listed according to an articulatory effort principle). We can observe how there is an overall increase quantitatively similar for the two children. In both cases, it is possible to see how the regression line goes up and down many times: we think that this is due to the undergoing process of the building of a phonological consciousness. As showed in the example of Albane pronouncing “tracteur” in many different ways in a bunch of seconds (see chapter on Phonetics and Phonology), acquisition of a consonantal minimal pair such as /r/ - /l/ opposition is a skill that requires time to be learnt. This is confirmed by the other two annexes regarding acquisition order of consonants

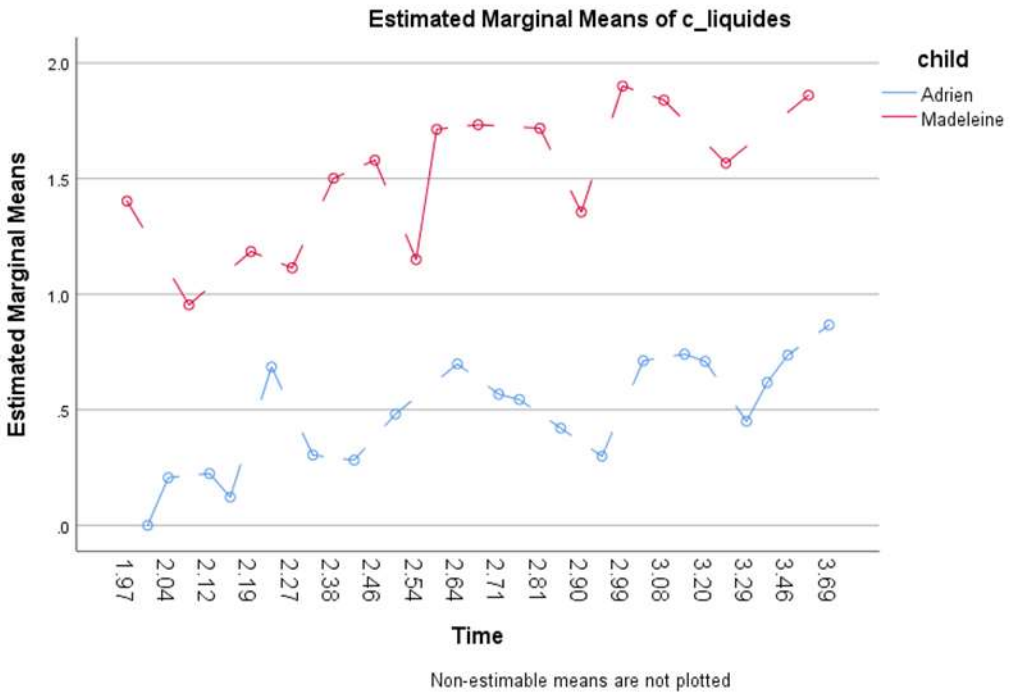



Figure 35

This analysis is phonetically less precise compared to a similar one made by Yamaguchi in her Phd thesis (see Annex A, p 309. Yamaguchi N., 2012) because she analysed the emergence and the role of the same opposition /r/ and /l/ adding the position occupied by these phonemes in either onset or coda. This is important because the same phoneme can be differently pronounced depending on the place it has in a given word or syllable.

For example, a child can properly pronounce the first form of these two couples, but not the second one, despite the syllable “clé” and “ra” are the same

“clé” and “bouclé”, “rat” and “train”

The position a phoneme occupies in a syllable is very important for its pronunciation and it must be taken in account, it is hard to create a script that is sensitive to this variation.

For an advanced phonological analysis I should have used PHON[144], (I would have had the opportunity to learn this software by attending a series of classes in Sorbonne University last spring, unfortunately this was not possible due to the pandemic).

This software provides multiples ways to analyse transcription in smaller parts (syllables, onset -nucleus-coda) by relying on a specific dictionary. By doing so, it is possible to consider the position of a phoneme and its consequent expected articulation (strong vs weak position and stress position)

Realising the potential of this software while writing the final chapters of this thesis it is a little bit frustrating. The only thing I can say is that learning it will be the next step of my research once this thesis will be finished.

I think that « automatic syllabification » and « automatic alignment function » would have given me the opportunity to save a lot of time spent in cleaning and filtering data and, at the same time, results would have been more accurate. I am realizing that I did manually (or by writing scripts) tasks that were already ready to use in PHON:

"Once the researcher has identified the domains of analysis, segmentation at the level of the syllable is performed automatically: Segments are assigned descriptive syllable labels (visually represented with colors) such as 'onset' or 'coda' for consonants and 'nucleus' for vowels. The program also identifies segmental sequences within syllable constituents (e.g. complex onsets or nuclei). Since controversy exists in both phonetic and phonological theory regarding guidelines for syllabification, the algorithm is parameterized to allow for analytical flexibility. The availability of different parameter settings also enables the researcher to test hypotheses on which analysis makes the best predictions for a given dataset[145]".

As well as for the automatic alignment

"After syllabification, a second algorithm performs automatic, segment-by-segment and syllable-by-syllable alignments of target and actual forms. Building on featural similarities

---

[144] Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G., Maddocks, K., O'Brien, P., & Wareham, T. (2006). Introducing *Phon*: A Software Solution for the Study of Phonological Acquisition. *Proceedings of the Annual Boston University Conference on Language Development. Boston University Conference on Language Development, 2006*, 489–500.

[145] Rose Y. et al., 2006, p7

and differences between the segments in each syllable and on syllable properties such as stress, this algorithm automatically aligns corresponding segments and syllables in target and actual forms[146]".

I think that I should see the glass half full: losing time with tables and scripts allowed me to train myself on formats and algorithms. In any case, if possible, I will try to present an application of PHON during the thesis defence, as it will be in a couple of months.

Going back to what we effectively done, in "Data mining" (chapter 11) , we have tried to create a series of algorithms to tackle this question by drawing inspiration from the previous work of computer scientists Agrawal & Skrikant[147] regarding itemsets and sequential pattern mining. This attempt has unfortunately unmatched the initial expectation as results obtained take in consideration the place a phoneme occupies in a given word (coded as "ph_début"; "ph_milieu", "ph_fin") but are almost impossible to interpret. Sequences obtained are quite messy and finding a correlation from them has been – at least for me – impossible. https://marine27.github.io/TER/site_aquisition_du_langage/pattern_mining1.html *********

To get an overview of the overall evolution of liquids see the corresponding graphs at the "proportion phonétique" window in this link https://marine27.github.io/TER/site_aquisition_du_langage/stackgraph.html

While to get an overview of /r/ and /l/ singularly, see the corresponding graphs at the "chronologie phonétique" window in this link https://marine27.github.io/TER/site_aquisition_du_langage/multistream.html

Data are both cases available for the six children part of CoLaJE.

To test whether the counts done by Yamaguchi and the counts done in this thesis are coherent between each other, considering that the procedure to get the value seems to be different, let's compare some values: the huge difference at 1.97 point time (x axis) correponds to Yamaguchi's table at the following values:

Adrien onset "l"= 0 "r"= 0 coda "l"=0 "r"= 0

[146] Rose Y. et al., 2006, p8

[147] Srikant R., Agrawal R., « Mining Sequential Patterns: Generalizations and Per-formance Improvements », Proceedings of the 5th International Conference on ExtendingDatabase Technology (EDBT'96), Avignon, France, September 1996, p. 3-17

Madeleine onset "l" = 133 "r"= 29 coda "l"= 27 "r"= 63

A glimpse on all the other values show that Madeleine's values are always higher than Adrien's values, we would then say that we could rely on these data because that they should correspond to approximately similar precise counting procedure leading to comparable values.

French is a language where "open" syllables such as CV and CCV sequences are more frequent (around 3/4 [148]) than "closed" syllables such as VC or CVC. Children are consequently more exposed to open syllables than to closed ones, so children should tend to display a higher variation rate in words containing closed syllables and, because of little input exposure, children need more time to master their articulation them: it is possible to observe how CHAID is partly able to partition the dataset in its SPVR form by creating segments containing the two differing syllables sequences (see chapter on CHAID)

By using the label "semico_difficile" we are referring to consonants considered to have a high articulatory effort according to the table showed in annex 5.

The opposition is the following: /ɥ/ and /w/ , to give an example is what distinguishes the first syllable fo the word "huile" from "oui". It is a slight difference: an Italian speaker of French L2 is not readily able to perceive and identify it during the first times as in Italian sound system there is not such a difference.

In this graph we cannot know if this couple of phonemes are correctly pronounced or not, but we would expect that as their occurrence increases, their rate of successful pronunciation increases too. It is possible to check the empirical validity of this affirmation by observing how an increase in types or in tokens (and of both of course) is followed by a decrease in SPVR too.

As it was for liquids consonants, even in this case Madeleine proves to be able to master her native language better than Adrien since the beginning of recording, but the latter reduces this difference again in the last sessions taken in exam.

[148] Sauvage J. Personal communication

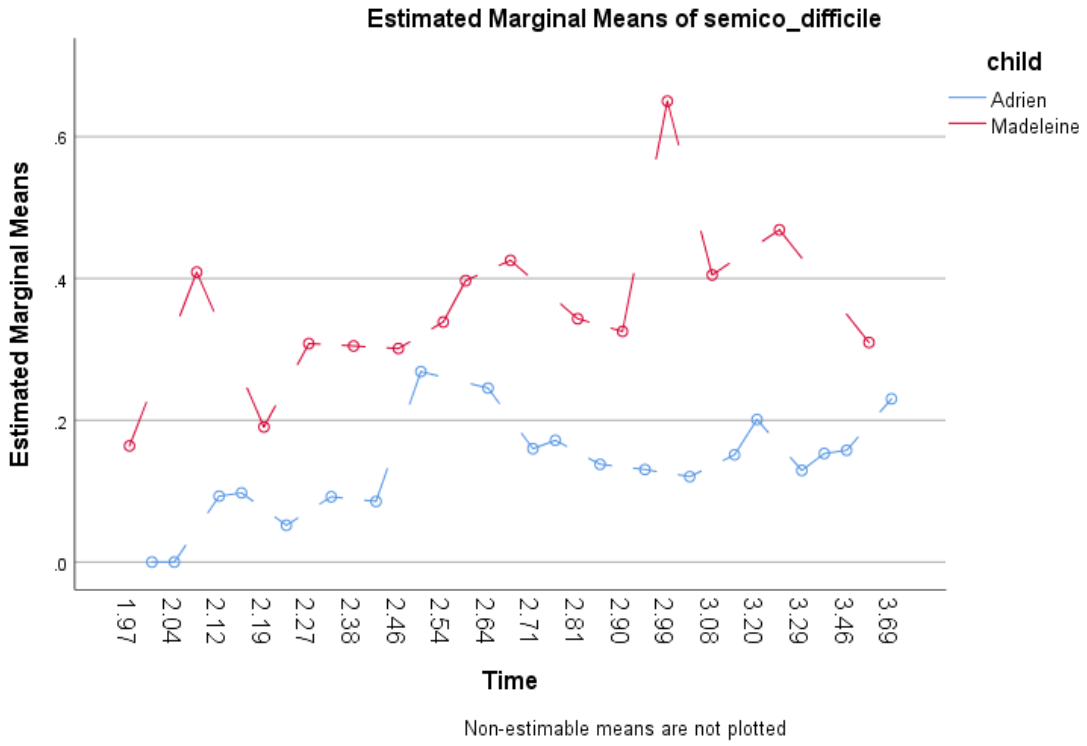

Estimated Marginal Means of semico_difficile
child
Adrien
Madeleine
Estimated Marginal Means
.6
.4
.2
.0
1.97
2.04
2.12
2.19
2.27
2.38
2.46
2.54
2.64
2.71
2.81
2.90
2.99
3.08
3.20
3.29
3.46
3.69
Time
Non-estimable means are not plotted

Figure 36

Table 17 POS tags. Comparison between Adrien and Madeleine

| c_time | | | ADJ | ADP | ADV | AUX | CONJ | CCONJ | DET | INTJ | NOUN | NUM | PART | PRON | PROPN | PUNCT | SCONJ | SYM | VERB | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.97 - 2.27 | Adrien | **Mean** | **0,08** | **0,01** | **0,24** | **0,04** | **0,00** | **0,00** | **0,08** | **0,05** | **0,65** | **0,00** | **0,00** | **0,00** | **0,01** | **0,00** | **0,00** | **0,00** | **0,16** | **0,00** |
| | | N | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 |
| | | Std. Deviation | 0,320 | 0,119 | 0,460 | 0,217 | 0,000 | 0,000 | 0,379 | 0,220 | 0,782 | 0,000 | 0,000 | 0,042 | 0,119 | 0,000 | 0,000 | 0,042 | 0,433 | 0,000 |
| | Madeleine | **Mean** | **0,12** | **0,27** | **0,25** | **0,26** | **0,05** | **0,00** | **0,45** | **0,14** | **0,65** | **0,01** | **0,00** | **0,69** | **0,05** | **0,00** | **0,02** | **0,01** | **0,54** | **0,15** |
| | | N | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 |
| | | Std. Deviation | 0,353 | 0,540 | 0,487 | 0,479 | 0,231 | 0,000 | 0,665 | 0,360 | 0,903 | 0,125 | 0,061 | 0,935 | 0,241 | 0,000 | 0,140 | 0,122 | 0,700 | 0,572 |
| 2.33 - 2.69 | Adrien | **Mean** | **0,07** | **0,02** | **0,43** | **0,03** | **0,03** | **0,00** | **0,09** | **0,19** | **0,55** | **0,01** | **0,00** | **0,12** | **0,06** | **0,00** | **0,01** | **0,00** | **0,19** | **0,01** |
| | | N | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 |
| | | Std. Deviation | 0,269 | 0,134 | 0,758 | 0,174 | 0,213 | 0,000 | 0,350 | 0,538 | 0,708 | 0,117 | 0,059 | 0,365 | 0,512 | 0,000 | 0,106 | 0,029 | 0,436 | 0,097 |
| | Madeleine | **Mean** | **0,14** | **0,40** | **0,41** | **0,26** | **0,15** | **0,00** | **0,50** | **0,12** | **0,72** | **0,04** | **0,00** | **0,74** | **0,08** | **0,00** | **0,11** | **0,00** | **0,73** | **0,30** |
| | | N | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 |
| | | Std. Deviation | 0,368 | 0,660 | 0,667 | 0,506 | 0,440 | 0,000 | 0,686 | 0,328 | 0,950 | 0,237 | 0,042 | 0,935 | 0,294 | 0,000 | 0,363 | 0,062 | 0,790 | 0,772 |
| 2.71 - 3.08 | Adrien | **Mean** | **0,12** | **0,12** | **0,39** | **0,10** | **0,01** | **0,00** | **0,12** | **0,18** | **0,40** | **0,01** | **0,01** | **0,26** | **0,06** | **0,00** | **0,01** | **0,01** | **0,37** | **0,04** |
| | | N | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 |
| | | Std. Deviation | 0,345 | 0,364 | 0,675 | 0,342 | 0,099 | 0,000 | 0,336 | 0,418 | 0,637 | 0,124 | 0,121 | 0,566 | 0,264 | 0,027 | 0,124 | 0,127 | 0,596 | 0,212 |
| | Madeleine | **Mean** | **0,15** | **0,39** | **0,36** | **0,25** | **0,17** | **0,00** | **0,54** | **0,14** | **0,71** | **0,07** | **0,00** | **0,84** | **0,09** | **0,00** | **0,14** | **0,01** | **0,83** | **0,44** |
| | | N | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 |
| | | Std. Deviation | 0,374 | 0,689 | 0,669 | 0,497 | 0,409 | 0,000 | 0,715 | 0,356 | 0,854 | 0,528 | 0,000 | 0,955 | 0,321 | 0,000 | 0,433 | 0,083 | 0,840 | 0,960 |
| 3.12 - 3.69 | Adrien | **Mean** | **0,11** | **0,16** | **0,49** | **0,22** | **0,06** | **0,00** | **0,23** | **0,09** | **0,44** | **0,03** | **0,01** | **0,57** | **0,05** | **0,00** | **0,04** | **0,00** | **0,59** | **0,07** |
| | | N | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 |
| | | Std. Deviation | 0,328 | 0,415 | 0,797 | 0,467 | 0,298 | 0,000 | 0,474 | 0,306 | 0,655 | 0,197 | 0,096 | 0,872 | 0,219 | 0,000 | 0,218 | 0,058 | 0,725 | 0,291 |
| | Madeleine | **Mean** | **0,16** | **0,44** | **0,43** | **0,27** | **0,22** | **0,00** | **0,51** | **0,20** | **0,77** | **0,03** | **0,01** | **0,78** | **0,11** | **0,00** | **0,14** | **0,01** | **0,76** | **0,53** |
| | | N | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 |
| | | Std. Deviation | 0,391 | 0,743 | 0,736 | 0,520 | 0,498 | 0,000 | 0,707 | 0,414 | 0,924 | 0,320 | 0,083 | 1,073 | 0,362 | 0,000 | 0,413 | 0,076 | 0,873 | 1,064 |

In the following graph (see supplementary file named Adrien_vs_Madeleine_2_09_2020) we have calculated the Pearson correlation and the p-value for the relation between SPVR and the phonemes and group of phonemes listed in Annex 5 (Articulatory effort), cell C14 can be read as: when vowels increase of a unit then SPVR increases by a 0.167 factor. C17 can be read as: when consonants increase of a unit, then SPVR increases of 0.264.

C260, Adrien at 2.71 – 3.08 at an increase of bilabials results in a decrease of SPVR, while in C269 an increase in “liquides” causes an increase in SPVR by a factor of 0.196.

| c_time | | | number_graphemes | voyelle | consonnes | groupe_consonatique | semi_consomnes | v relache | v tendu | v nasal | c_occlusives_bilabiales | c_a_occlusives | c_constrictives_precis | c liquides | c_constrictives_large | groupe_consn trdr | groupe_consn kgr | semico_facile | semico_difficile |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.97 - 2.27 | Adrien | Mean | 4,12 | 1,86 | 2,19 | 0,00 | 0,07 | 0,64 | 0,69 | 0,54 | 0,71 | 0,91 | 0,10 | 0,42 | 0,05 | 0,00 | 0,00 | 0,01 | 0,06 |
| | | N | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 | 555 |
| | | Std. Deviation | 2,746 | 1,086 | 1,690 | 0,000 | 0,327 | 0,869 | 1,032 | 0,647 | 1,304 | 1,158 | 0,385 | 0,689 | 0,326 | 0,000 | 0,000 | 0,104 | 0,312 |
| | Madeleine | Mean | 8,78 | 4,05 | 4,25 | 0,00 | 0,40 | 1,75 | 1,74 | 0,56 | 0,94 | 1,06 | 0,92 | 1,13 | 0,20 | 0,00 | 0,00 | 0,11 | 0,28 |
| | | N | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 |
| | | Std. Deviation | 6,689 | 3,078 | 3,551 | 0,000 | 0,615 | 1,708 | 1,779 | 0,850 | 1,256 | 1,306 | 1,065 | 1,379 | 0,502 | 0,000 | 0,000 | 0,365 | 0,514 |
| 2.33 - 2.69 | Adrien | Mean | 5,04 | 2,43 | 2,41 | 0,00 | 0,19 | 0,85 | 0,88 | 0,70 | 0,87 | 0,84 | 0,22 | 0,43 | 0,04 | 0,00 | 0,00 | 0,01 | 0,18 |
| | | N | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 | 1154 |
| | | Std. Deviation | 3,336 | 1,591 | 1,884 | 0,000 | 0,545 | 1,133 | 1,140 | 0,912 | 1,260 | 1,044 | 0,504 | 0,883 | 0,251 | 0,000 | 0,000 | 0,226 | 0,501 |
| | Madeleine | Mean | 11,42 | 5,20 | 5,70 | 0,00 | 0,45 | 2,31 | 2,22 | 0,68 | 1,10 | 1,51 | 1,16 | 1,54 | 0,38 | 0,00 | 0,00 | 0,10 | 0,35 |
| | | N | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 | 2309 |
| | | Std. Deviation | 9,406 | 4,203 | 5,119 | 0,000 | 0,646 | 2,113 | 2,336 | 0,988 | 1,416 | 1,745 | 1,344 | 1,675 | 0,725 | 0,000 | 0,000 | 0,330 | 0,564 |
| 2.71 - 3.08 | Adrien | Mean | 5,56 | 2,76 | 2,80 | 0,00 | 0,21 | 1,13 | 1,20 | 0,43 | 0,61 | 0,89 | 0,42 | 0,55 | 0,13 | 0,00 | 0,00 | 0,07 | 0,14 |
| | | N | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 | 1420 |
| | | Std. Deviation | 3,647 | 1,759 | 2,039 | 0,000 | 0,463 | 1,263 | 1,182 | 0,738 | 0,942 | 1,107 | 0,748 | 0,826 | 0,379 | 0,000 | 0,000 | 0,308 | 0,369 |
| | Madeleine | Mean | 12,95 | 5,96 | 6,35 | 0,00 | 0,55 | 2,26 | 2,78 | 0,92 | 1,22 | 1,78 | 1,28 | 1,71 | 0,35 | 0,00 | 0,00 | 0,18 | 0,37 |
| | | N | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 | 1159 |
| | | Std. Deviation | 9,882 | 4,539 | 5,243 | 0,000 | 0,747 | 2,024 | 2,524 | 1,226 | 1,440 | 1,898 | 1,470 | 1,869 | 0,712 | 0,000 | 0,000 | 0,446 | 0,569 |
| 3.12 - 3.69 | Adrien | Mean | 7,40 | 3,57 | 3,57 | 0,00 | 0,26 | 1,53 | 1,40 | 0,64 | 0,66 | 1,17 | 0,86 | 0,72 | 0,16 | 0,00 | 0,00 | 0,08 | 0,18 |
| | | N | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 | 2040 |
| | | Std. Deviation | 4,891 | 2,341 | 2,614 | 0,000 | 0,521 | 1,458 | 1,451 | 0,855 | 1,034 | 1,186 | 1,039 | 0,979 | 0,432 | 0,000 | 0,000 | 0,324 | 0,433 |
| | Madeleine | Mean | 13,09 | 6,04 | 6,33 | 0,00 | 0,60 | 2,37 | 2,88 | 0,79 | 1,23 | 1,79 | 1,26 | 1,71 | 0,33 | 0,00 | 0,00 | 0,22 | 0,39 |
| | | N | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 | 865 |
| | | Std. Deviation | 11,597 | 5,402 | 5,994 | 0,000 | 0,823 | 2,428 | 2,851 | 1,170 | 1,576 | 2,034 | 1,556 | 1,959 | 0,711 | 0,000 | 0,000 | 0,489 | 0,618 |

Table 18 Phonemes' evolution. Comparison between Adrien and Madeleine

Are these results in line with Morgenstern & Parisse (2012) and, more generally speaking, to current literature?

# Chapter 11 - Data mining

In this chapter are presented and discussed several data mining techniques that have been learnt, conceived and applied on CoLaJE *corpora* thanks to a collaboration with the data science lab of the "Paul-Valéry" University. This close collaboration consisted in weekly meetings, lessons and workshops lasted all the past academic year (from september 2019 until may 2020) and finished with a successful final online exam during on of the last lockdown days.

Since I am not a statistician or a data scientist, it would have been impossible to me to make all these statistics, to use such a diversity of computational and graphic tools and to develop a number of scripts in Python language to address all of the questions arised on how to look for patterns, which was the best (and feasible) level of analysis, which format would have fit the best to the graph available and so on.

For this reason I have chosen to write this chapter in the first plural form, to make explicit that this work derived from an equal effort made by me and four master students (see Aknowledgments) and thanks to the supervision of lab director professor Sandra Bringay and professor Sallaberry.

All the scripts, statistics, graphs and references are available on a GitHub page at the following link https://marine27.github.io/TER/index.html

## 11.1 Restructuring the data with Python

On the CoLaJE site, we have downloaded data child by child, and registration by registration. This represents 236 files. Each of them contains the transcripts of a child's video recording for a period of approximately one hour.

We have downloaded these files in *.chat* format: we thought that this format is better than .csv for the aim we had in mind and for automatic processing too.

In order to be able to use these files, we have converted them to the *.txt* format thanks to the converter proposed by Ortolang[149] . We obtained 236 *.txt* files with an average of 3000 lines each.

We then select data and establish a common structure for all of them. To do this, we set up a model to unify the transcripts:

| Libelle | ENFANT | AGE | TYPE | TEMPS_DEBUT | TEMPS_FIN | CONTENU |
|---|---|---|---|---|---|---|
| Typage (format) | string | timestamp | string | Int | Int | string |
| Obligatoire | oui | oui | oui | facultatif | facultatif | oui |

Figure 37 Data dictionary

Description of the fields :

- CHILD : The name of the child.

- AGE : The age of the child to allow us to apply treatments according to time.

- TYPE: The type of the transcription, to allow us to apply treatments according to the transcriptions.

- START TIME: The start time of the transcription on the video in seconds (for example 200 seconds after the start of the video recording).

- END TIME: The end of the transcription on the video in seconds (e.g. 200 seconds after the beginning of the video recording).

- CONTENT: The content of the transcript.

By grouping the .txt files by children and applying the above model, we obtain 7 .csv files containing the harmonized data.

Here below an example:

[149] http://ct3.ortolang.fr/teiconvert/index-en.html URL consulted on 22/10/2020

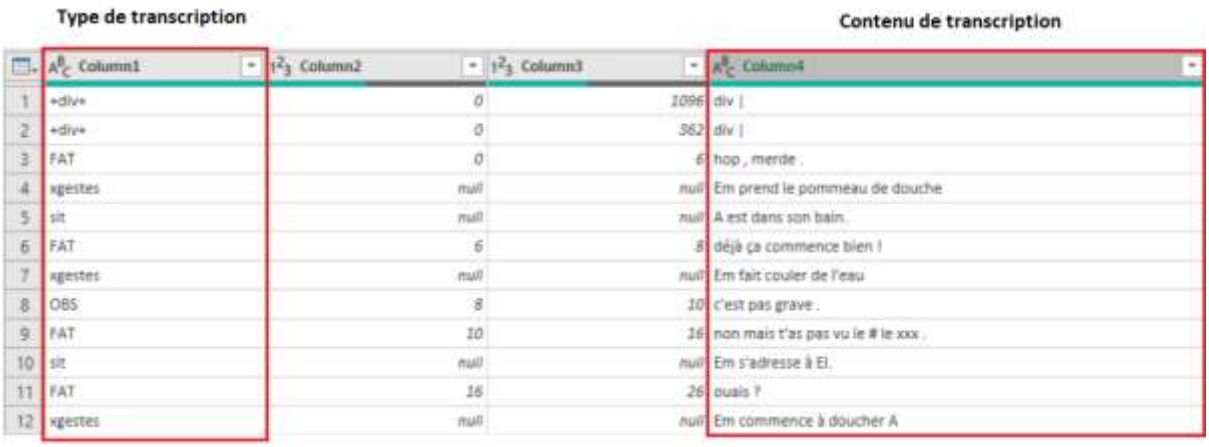



Figure 38 : Example of the .csv file

We have decided not to keep into account the dataset of the child called Léonard. This dataset contains very few transcriptions and has not been produced under the same conditions. This choice should avoid potential bias in the results and allows final results to be fully comparable.

PHO-type transcripts were 'cleaned' by taking off all the non-verbal symbols (i.e those contained in % lines) and encoding work was carried out to replace particular phonemes with a numeric code as it was difficult to allow algorithms recognise them. Plosive-liquids phonemes were coded too because the project would like to focus especially on these occurrences.

```
Dic={'0':'ɑ̃','1':'ɔ̃','2':'ɛ̃','3':'œ̃','4':'tr','5':'kr','6':'dr','7':'gr'}
```

Figure 39 - Phoneme coding

## 11.2 Phonemes proportions (stackplot)

We first carried out an exploratory analysis to extract a global vision of the data and then carried out two more in-depth studies, the first based on “Pattern Mining techniques” to extract phonetic subsequences to explore the phonetic/phonological evolution of children and

the second based on vectorial phonemes' representation and articulatory features to highlight the particularities of language acquisition.

By using Jupyter Notebook, we analyse the proportion of phonemes for each child through these following steps:

- Data source: The phonetic transcriptions were taken from the previous elaboration treatment: in this form characters are coded with the International Phonetic Alphabet.

- Phoneme extraction: we use several Python libraries such as *pandas*, *numpy*, *joblib* and *counter*. To help the Python interpreter we need to give a specific alphanumerical code to certain phonemes that the programming language was not able to recognise.

- Phonemes grouping: The phonemes have been organised according to the two-level hierachy described in the following table:

| Voyelles | | | Consonnes | | | | | Groupes conson-antiques | Semi-consonnes | |
|---|---|---|---|---|---|---|---|---|---|---|
| Relâcheés | Tendues | Nasales | Occlusives bilabiales | Constructives avec un point d'articulation précis | Constructives avec un point d'articulation large | Liquides | Autres occlusives | Oppositions particulières | Facile | Oppositions difficiles |
| a | i | ɑ̃ | b | f | l | ʒ | d | tʁ | j | ŋ |
| ɑ | e | ɔ̃ | p | v | ʁ | | t | kʁ | | w |
| ɛ | ø | ɛ̃ | m | s | | | k | dʁ | | |
| ɔ | ə | œ̃ | | z | | | g | gʁ | | |
| œ | o | | | | | | n | | | |
| | y | | | | | | ɲ | | | |
| | u | | | | | | | | | |

Figure 40 Phonetic units hierarchy (see annex 5)

- Standardization: The measured quantities of phonemes were normalized to obtain the proportion of each phoneme group according to each recording:

$$P_{pho} = \frac{N_O}{N_T}$$

(8)

In the formula we can observe:

- Ppho : Proportion of phonemes;

- NO : Occurrences of one phoneme per recording;

- NT : Total number of phonemes during the recording.

Finally, the data was aggregated and saved in a file for each child. The results of this study are shown below.

Let's look at the results produced by the Madeleine and Théophile children. The results of a previous study on the quantity of phonemes over time showed that the two children had a noticeable evolution in the quantitative number of phonemes, but that there were proportionally dissimilarities. results For example, Madeleine early develops many different phonetic groups, whereas Théophile has only a few. But from a broad perspective, the study showed similarities:

- The “constrictive larges” (fricatives in English) group increases gradually;

- Constrictives (fricatives) increase abruptly;

- Nasal vowels and tense vowels increase and then decrease;

- Semi-consonnes difficiles (ifficult semi-consonants) increase the equation.

However, when reading the “Stackplot” from the analysis of the proportions of phonemes produced on the next page (figure 13, figure 14), one can notice some markers that had passed under track:

- The “bilabial plosives” are less and less present until they have a constant proportion. One might therefore wonder whether the acquisition of these phonemes is not very early and marks their definitive presence in language.

- Vowels gradually decrease

- Madeleine develops the acquisition of phonetic groups much faster. Some of them are even present very early in her evolution, whereas in Théophile, the evolutions seem to be progressive. In Madeleine's case, learning is much more abrupt (although she has developed a much more varied lexicon).

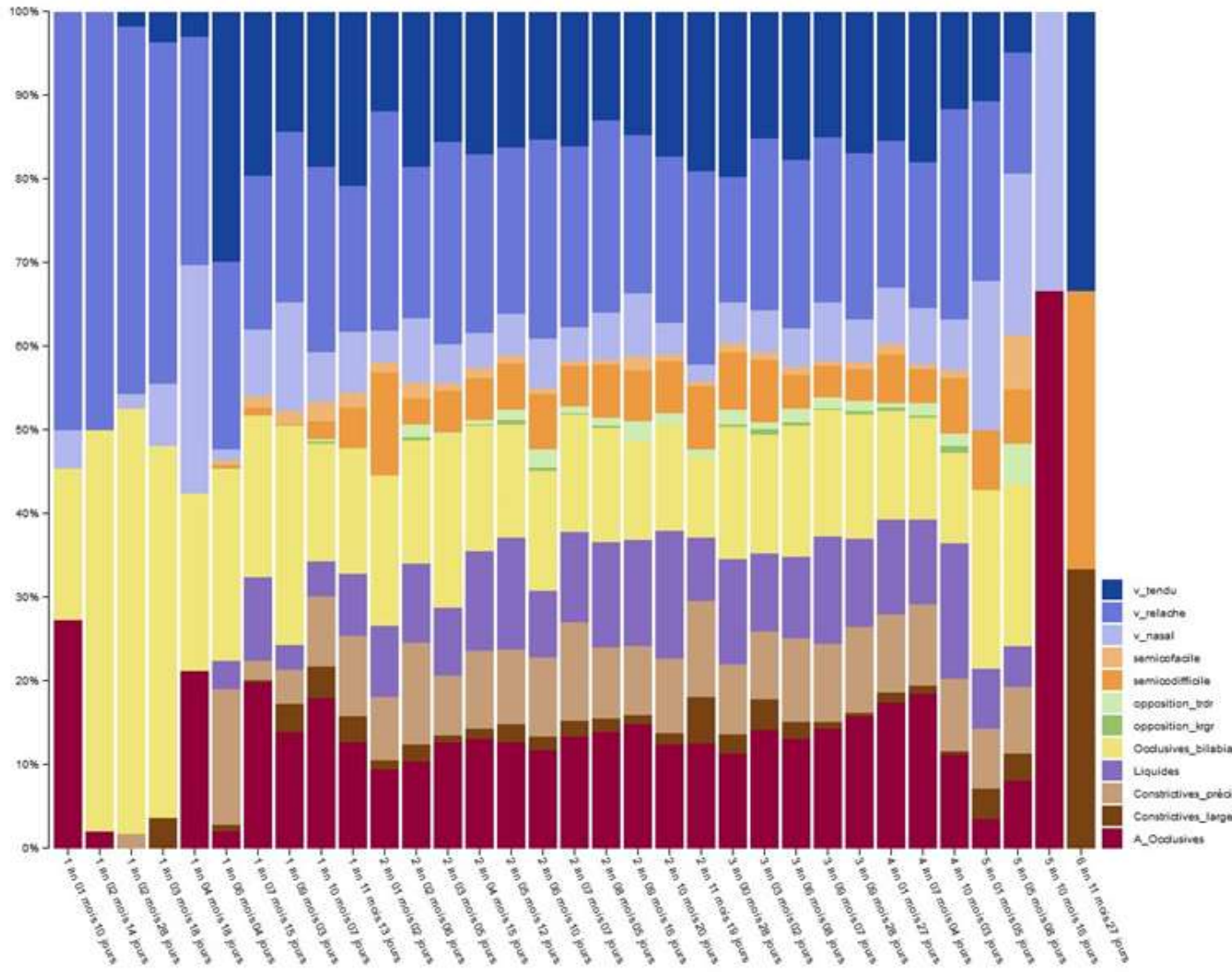



Figure 41 Histogram showing Madeleine's phonemes' evolution (relative values)

Here below the same version of the graph showing absolute values:

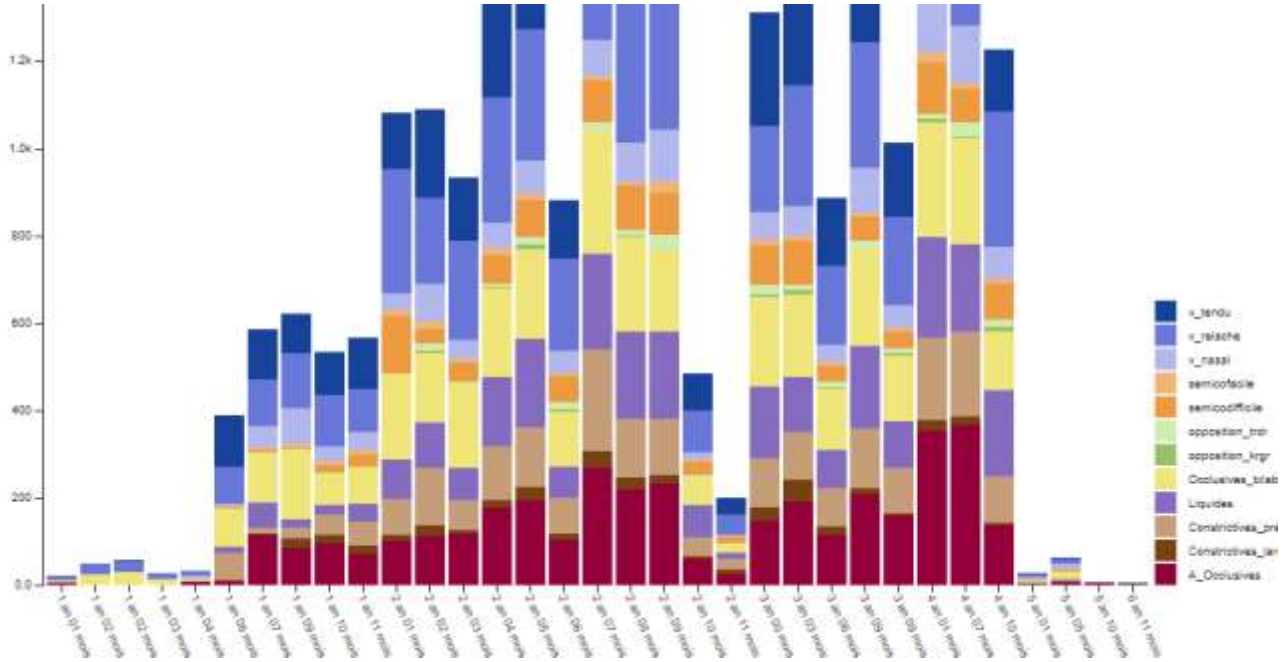


Figure 42 Histogram showing Madeleine's phonemes' evolution (absolute values)

Here below are Théophile's relative values

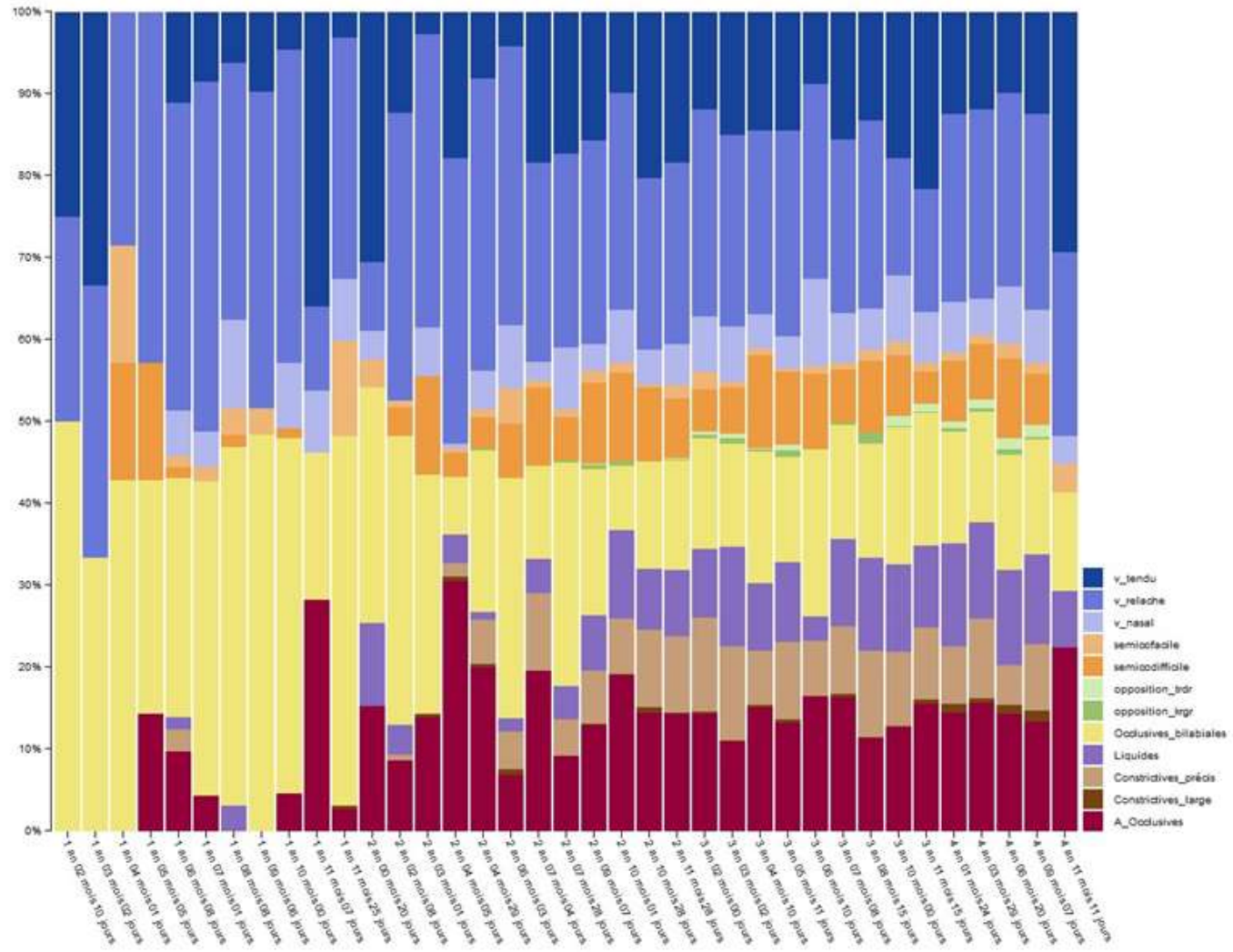



Figure 43 Histogram showing Théophile's phonemes' evolution (relative values)

https://marine27.github.io/TER/site_aquisition_du_langage/stackgraph.html

These results are coherent to the tables presented in Yamaguchi's thesis (Yamaguchi, 2012, pp309-312).

We then looked at the evolution of the number of pronounced phonemes over time. To do this, we used an interactive tool: the Multistreamgraph.

## 11.3 Multiresolution Streamgraph

We used a Streamgraph visualization to represent the temporal evolution of multiple time series, *i.e.* a set of multiple quantitative variables occurring in a same time temporal interval and interacting with each other.

We have chosen to use the representation developed by professor Sallaberry and colleagues (Cuenca A., 2018) because it has the advantage of being able to represent several time series organised in a hierarchical structure which facilitates the exploration and comparison of every stream with the others as well as to explore any given time window through different temporal granularities. By using a 2D Cartesian coordinates system time is on the X horizontal axis and the quantitative dimension is represented on the Y vertical axis.

According to the authors:

"a hierarchical structure in multiple time series can be expressed as on ordered set of time series, where individual time series are grouped hierarchically according to their proximity[150]".

In this case, proximity turns to be the progressive articulatory effort represented by a list of phonemes ordered by this criterion (see annex 5). "Multiresolution streamgraphs" has been conceived to tackle the issue of scalability in current visual representations aimed at modelling time series:

"multiple time series can be aggregated into a hierarchical structure to depict the information at different levels of abstraction[151]"

Despite CoLaJE datasets are not exactly a time series, they can probably be used as if they were time series because they do share key core features:

4 They are a sequence of discrete-time data
5 They have been taken at successive and equidistant temporal points (monthly CoLaJE sampling)

[150] Cuenca A. et al. P3160

[151] Ibidem, p 3160

6 They can be used to make predictions on future performances (see the Neural network below), that is to make statistical inferences based on previously observed values)

7 Many intertwined variables play a role in the definition of the final outcome

Here, the chronological data shown correspond to the number of phonemes over time. The phonemes are organised in a two-level hierarchy. The first level corresponds to broad categories: vowels, consonants, semi-consonants, consonant groups and the second level corresponds to the single phoneme ordered by a supposed articulatory effort (see previous figure).

As a result, we had to change the format of CoLaJE data to match the structure needed for this visualization: data were transformed by using the Python libraries *pandas*, *numpy*, *gmtime* and *strftime*.

During this formatting process a 'key step' was to add a common date to our phonetic data which was only connected to the specific child's age: by doing so we pretended that all the children were born on the 1/1/2000 allowing in this way an easier comparisons.

We made this task through a script able to convert the age into a number of days and add this number to the date of 01/01/2000. Another constraint was that the number of data (phonemes) per date (successives monthly records) was too large and this made the code inoperable. Once the reason for this bug was found, to rectify this we inserted a time that differed by one second between each phoneme of a recording: this solution has been made to make the use of the tool possible, although we think that the differences in visualisation quality between the “evolution of music genres graph” (an original and good quality one provided by the creators to show the functionalities of the tool) and our version would suggest us that the “one-second-trick” used to avoid the problem finally resulted in a quality loss in visualisation.

The original graph http://advanse.lirmm.fr/multistream/visualize.php

Our version https://marine27.github.io/TER/site_aquisition_du_langage/multistream.html

Let's have a look at the results produced by Adrien :

Appearance of consonant groups:

- We can see the emergence of the phoneme 'dK' (called 'kr') 8 times at 23 months, then 96 times at 24 months. For the following recordings, the 'dK' appears more than 100 times. It is the most present consonant phoneme of the group.

- In contrast, the phoneme 'kK' has been recorded 5 times at 24 months. For the following recordings it appears no more than 17 times.

- The phoneme 'tK' occurs 33 times at 24 months. For all recordings, he does not occur more than 56 times.

- The phoneme 'gK' is never pronounced.

“Þ” This phoneme group is particularly interesting for language experts. These are phonemes which combine two consonants and are more difficult to pronounce correctly.

Appearance of semi-consonants: By choosing the 'semi-consonants' - The phoneme 'j' is present from the first recording, 15 times at 15 months.

- The phoneme 'w' is also present from the first recording, 13 times at 15 months.

- We notice that the phoneme '6' is never pronounced.

This tool thus allows an in-depth analysis of the quantity of phonemes over time.

### 11.3.1. Confirming Clements’s “markedness avoidance principle” through Multistream graph

If we shift the mouse cursor on the graphs https://marine27.github.io/TER/site_aquisition_du_langage/multistream.html and we focus on the green part (consonants), we can test whether the non marked value is higher than the marked one in couples of phonemes that are distinguishable only by the + or – voicing contrast. According to Clements’s Theory, non marked traits are acquired before marked traits and by consequence consonants with non marked traits will be more common than their counterparts.

That is “p-b; t-d; k-g; f-v; s-z”: so the first consonant should display an higher value, and this is true for the majority of cases.

Again, it is possible to verify the same data for Adrien and Madeleine in Yamaguchi's thesis (Yamaguchi, 2012, pp309-312).

A discourse a part should be made for the /f/ - /v/ opposition, because in this case the voiced consonant /v/ is more frequent than its non-voiced counterpart, especially at later ages. If we look at the table provided by prof. Adda- Decker (Adda-Decker, 2006, p883, figure 4) /v/ is the only voiced consonants that stands above its non-voiced counterpart, meaning that in French adult language is more frequent than /f/.

## 11.4 Levenshtein Distance

We have chosen a set of words we suppose to be representative to children language in terms of frequency and morphological diversity (simple and complex). These words are "maman", "papa", "merci", "voiture", "prendre", "chercher", "derrière", "peut-être", We have calculated the difference between the subsequent variated forms and the expected pronunciation by adapting the Damerau-Levenshtein distance (this distance has already been used in Chapter 5) .

This distance is defined as follows: let two sequences of characters (in our case, sequences made up of phonemes) A and B. and a set of *n* actions to transform a given sequence of characters A into another sequence of characters B, which it is expressed with the following formula:

$$E_n(A, B) = \{e_i, i = 1...n\} \tag{9}$$

where each action can represent :

- an insertion of a character

- a deletion of a character

- a substitution of one character by another character

- transpositions of two successive characters.

A distance from Damerau-Levenshtein is defined as :

$$DL(A,B) = Card\Big(\underset{n>=0}{argmin}(E_n(A,B))\Big)$$

(10)

To put the formula in words: it gives the minimum number of operations necessary to transform one sequence of characters into another. For our study we needed a normalised score: so we set a range of possible values between 0 and 1, where 1 means perfect pronunciation (and therefore a Damerau-Levenshtein distance of 0) and 0 means totally incorrect pronunciation. Here below the formula for a normalised distance:

$$NDL(A,B) = 1 - \frac{DL(A,B)}{max(Card(A), Card(B))}$$

(11)

Where Card(A) and Card(B) represent the number of characters in word A and word B respectively. A represents the word that the child should have said and B the word that he actually pronounced. In order to retrieve this information, we referred to the CHI (words in the standard orthographic) and PHO (word said by the child in IPA) tiers. In order to retrieve the CHI tiers in IPA characters, we used the Python library “Wiktionary parser” because it provides IPA phonetic translation of a given word. We retrieved 227 words to compare for all the children (considering that each word was said at least 30 times throughout the recordings).

The normalized Damerau-Levenshtein distance allows us to evaluate the accuracy of a word pronounced by a child along his/her development.

In the following link it is possible to see all the results obtained by using the Levenshtein distance https://drive.google.com/drive/folders/1R532TzQhq-DqdSimYIHfrHNYE5TB-IGd

A summary is provided here https://marine27.github.io/TER/site_aquisition_du_langage/distance_dl.html

Here below an example

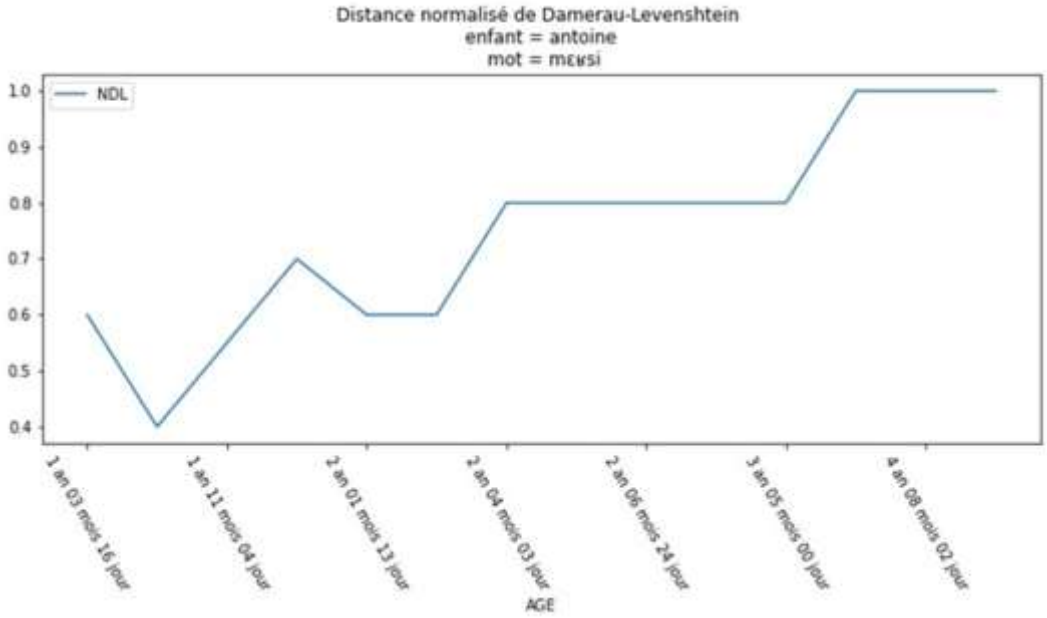



Figure 44 : Normalized Damerau-Levenshtein distance

And a summary of the evolution of mean and standard deviation related to these elaborations:

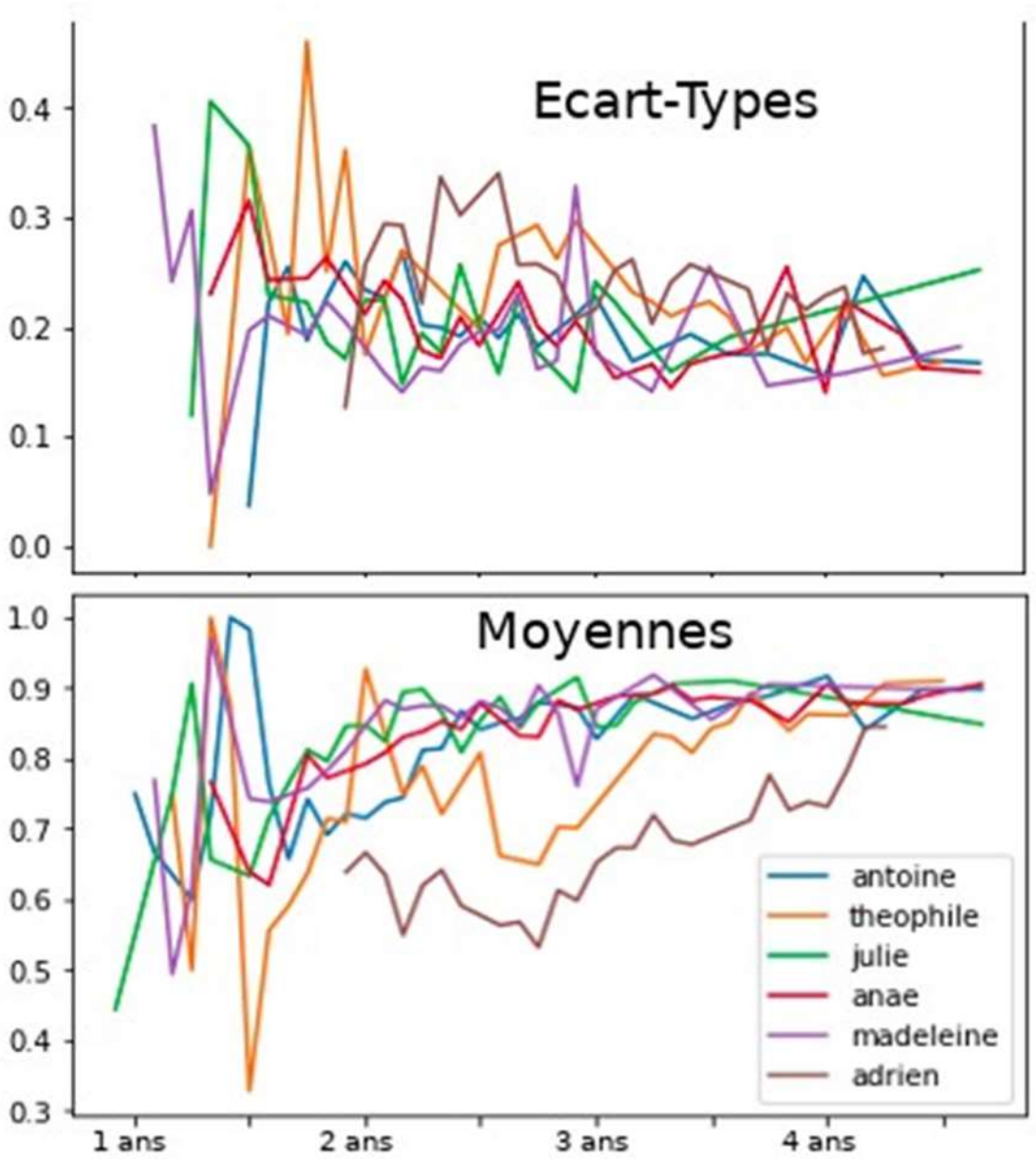



Figure 45 : Evolution of mean by time

Here again we can confirm similar results obtained by Morgenstern & Parisse in their 2012 article summarizing CoLaJE datasets into graphs: in the “moyennes” part Madeleine is always at the top and the boys begin by showing higher variation rate but around 4-5 years old they show a similar mastery of their native language.

In the graph displaying Standard Deviation values over time we can observe how the range of possible variations narrows quite proportionally over time in a quite similar way for all the children.

All the results obtained by using Damerau-Levenshtein Distance on single words can be quite easy verified by using the query provided by CoLaJE http://ct3xq.ortolang.fr/ct3xq/interro

## 11.5 Pattern Mining

Pattern mining is a technique for extracting patterns, by using this term we refer to any potentially informative and significant recurrence of sequences from a given databases aimed at improving the understanding of a data structure. In other words, to find a pattern from a data structure means to point out something that was not previously visible, in the same sense in which φ= 1,618 is the underlying pattern of the Fibonacci's sequence "1,1,2,3,5,8,13,21,34.."

The following studies were carried out with the help of M. Alatrista-Salas, professor at the Universidad de Lima, by using the Python open access library "pymining".

### 11.5.1 Sequential patterns

For this first pattern research study, we wanted to extract the most common phonemes by age and by child. We thought that knowing the frequencies would allow us to get an useful developmental information. As a result, we extracted a particular type of sub-equivalence called sequential patterns as it has been defined by Srikant and Agrawal [152].

Each phoneme was indexed by a letter symbolising the position of the phoneme in the word within which it was pronounced:

[152] R. Agrawal; T. Imielinski; A. Swami . " Mining Association Rules Between Sets of Items in Large Databases". SIGMOD Conference 1993 : 207-216

d : the phoneme appears at the beginning of the word;

m : the phoneme appears in the middle of the word;

f : the phoneme appears at the end of the word.

We have considered the children's recordings as equations. This was done to extract sub-equations of the phonemes that are common in this set of equations. The items are the words containing the indexed phonemes and the itemsets are the set of words spoken during a recording. To put a filter on this huge amount of data, the patterns that have been extracted need to have a frequency higher than 170 occurrences (this threshold has been chosen in relation to the total amount of phonemes found)

This study is also proposed by providing an application of a filter on the results, allowing us to focus the study on words containing phonemes from results known for their pronunciation difficulties in children

Thus, this study allowed us to observe a child's phoneme sequence throughout his or her learning process. We went further in the search for frequent patterns by considering the order of appearance of the phonemes.

Here is the link. Mind the “avec filtre/sans filtre” button https://marine27.github.io/TER/site_aquisition_du_langage/pattern_mining1.html

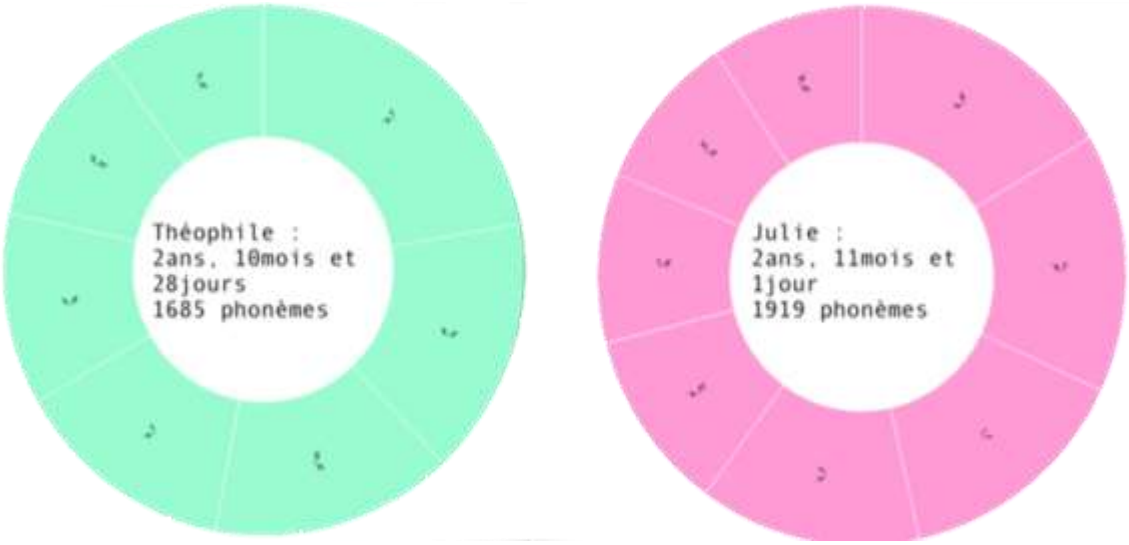



Figure 46 : Graphs of the sequential patterns

### 11.5.2 Sequential patterns two

This second study is based on sequential patterns and analyses a different type of equation. It does not consider the child or the child's age, but simply the most frequent order in which the phonemes appear and their positions in words over the years.

We have considered the sequence of recordings per child to be equivalent (as we did for aligning the Multistreamgraph). The items are the indexed phonemes and the itemsets are the set of phonemes pronounced during a recording. To make the computation easier, we made the following assumption: in an itemset, the phonemes are considered to be non-ordered.

Goal is to extract all the sub-equivalences whose support (number of occurrences) is greater than a threshold set by the user (170).

Currently, the base contains 6 sequences for 6 children. In order to have a more consistent support, each sequence has been arbitrarily divided into 160, which has allowed us to work with 960 sequences. We were thus able to extract subsequences of frequent phonemes according to their frequency of occurrence which we set at 890 occurrences, this means that the subsequence of phonemes must be present in 93% of the sequences. This threshold was arbitrarily chosen by the team after a discussion.

These patterns can help in analysing phonemes emergence in children. We wanted to deepen the discovery of the relationships between phonemes with the “association rules” method to better understand their association within words.

Important remark: unfortunately this method did not worked at all: we underestimated the number of possible combinations between phonemes and, conversely, we overestimated the accuracy of the results. We tried to interpret it in many ways but – as far as we know – it is impossible to extract any useful information from that.

Here the link
https://marine27.github.io/TER/site_aquisition_du_langage/pattern_mining2.html

Drawing any conclusion from these kind of graph seem impossible: the available zooming option for each child would seem to help but, in the end, the amount of data remains too high to deduce something even at a smaller time span.

## 11.6 Association rules

Here we try to explore how a given phoneme occurs to be with another given phoneme and the relative frequency of this association by using a set of “association rules” defined by Agrawal (Agrawal, 1993):

Let $I = \{i_1,i_2,...,i_n\}$ a set of items, and $T = \{t_1,t_2,...,t_n\}$ a set of transactions, such that “t” is a subset of $I$ ($I\ et_i \subseteq I$). An association rule is expressed as follows in the form:

$$X \rightarrow Y, \text{ or } X` \in T, Y \in T, \text{ et } X \cap Y\ 6= Ø \tag{12}$$

We call itemset a subset X of item ($X \subseteq I$).

In this specific case association rules would allow us to identify correlations of phonetic subsequences within the words pronounced by children.

Items are the words containing the indexed phonemes as explained above and itemsets are all the words pronounced during a recording. The association rules are extracted according to two criteria: their frequency of occurrence (in our case greater than 100 occurrences) and the confidence in the event, *i.e.* the proportion of words containing the first phoneme that also contain the second phoneme (greater than 0.8).

This method makes it possible to highlight the links between phonemes during child's language learning.

## 11.7 Deep Learning (Neural Network based on CoLaJE data)

https://colab.research.google.com/drive/1fIa0ak1k-yWFmsCx1FZpl6VdYEY_PpwS

or alternatively this one https://colab.research.google.com/drive/1C9_eu-kUgOAziAYvBsYzlNL-seEZt6cg

articulatory: study : https://colab.research.google.com/drive/1C9_eu-kUgOAziAYvBsYzlNL-seEZt6cg?usp=sharing

prediction : https://colab.research.google.com/drive/1fIa0ak1k-yWFmsCx1FZpl6VdYEY_PpwS?usp=sharing

*As data available for each child was insufficient to adequately train the model, we chose to train the neural network with all the data from 5 children (Antoine has been excluded because for him transcriptions were different) by mixing it. By doing so, the prediction of this net is not child-specific because it is based on all the occurrences pronounced by the children part of the CoLaJE project. We can interpret the result as an average of the expected accuracy of pronunciation for any given word that CoLaJE's children should have at 2,3 or 4 years old.

Tips for using the Colab: on the main bar, click on the button "tout exécuter" to start, choose an age, type a word by using the menu showing a list of French phonemes in IPA characters. See the result. Try to look whether the same word increases its accuracy value over time or

not. Check if there are significant differences between words containing easy and difficult syllables (e.g “tasse”, “tracteur”

(*please don’t mind the “authoring” warning, once logged in with your Google account, goes on by clicking on “yes, continue”) *

Neural networks models represent a growing domain in first language acquisition since several years: the attempt to simulate a complex process with many interacting variables that change over time and place is computationally difficult.

In this paragraph we will just introduce a simple neural network that could be considered as a sketch, a way of train ourselves on a already available dataset on which – as far as we know – nobody to date has implemented on it a neural network. Being aware of the huge quantity of articles published on this topic and find some new machine learning techniques to improve current methods could represent a brilliant thesis itself.

Here the modest goal is to write some lines on one of the latest development of language modelling regarding phonology and to create a simple model by using TensorFlow[153], a free and open-source software library for machine learning.

To say that we model language acquisition through networks is a slippery description: essentially, what we have done is to use CoLaJE data to train a model and to test the validity of the prediction based on this model on the same data on which we have built up the model. The way the network is conceived in terms of structure and processes is not related to the way in which a child’s cognition is made of.

“The design and implementation details of any computational model will of course differ dramatically from the mental architecture and processes of a child. Yet, the success of a model in learning from the same input as a child provides evidence that the child may employ similar learning strategies[154].

This means that if the model works, we may have found an underlying structure in CoLaJE longitudinal data that allow us to predict the accuracy of a given word (defined by its

[153] https://www.tensorflow.org/learn?hl=fr URL consulted on22/10/2020

[154] Roy D. et al. (2006). “The Human Speechome Project”. Proceedings of the 28th Cognitive Science Society Conference

constitutive phonemes) at any given age based on the accuracy of the same sequence of phonemes during the same age period. This does not imply that we have discovered how language learning has developed in Adrien or Madeleine, but simply that the way by which data are subsequently structured along the recordings combined by the way by which data are computationally transformed by the neural network could be analogous in some aspects in virtue of this similarity.

A huge limit of the conceived neural network is that it does not take into account the input a child receives from their parents and environment: to simplify our task, we choose not to consider it despite we are obviously aware that is of primarily importance.

An up-to-date and complete example of a neural network capable of modelling language acquisition focusing on the phonetic/phonological interface is proposed by Boersma et al. (2020) in a recent paper:

"We provide a first proposal of a neural network model that can handle two important aspects of the transmission of a sound system from one generation to the next, namely category creation and auditory dispersion, and we simulate the model on a range of synthetic data[155]"

This model takes into account what we left out in our simplified model, input from adults and the consequent perception related issues:

"The model therefore addresses the hitherto unsolved problem of how symbolic-looking discrete language behaviour can emerge in the child from gradient input data from her language environment"

Here below a schematic diagram representative of the what the authors thought to be the two levels of representation and stored knowledge in a hierarchically structured model of phonology and phonetics

[155] Boersma P.; Benders T.; Seinohorst K. (2020). "Neural networks models for phonology and phonetics". Journal of Language Modelling Vol 8, No 1 pp. 103–177. P104

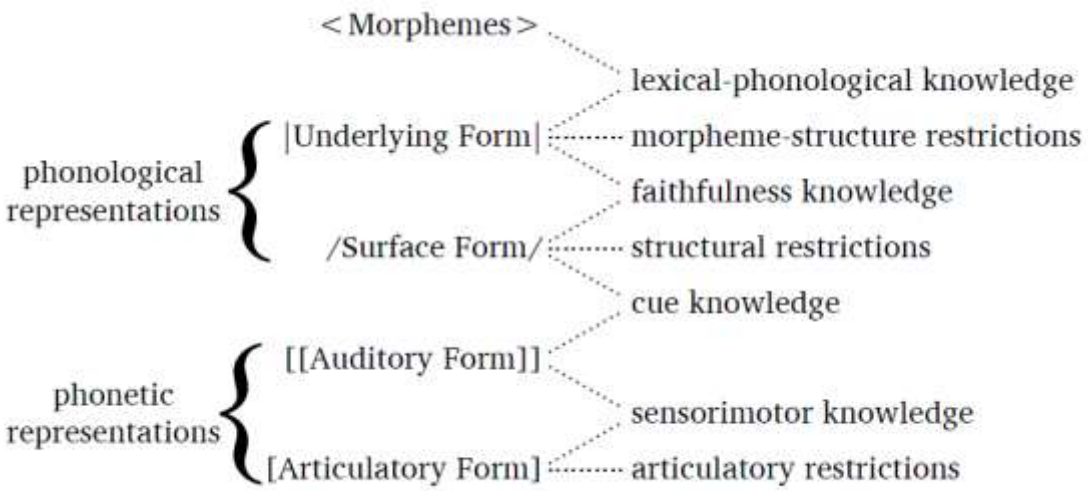



Figure 47[156] : Schematic diagram

While the core algorithm to simulate gradual learning is different:

"[..] weights can learn from experience: they change only slowly over the months and years as the child is acquiring her language. In this section we identify a learning rule for our stochastic bidirectional artificial networks: we show that out of a family of Hebbian-like learning rules the only rule that meets the requirements of stochasticity and symmetric bidirectionality is what we call *inoutstar* [..][157]"

Neural networks are used to model and simulate first language acquisition even to attempt to contribute to the long-lasting debate between Nativism and Constructivism: the pioneer work by Deb Roy cited in Chapter 5 helped to adress this issue by providing a near complete empirical role that nurture plays in learning:

"[..] what are the set of ontological constraints that must be built into a model for it to successfully learn aspects of language? If a machine can be shown to acquire some capability

[156] Ibidem, p 105

[157] Boersma P.; Benders T.; Seinohorst K. (2020). "Neural networks models for phonology and phonetics". Journal of Language Modelling Vol 8, No 1 pp. 103–177. P124

or structure *X* without corresponding innate preconditions, this provides evidence that the child's environment provides *X* – and thus need not be innate" (Roy D., 2006, p2)

Our inexperience on this field and the already scheduled timing of the master project lead us to use other algorithms that were already in use in Montpellier's data science lab without taking into consideration state of the art paper such as the on of Broesma et al.

### 11.7.2 Articulation study

We created a code based on articulatory features for each possible phoneme (see Appendix: Articulatory Features.) Each phoneme is coded according to its articulatory features components:

Anterior Oral

Open-,

Closed-,

Rounded-,

Occlusive,

Liquid

Fricative,

Bilabial,

Labiodental,

Dental,

Palatal,

Lateral,

Voiced

Each word is therefore represented by a matrix whose columns are the articulatory elements of a phoneme and whose rows are the phonemes constituting the word.

In order to better assess a child's ability to pronounce a word correctly, we have decided to set up several classification models: a model by age based on a convolution neural network (CNN).

A two-dimensional convolution is organized as follows: a kernel (an n × k weighted matrix) performs a convolution product with an m × p matrix by projecting itself onto it

A padding can be added to the input matrix, to make sure that the same dimensions are kept after the convolution product. Usually this layer is filled with 0 values. The stride in a convolution is the 'step' of the core displacement during the convolution product with the input matrix. Usually a value of 1 is chosen, so the nucleus moves by one cell at each step.

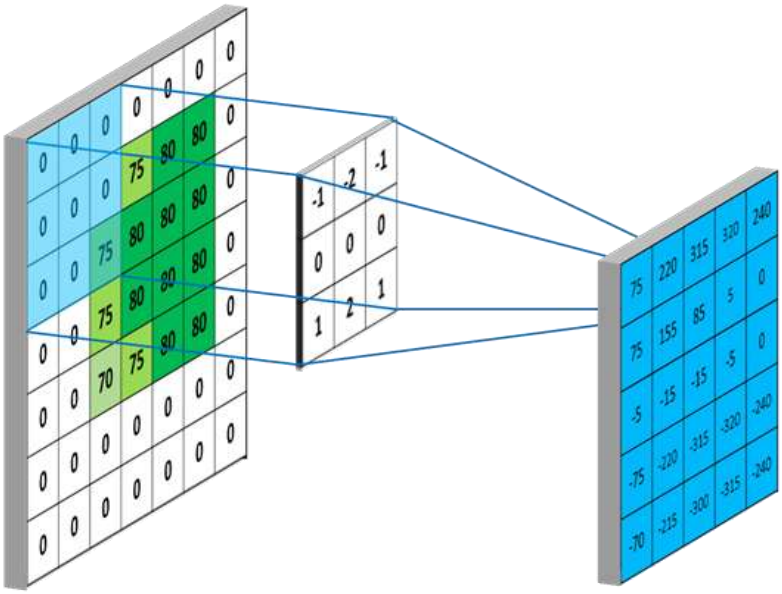


Figure 48 : Convolution with padding[158].

[158] https ://mlnotebook.github.io/post/CNN1 URL consulted on 5 may 2020

It is assumed that there is no padding and that the stripe is 1. The result is a matrix of size: (n - m + 1) × (k - p + 1).

A set of filters with (a priori different) cores will therefore perform this operation on an input data (usually an image) and return after activation, a feature map. In a neural network, a set of filters (or neurons) represents a layer.

In order to limit the number of parameters and to avoid overlearning, a pooling layer is usually applied which will reduce the size of the filters of a layer by aggregation. In our case, we will use max pooling layers that will aggregate the spatial information by maximum value.

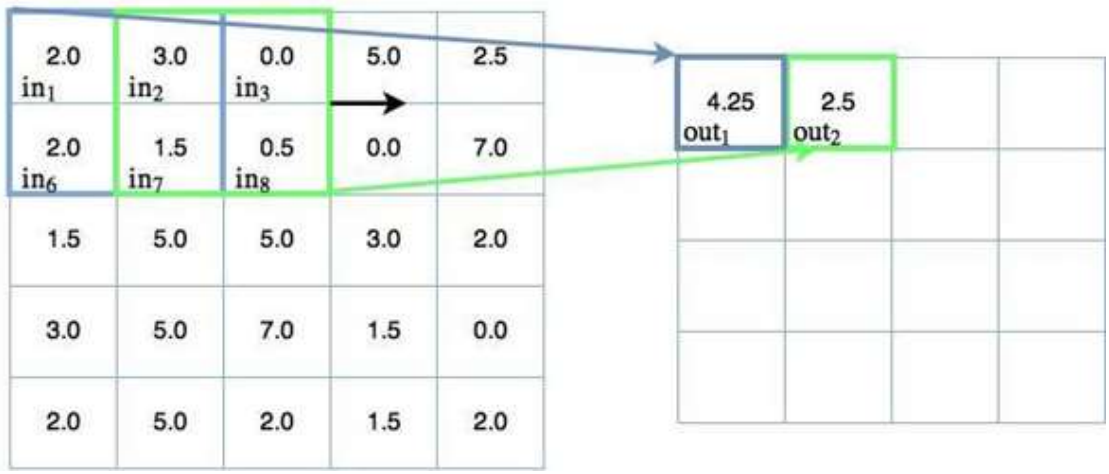



Figure 49 - Schematisation of a convolution operation[159]

[159] https ://adventuresinmachinelearning.com/ URL consulted the 5 may 2020

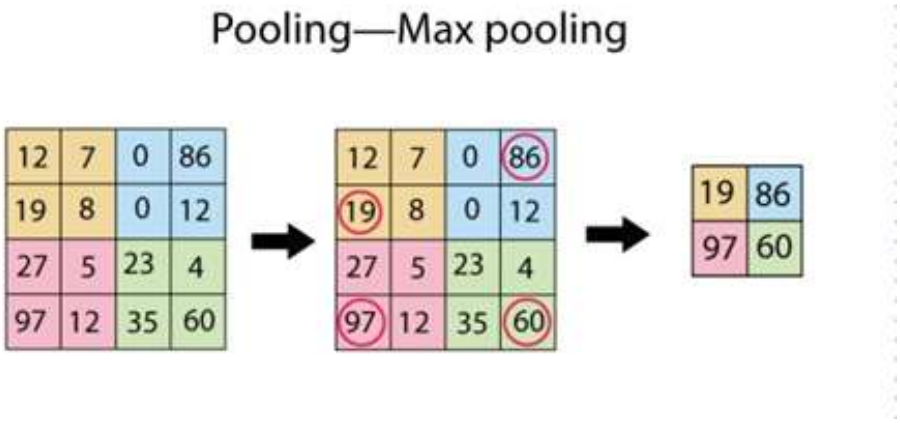



Figure 50 : Schematic diagram of a 2 × 2 max pooling system[160]

The model will perform convolutions and poolings on two different axes in parallel:

- On the phonemes axis, in order to retranscribe the articulation of the phonemes between them Convolution 2 × 1

- On the axis of the articulatory components, in order to retranscribe the importance of articulatory places.

Convolutional × 2

[160] principlesofdeeplearning.com/

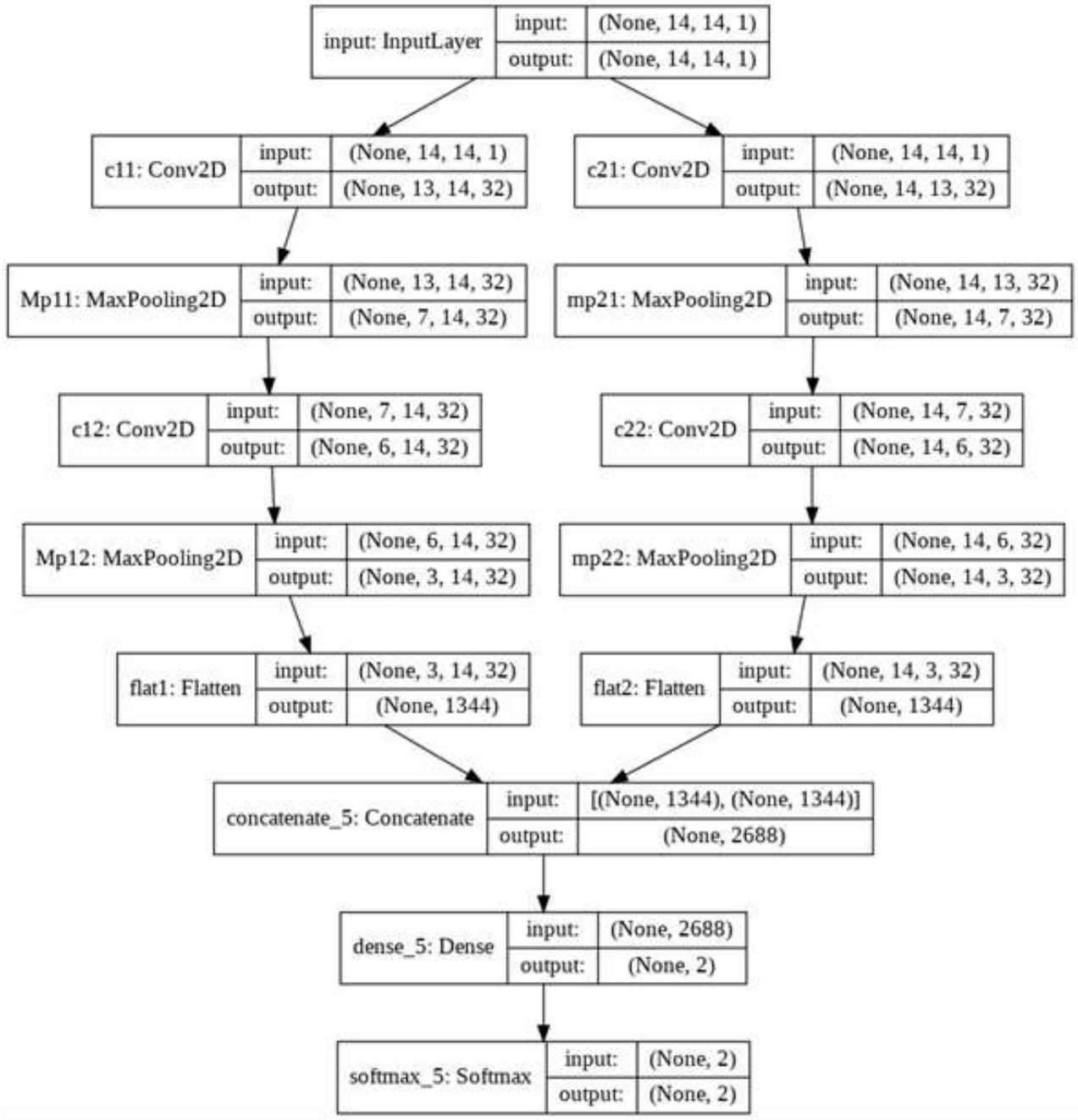



Figure 51 : Architecture of the model via Keras

The layers of these parallel networks are then 'flattened' (from matrix to vector and concatenated together). Finally, an output layer containing 2 neurons represents the two possible states. The activation function of the last layer is a *Softmax* in order to evaluate the probability that a child can express a word according to his/her age. If we put x0 and x1 as the output values of the last layer, the activation function will be :

$$\underset{i=0,1}{Softmax}(x_i) = \frac{e^{x_i}}{e^{x_1} + e^{x_2}}$$

(13)

The activation function used for each layer (except the last one) is the Read function (simply defined by : Relu(x) = max(x,0)). The idea behind this structure is to set up a model whose feature maps are able to represent the difficulties of coarticulation of the articulatory phonemes, in terms of manner of articulation and place of articulation.

In order to bring new solutions and perspectives to the corpus we have decided to apply two deep learning models, each with a different objective and structure. These models will allow us to have at our disposal a tool that will try to represent as accurately as possible a child's pronunciation difficulties at a given age according to the different phonetic groups

The model has been trained on

- 13171 words for children aged 2;

- 9828 for children aged 3;

- 6610 for children aged 4.

Of all the data, 90% have been saved for training and 10% for validation.

On the validation set, the following results were obtained:

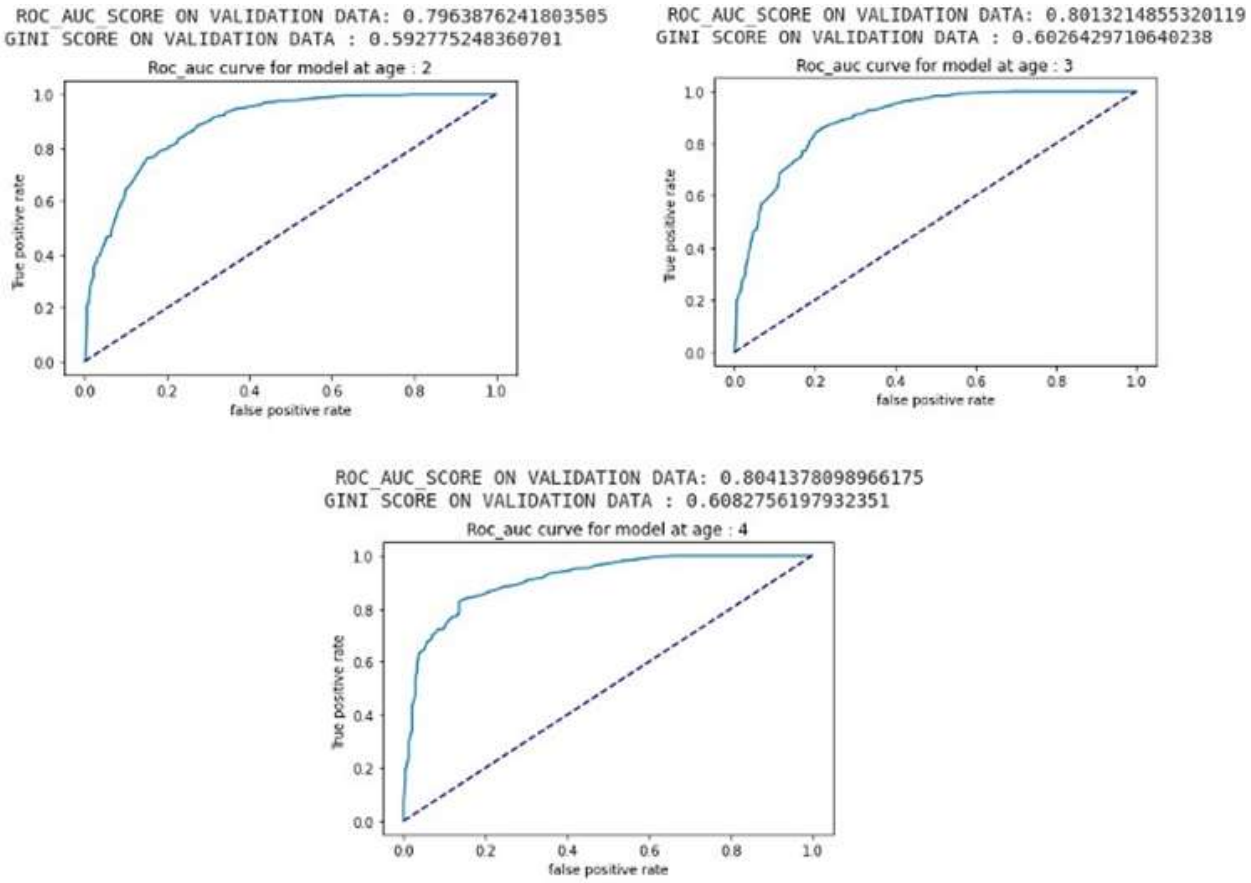



Figure 52 : Correctness's measure of the model at different ages

We can appreciate that the model predicts how the child will properly pronounce a given word in a quite fair way. Generalizations that have been made on the training set would indicate that a greater sample will further improve predictions' accuracy.

| age | precision | recall | f1-score | Support 0 | Support 1 |
|---|---|---|---|---|---|
| 2 | 0.83 | 0.83 | 0.82 | 562 | 902 |
| 3 | 0.83 | 0.83 | 0.83 | 399 | 693 |
| 4 | 0.83 | 0.84 | 0.83 | 243 | 492 |

Table 18 : Prediction's accuracy

This model is available through a Google Colab[161]. The only parameter to set in order to have a valid result is that that word should contain at least two phonemes

This model is therefore a tool for the articulatory analysis of phonemes in children (at 2, 3 and 4 years old). For example, we expect that the word "trak" will not be easily expressed by a 2 year old child because of the presence of an occlusive-liquid at the onset of the syllable: the model has in fact a very low confidence level and we can therefore conclude that the word will not likely to be said correctly at 2 years old. While at 4 years-old the model is almost sure that this particular articulation will be properly pronounced.

This model is a prototype and it is based on a particular corpus, it would be interesting to add different *corpora* to reinforce its robustness. A consistent amount of data could allow an increased capacity to generalize the model's predictions.

### 11.7.3 Phonetic embedding

Word embedding is a method of automatic language processing that aims to represent words in a vectorized form. Each vector is expressed in *Rn* where *n* is the dimension of the embedding. The goal of such a method is not to attribute random values to each word, but to define a representation space in which the words with the same "context" are close. Inspired by this approach, we decided to apply the algorithm to the phonemes and to represent each phoneme according to its context. One of the most optimal methods to achieve this result is to use a neural network. We used a Word2Vec Skip Gram[162] type structure.

[161] https ://colab.research.google.com/drive/1fIa0ak1k-yWFmsCx1FZpl6VdYEYPpwS

[162] Hu, Jie Li, Shaobo Yao, Yong Yu, Liya Guanci, Yang Hu, Jianjun. (2018). Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. Entropy. 20. 104.

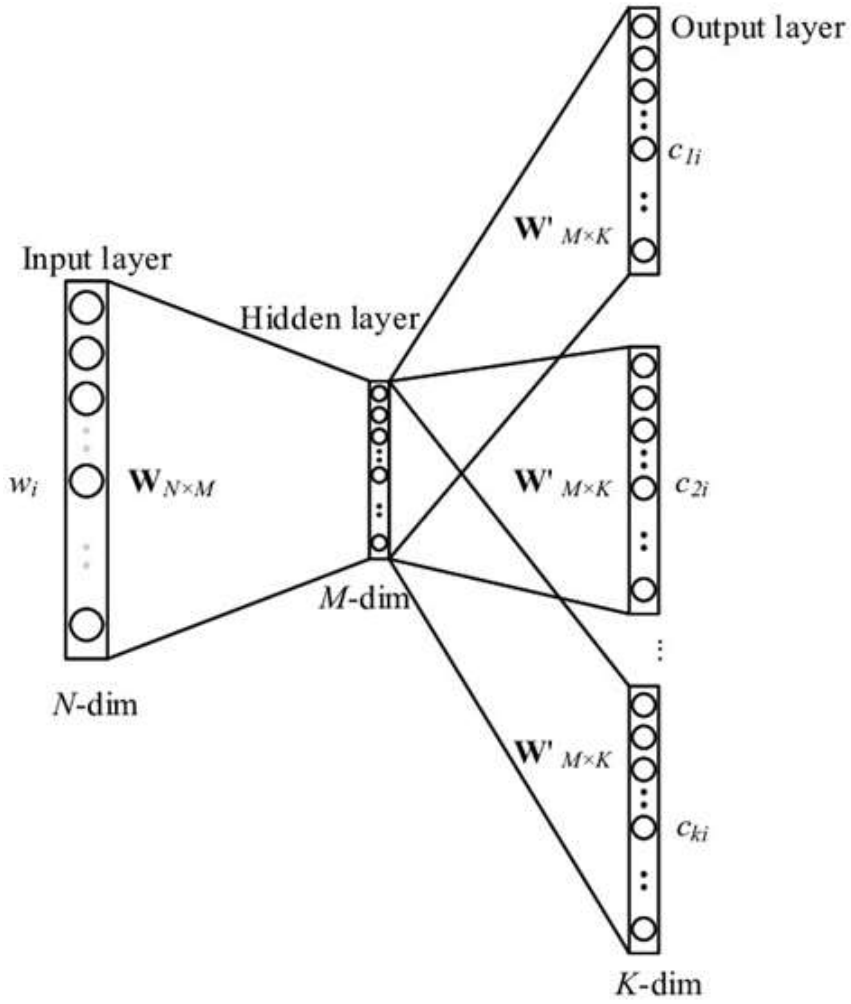



Figure 53 Skip Gram architecture

The purpose of such a structure is to predict the context of an element. The entry here will be a phoneme and the trained network will have to be able to 'predict' this specific phoneme's context. To give an example: the word 'dad' if we take a 'direct' context (a single character sliding window) we will have the following relationships:

| target_phoneme | context |
|---|---|
| 'p' | ['blank', 'a'] |
| a' | ['p', 'p'] |

| | |
|---|---|
| 'p' | ['a', 'a'] |
| a' | ['p', 'empty'] |

Here each phoneme has a particular context and the aim of our network is to predict the most likely one. We can see that for the phoneme 'p', the most probable context is: 'a' after the 'p', and before the 'p' all the other phonemes present in our context have the same frequency.

Our data will be “hot-encoded”: each word will be represented by a vector of the size of our vocabulary, in this case French phonemes , and this vector will be filled with 0, except at a position where it will be 1. The position corresponds to the label of the phoneme. For example, if we code our phonemes in integer from 0 to 40, the phoneme n◦20 will be represented by a vector of size 41 filled with 0, except at the 20th position where there is a 1.

Our neural network will therefore try to predict the context of our target value which will be the input. The output for each 'context element' will be a *Softmax* that will give us the probability that an element will be part of the context. The elements having the same context, will thus have a very "close" projection on the hidden layer during the re-propagation (Hidden Layer in the figure above)

By retrieving the hidden layer, we obtain a projection of our “vocabulary” according to its context. The number of neurons in the hidden layer represents the dimension in which we will make our projection. Usually we can consider that taking a dimension equal to the fourth root of the vocabulary size is an acceptable approach. In our case, with a vocabulary ranging from 38 to 44 distinct phonemes (size depends on the child and the related age), we have chosen a dimension equal to 3. The coding of all this procedure was done with “Tensorflow”. to highlight the network structure and to have more information about the architecture.

We will use this vector representation to bring phonemes together in their new layout placed in the new representation space. As it is interesting to know the possible connections between phonetic groups, we decided to use a method of hierarchical ascending classification to have a visualization of the possible groupings. We will use the "Farthest-first traversal[163]" method

[163] stackoverflow.com/questions/48479915/what-is-the-preferred-ratio-between-the-vocabulary-sizeand-embedding-dimension

and we will use the Euclidean distance to define the aggregations between groups of phonemes.

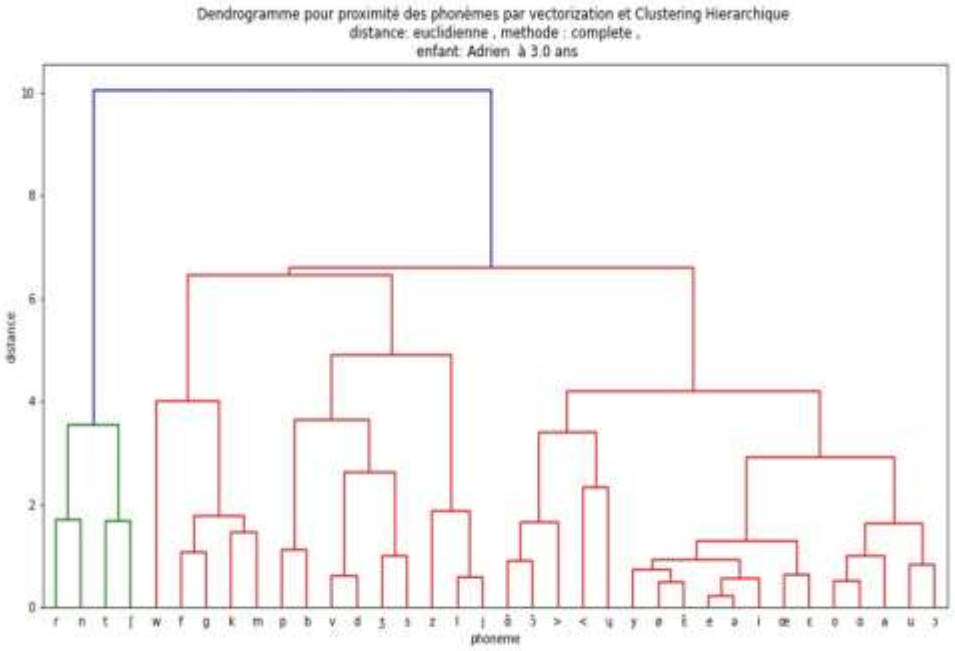



Figure 54 Dendrogram

After having trained our model and applied a method of hierarchic agglomeration on it, we were able to make an interesting observation: despite the obvious limits of a first attempt, we can see how vowels are grouped in the same area and consonants in another specific area. Inside the consonant area, only “p” and “b” seem to be grouped according to a certain criterion, while all the other groups seem random.

---

[163] . colab.research.google.com/drive/1dOkD50-mnfAYjFppTIk4ezbphBDrFtuk

[163] Mukesh K. (2013). An optimized farthest first clustering algorithm. 1-5. Proceedings of the Nirma University International Conference on Engineering (NUiCONE)

## Conclusions and future directions

This thesis mainly consists in an attempt to study first language acquisition with a variety of quantitative methods and to provide new ways of visually representing the evolution of language acquisition over time.

I hope what I wrote would be enough originally and interesting to be used by other researchers in the domain of first language acquisition: CHAID, EM clustering, Multistreamgraphs and all the different statistics and graphs provided would serve as complimentary tools to improve current studies on child language as well as integrate already achieved study on child language, both in French and other languages.

For example, a visual tool such as Multistreamgraph could be improved in some technical details as well as in the final rendering: it would then be ready to be used as an additional information to every longitudinal corpus in CoLaJE. Because, as data formats are the same (or almost the same), it would be possible to apply the same procedure consisting in transforming raw data contained in XML files into an interactive interface in which every researcher would be able to have an idea of the path of every consonant and vowel over all the years during which the children have been recorded. This graph can be modified in many different ways: for example it is possible to focus only on one phoneme by simply leave aside by a drag-and-drop mouse move all the other phonemes (this is done in the left side part of the page).

A final version of Multistreamgraph has this degree of accuracy: http://advanse.lirmm.fr/multistream/visualize.php

While results presented for the six CoLaJE children still present some problems in visualisation that need to be fixed: we have not been able to solve the problem of shrinking streams from one session to the next, so the flow seem to reduce in phonemes quantity in the passage between successive records while it is the opposite phenomenon that really occurs: almost every phoneme increases its absolute value from one month to the next (although the relative proportion between them can vary a lot, as is the case for /m/: its importance is great between 1 year and two-year old and then decrease in terms of relative value, but this detail is clearly well represented in the Multistream regarding music genres evolution).

A very smart suggestion has been given to me from professor Yamaguchi during a meeting in Sorbonne University: the list of French phonemes on which CoLaJE *corpora* are currently analysed could be replaced by an adapted list of phonological traits (according to Clements's theory. Clements & Hume, 1995 ) as the one showed in her thesis (see annex 8).

By doing so, it would be possible to observe what now is possible to infer in an approximate manner only through indirect paths: the learning of a trait over time.

If we would analyse the *corpora* according to the nine traits listed here

| **Traits** | p | b | t | d | k | g | f | v | s | z | ʃ | ʒ | m | n | ɲ | l | ʁ | j |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [±sonant] | − | − | − | − | − | − | − | − | − | − | − | − | + | + | + | + | + | + |
| [±approx.] | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | + | + | + |
| [±continu] | − | − | − | − | − | − | + | + | + | + | + | + | | | | | | |
| [±voisé] | − | + | − | + | − | + | − | + | − | + | − | + | | | | | | |
| [LABIAL] | ✓ | ✓ | | | | | ✓ | ✓ | | | | | ✓ | | | | | |
| [CORONAL] | | | ✓ | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ |
| [±post.] | | | | | | | | | − | − | + | + | | − | + | | | |
| [DORSAL] | | | | | ✓ | ✓ | | | | | | | | | | | ✓ | |
| [±latéral] | | | | | | | | | | | | | | | | + | | − |

Table 20 : Corpora according to the nine traits

it would then be possible to provide a more straightforward way to confirm tenets of Clements' theory such as feature hierarchy and economy.

This because if we were able to put the nine traits that specify French phonemes on the left part of the Multistreamgraph, we would then put every single phoneme on the stream, this would allow us to see its increase over time. Two additional information that are missing in the current graph would be the variation rate (whether the phoneme has been correctly said or not) and the position it occupies at the syllabic level, to see how this influences its pronunciation. But an important information would still miss: which phoneme influences the pronunciation of another, in which position, and how?

What would become times more difficult is the programming of all this analyses. As explained before, we have focused on a sentence level (see CHAID and Expectation-Maximization analysis) because we soon realised that dealing with nearly ten thousands sentences per child was already a challenge.

I would need to master Python programming language as an advanced user to hope to do so in relatively short time. The attempts we made to model the phonemes acquisition below the word level such as in "sequençage phonétique", "lien entre phonèmes" and émergence phonétique" seem to have been failed to reach the expected descriptive target[164].

Adding more intertwined levels of analysis (feature, phoneme, position, syllable [onset + nucleus + coda], word) would give a much more clearer picture of acquisition than the current, but it would add many complex steps in the data structure manipulation phase and in the programming phase. For this reason, as previously said, I will start to use PHON to solve all of these issues.

This thesis has been mostly written – for many reasons – in the last year of my Phd: the overall aspect is still confusing, but the work behind the results has been huge. I hope to have demonstrated how quantitative analyses can improve the understanding of language acquisition, in fact

"Frequency effects are observed across a variety of different domains, levels (e.g. lexical vs. abstract; type vs. token, absolute vs. relative), and outcome measures (e.g. age of acquisition, rates of error/correct use, types of error), and therefore constitute a phenomenon that demands explanation under any theoretical account"[165]

It is hard to predict what the ongoing data revolution will bring us

Would it be possible that the next advances in deep learning and computational power will allow linguists and computer scientists to create a model of language acquisition capable of considering all the variables at play that will – in turn – simulate in a plausible and realistic way what acquisition really is?

Will corpus linguistics still continue to play the great role it had until now in explaining aspects of first language acquisition or will new technologies reduce its importance?

I am not really able to understand the pioneering works on modelling language through neural networks, I would need an intensive math course to reach this level.

---

[164] See for example the graph here https://marine27.github.io/TER/site_aquisition_du_langage/pattern_mining3.html

[165] Amber B. et al., 2015, p 264

Next step will be to try to review current literature on this topic, but my first impression is that technological and scientific advances seem to have a tendency toward a global explanation: as in building up robots, the final aim seems to reproduce a model of a child and predict what s/he will learn on the basis of the given input and the pre-existing structure that will compute that input.

This reminds me High school classes in Philosophy and Laplace's famous definition of determinism, in his 1814 "Essai philosophique sur les probabilités":

" Nous devons donc envisager l'état présent de l'Univers comme l'effet de son état antérieur et comme la cause de celui qui va suivre. Une intelligence qui, pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'Analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle et l'avenir, comme le passé serait présent à ses yeux[166]".

[166] "We ought to consider the present state of the universe as the effect of its previous state and as the cause of that which is to follow. An intelligence that, at a given instant, could comprehend all the forces by which nature is animated and the respective situation of the beings that make it up, if moreover it were vast enough to submit these data to analysis, would encompass in the same formula the movements of the greatest bodies of the universe and those of the lightest atoms. For such an intelligence nothing would be uncertain, and the future, like the past, would be open to its eyes". Personal English translation

# Bibliography


Adda-Decker M.; (2006). “De la reconnaissance automatique de la parole à l’analyse linguistique de corpus oraux”. JEP2006 - XXVIes Journées d’Étude sur la Parole, 12-16 juin 2006, Dinard (Proceedings)

Agrawal R.; Imielinski T.; Swami A. (1993). “Mining Association Rules Between Sets of Items in Large Databases”. SIGMOD Conference 1993 : 207-216

Alwan A, Jiang J, Chen W. (2011). “Perception of place of articulation for plosives and fricatives in noise”. Speech Commun. 53:195–209

Ambridge B; Kidd E.; Rowland C. F; Theakston A. (2015). “The ubiquity od frequency effects in first language acquisition”. J. Child Lang. 42. 239-273. Cambridge University Press

Biggs, D., B. De Ville, and E. Suen (1991). “A method of choosing multiway partitions for classification and decision trees”. Journal of Applied Statistics 18 (1), 49–62.

Boersma P; Benders T.; Seinhorst K. (2020). “Neural networks models for phonology and phonetics”. *Journal of Language Modelling* Vol 8, No 1 pp. 103–177

Briglia A.; Mucciardi M.; Sauvage J. (2020). « Identifying the speech code through statistics : a data-driven approach ». Pearson Edition. Proceedings of 50th Italian Conference of the Statistical Society. Pisa.

Chevrot J-P; Fayol M. (2001). “Acquisition of French Liaison and Related Child Errors”. Research on Child Language Acquisition, vol. 2, M. Almgren, A. Barreña, M.J. Ezeizabarrena, I. Idiazabal and B. MacWhinney (eds), Cascadilla Press, pp.760-774

Colombo, M. (2018). « Bayesian cognitive science, predictive brains and the nativism debate » Synthese 195 : 4817-4838

Colombo M. ; Wright C. (2018). « First principles in the life sciences : the free-energy principle, organicism, and mechanism ». Synthese. Springer

Colombo, M., & Seriès, P. (2012). “Bayes in the brain—On Bayesian modelling in neuroscience”. *The British Journal for the Philosophy of Science, 63,* 697–723

Childers J. ; Tomasello M. (2001) « *The Role of Pronouns in Young Children's Acquisition of the English Transitive Construction* » Developmental Psychology, Vol. 37. No. 6, 739-748.

Clements, G. N.(1985).“The geometry of phonological features”. Phonology yearbook 2. 225-252

Cuenca E., Sallaberry A., Wang Y., Poncelet P. (2018). « *MultiStream : A Multiresolution Streamgraph Approach to explore Hierarchical Time Series* ». IEEE Transactions on visualization and computer graphics, vol.24, no. 12.

Damerau F. (1964). “A technique for computer detection and correction of spelling errors”. ACM Communications

Darwin C. (1877). « A biographical sketch of an infant ».Mind 2. 285-294

Dempster A.P. ; Laird N.M. ;Rubin D.B. (1977). « *Maximum likelihood from incomplete data via the EM algorithm* ». Journal of the Royal Statistical Society. Series B: Methodological 39: 1–38

Dodane C.; Martel K. (2009). “Évolution de l’inventaire de contours de Fo chez deux enfants français de 10 à 12 mois: l’importance du contexte pour décrire le stade pré-linguistique” Enfance, 305-316

Dodane C. ; Massini-Cagliari G. (2010). « La prosodie dans l’acquisition de la négation: étude de cas d’une enfant monolingue francaise ». Alfa, Revista de Linguistica.

Dos Santos C. (2007). « Développement phonologique en français langue maternelle : une étude de cas”. Phd thesis Université Lumière Lyon2

Ferrer i Cancho R.; Solé R. V. “The small world of human language”. Proceedings Royal Society of London B (2001). 2260-2265

Fikkert P. (1994).“On the acquisition of prosodic structure”. Radboud University Nijmengen Phd thesis Repository.

Fort M., Brusini P., Carbajal M., Sun Y., Peperkamp S. (2017). « A novel form of perceptual attunement: Context-dependent perception of a native contrast in 14-month-old infants ». Developmental Cognitive Neuroscience 26. 45-51

Frey J.C. (2020). “Using data mining to repurpose German language corpora. An evaluation of data-driven analysis methods for corpus linguistics”. Phd thesis. Università di Bologna repository

Friston, K. (2010). “The free-energy principle: a unified brain theory?”. *Nat Rev Neurosci* **11,** 127–138

Friston K. (2009). "The free-energy principle: a rough guide to the brain?". Trends in Cognitive Sciences, Volume 13, Issue 7, Pages 293-301

Gembillo G. (2011) “Filosofia della complessità”. Le Lettere, Firenze

Goodman, L. A. (1979). “Simple Models for the Analysis of Association in Cross Classifications Having Ordered Categories”. Journal of the American Statistical Association, 74, 537-552.

Gould S.J.; Lewontin R.C. (1979). « The Spandrels of San Marco and the Panglossian paradigm: a critique of the adaptionist programme ». Proceedings of the Royal Society of London. Series B, Biological Sciences, pp. 581-598

Hu, Jie Li, Shaobo Yao, Yong Yu, Liya Guanci, Yang Hu, Jianjun. (2018). Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. Entropy. 20. 104

Hu, Jie Li, Shaobo Yao, Yong Yu, Liya Guanci, Yang Hu, Jianjun. (2018). (Patent) “Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification”. Entropy. 20. 104

Jakobson, R. (1941/68). “Child language, aphasia, and phonological universals”. The Hague: Mouton. English translation of “Kindersprache, aphasie und allgemeine lautgesetze*”*. Uppsala.

Kass, G.V. (1980) “*An Exploratory Technique for Investigating Large Quantities of Categorical Data*”. App. Statist 29(2):119-127

Kiebel S.; Friston K. (2009). “Predictive coding under the free-energy principle” *Phil. Trans. R. Soc. B* **364** 1211–1221

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S. & Iverson, P. (2006). » Infants show a facilitation effect for native language phonetic perception between 6 and 12 months ». Developmental Science, 9, F13-F21

Laplace P.S. (1814). “Essai philosophique sur les probabilités” . English transl. by A.I. Dale, “Philosophical essay on probabilities”. Springer, 1995

Lestrade S. (2017). “Unzipping Zipf’s law”. PlosOne

Lorenz E.N. (1963) “Deterministic non periodic flow”. J. Atmosph. Sci. 20, 130–141

Lorenz E.N. (1972). “Predictability: does the flap of a butterfly’s wings in Brazil set off a tornado in Texas?” 139th Annual Meeting of the American Association for the Advancement of Science.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. 3rd edition*. Mahwah, NJ: Lawrence Erlbaum Associates

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T., Xu, F., & Clahsen, H. (1992). Overregularization in Language Acquisition. Monographs of the Society for Research in Child Development, 57(4), I-178

Martel K; Dodane C. (2012) “Le rôle de la prosodie dans les premières constructions grammaticales: étude de cas d'un enfant français monolingue”. Journal of French Language Studies Volume 22 Issue 1 Pages13-35 Cambridge University Press.

McLeod S.; Crowe K. (2018). “Children’s Consonant Acquisition in 27 Languages: A Cross-linguistic Review”. American Journal of Speech-Language Pathology. 1-26.

Monod J. (1970). “Le hasard et la nécessitté. Essai sur la philophie naturelle de la biologie moderne”. Editions du Seuil. Paris

Morgenstern A. (2006) ”Un ”Je” en construction: genèse de l’auto-désignation chez le jeune enfant”. OPHRYS Editions. Paris

Morgenstern A. ; Parisse C. (2007). « Codage et intérpretation du langage spontané d’enfants de 1 à 3 ans ». Corpus 6. Intérpretation, contextes, codages, pp 55-78

Morgenstern A. ; Parisse C. (2012), « *The Paris Corpus* ». French language studies *22*. 7-12. Cambridge University press. Special Issue

Morgenstern, A.; Parisse, C. (2012). “Constructing "basic" verbal constructions: a longitudinal study of the blossoming of constructions with six frequentverbs”. In Bouveret, M. & Legallois, D. (Eds.) Constructions in French. Benjamins.

Morgenstern A.; Sekali M. (2009) “What can child language tell us about prepositions ?” Jordan Zlatev, Marlene Johansson Falck, Carita Lundmark and Mats Andrén. *Studies in Language and Cognition*, Cambridge Scholars Publishing, pp.261-275

Morin E. (1990). “Introduction à la pensée complexe”. Paris. ESF.

Morin E. (1994). “La complexité humaine”. Textes choisis. Paris. Flammarion

Morin E. (1999).”Relier les connaissances”. Paris. Seuil

Orvig A. ”Dialogical beginnings of anaphora: the use of third person pronouns before the age of three”. Journal of Pragmatics 42 (7). 1842-1865.

Parisse C.; Le Normand M-T. (2000).“How children build their morphosyntax: The case of French. Journal of Child Language, Cambridge University Press (CUP), 27, pp.267-292.

Piantadosi S. (2014). “Zipf’s word frequency law in natural language: A critical review and future directions”. Psychon Bull Rev.; 21(5): 1112–1130.

Piattelli-Palmarini M. (1983) “Language and Learning: The Debate Between Jean Piaget and Noam Chomsky**”.** *Mind* 92 (365):138-140

Pinker S. (2003). “The blank slate. The modern denial of human nature”. Penguin Books, New York

Plebe A.; De la Cruz V. (2016). “Neurosemantics. Neural processes and the construction of linguistic meaning”. Volume 10. Studies in Brain and Mind. Springer

Qi P.; Zhang Y.; Zhang Y.; Bolton J.; Manning C. D. (2020). “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages”. Association for Computational Linguistics (ACL) System Demonstrations.

Ramus F.; Peperkamp S.; Christophe A.; Jacquemot C.; Kouider S.; & Dupoux E. (2010) “A psycholinguitic perspective on the acquisition of phonology”. In Fougeron C.; Kuhnert B.; D’Imperio M.; & Vallée N. (Eds), “Laboratory Phonology 10: Variation, Phonetic Detail and Phonological Representation”. (pp 311-340). Berlin: Mouton de Gruyter.

Ratner, B. (2017). Statistical and Machine-Learning Data Mining Techniques for Better Predictive Modeling and Analysis of Big Data, Third Edition, CRC Press

Rochat P.; Broesch T.; Jayne K. (2012) ”Social awareness and early self-recognition. Consciousness and cognition 21. Elsevier. 1491 – 1497.

Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G., Maddocks, K., O'Brien, P., & Wareham, T. (2006). Introducing *Phon*: A Software Solution for the Study of Phonological Acquisition. *Proceedings of the Annual Boston University Conference on Language Development. Boston University Conference on Language Development*, *2006*, 489–500.

Roy D. et al. (2006). « *The Human Speech Home project* ». Proceedings of the Annual Conference Cognitive Science Society.

Saffran J. (2003). “Statistical language learning: mechanisms and constraints”. Current directions in psychological science. Vol 12. No 4. P. 110-114

Saffran J. R ; Aslin R. N ; Newport E. L (1996). « Statistical learning by 8-Month-Old infants », Science, vol. 274. 1926-1928

Sander E. K. (1972). *“When are speech sounds learned?”.* Journal of Speech and Hearing Disorders, 37(1), 55-63.

Sauvage J. (2015).« L’acquisition du langage : un système complexe », L’Harmattan, Louvain la neuve

Sekali. M.(2012). “First language acquisition of french grammar (from 10 months to four years old). Introduction ». French Language Studies 22 (2012), 1–6, Cambridge University Press

Sievers C.; Wild M.; Gruber T. (2017). “Intentionality and flexibility in animal communication”. Routledge Handbook for the Philosophy of Animal Minds. London

Simon, H.A. (1981) « The Sciences of the Artificial ». MIT Press: Cambridge

Soderstrom M.; Seidl A.; Kemler Nesìlson D.; Jusczyk P. (2003) “The prosodic bootstrapping of phrases. Evidence from pre-linguistic infants”. Journal of Memory and Language**.** Volume 49, Issue 2. Pages 249-267

Suire, A., Raymond, M., Barkat-Defradas, M. (2019). “Male vocal quality and its relation to females' preferences”. Evolutionary Psychology

Suire, A., Tognetti, A., Durand, V., Raymond, M., Barkat-Defradas, M. (2020). “Speech acoustic features: a comparison of gay men, heterosexual men, and heterosexual women”. Archives of Sexual Behaviour

Stapel D.; Semin G. (2007). “The magic spell of language: linguistic categories and their perceptual consequences”. Journal pf personality and social psychology. Vol. 93 No1. 23-33

Tomasello, M.; Stahl, D. (2004). « *Sampling children’s spontaneous speech: How much is enough?* » . Journal of Child Language, 31:101–121.

Vihman M. M. (2014). “Phonological Development: The First Two Years” Second Edition.

Vihman, M. M., & McCune, L. (1994). “When is a word a word?” Journal of Child Language, 21(3), 517–542

Westermann G.; Miranda E. (2004). « A new model of sensorimotor coupling in the development of speech » . Brain and language. Elsevier.Wiley & Sons, Inc.

Wintner S. (2010) Computational Models of Language Acquisition. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2010. Lecture Notes in Computer Science, vol 6008. Springer, Berlin, Heidelberg

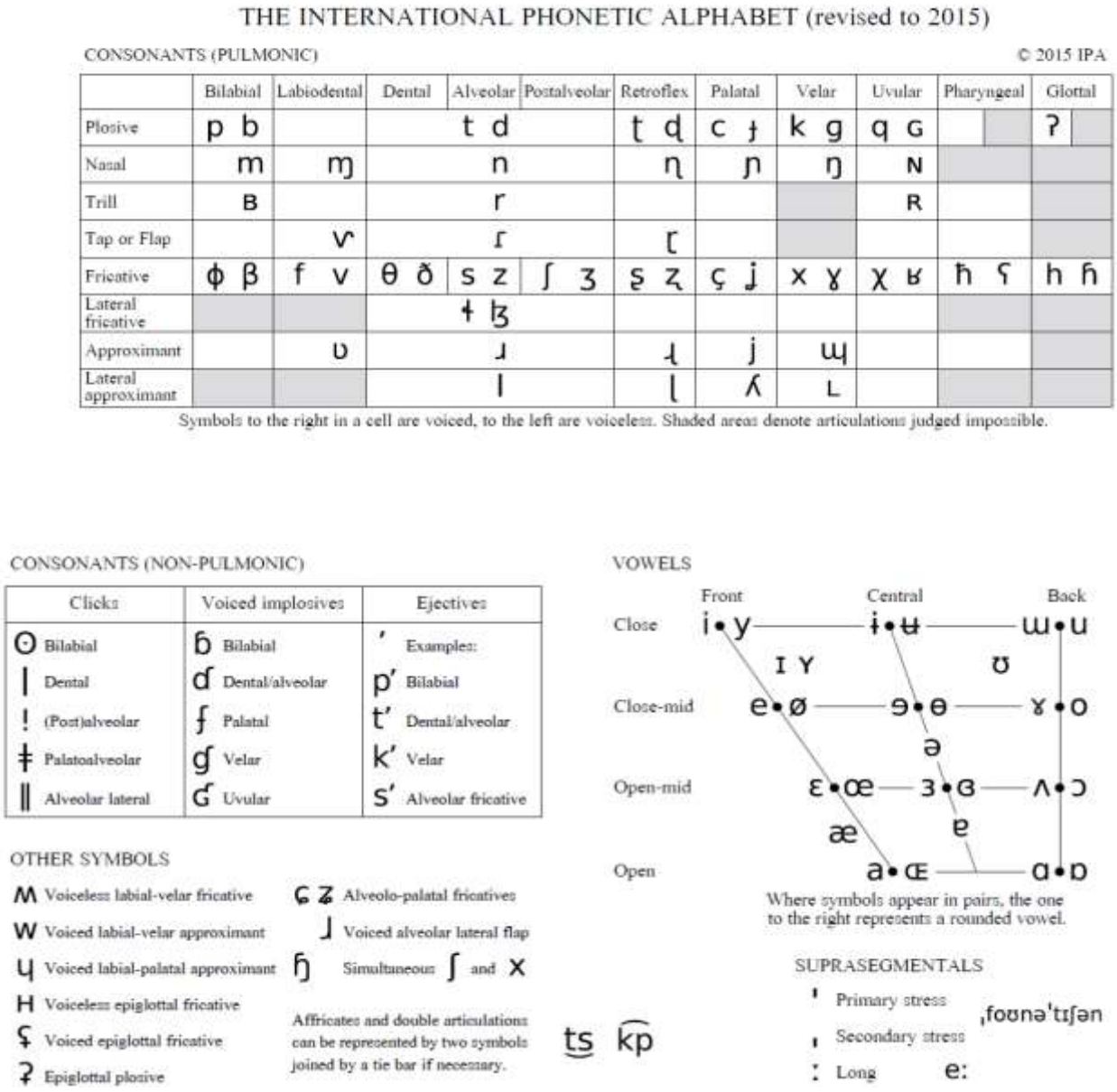
Yamaguchi N. (2012). « Parcours d'acquisition des sons du langage chez deux enfants francophones ». Phd thesis, Sorbonne University (Paris 3). Available in archives-ouvertes.fr

Yamaguchi N.; "What is a representative language sample for word and sound acquisition?". Canadian Journal of Linguistics / Revue canadienne de linguistique, University of Toronto Press, 2018, 63 (04), pp.667-685.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Addison-Wesley Press

# List of Annexes

Annexe 1) IPA Chart



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

| | Bilabial | Labiodental | Dental | Alveolar | Postalveolar | Retroflex | Palatal | Velar | Uvular | Pharyngeal | Glottal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Plosive | p b | | | t d | | ʈ ɖ | c ɟ | k g | q ɢ | | ʔ |
| Nasal | m | ɱ | | n | | ɳ | ɲ | ŋ | ɴ | | |
| Trill | ʙ | | | r | | | | | ʀ | | |
| Tap or Flap | | ⱱ | | ɾ | | ɽ | | | | | |
| Fricative | ɸ β | f v | θ ð | s z | ʃ ʒ | ʂ ʐ | ç ʝ | x ɣ | χ ʁ | ħ ʕ | h ɦ |
| Lateral fricative | | | | ɬ ɮ | | | | | | | |
| Approximant | | ʋ | | ɹ | | ɻ | j | ɰ | | | |
| Lateral approximant | | | | l | | ɭ | ʎ | ʟ | | | |

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

| Clicks | Voiced implosives | Ejectives |
|---|---|---|
| ʘ Bilabial | ɓ Bilabial | ʼ Examples: |
| ǀ Dental | ɗ Dental/alveolar | pʼ Bilabial |
| ǃ (Post)alveolar | ʄ Palatal | tʼ Dental/alveolar |
| ǂ Palatoalveolar | ɠ Velar | kʼ Velar |
| ǁ Alveolar lateral | ʛ Uvular | sʼ Alveolar fricative |

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative
w Voiced labial-velar approximant
ɥ Voiced labial-palatal approximant
ʜ Voiceless epiglottal fricative
ʢ Voiced epiglottal fricative
ʡ Epiglottal plosive

ʗ ʖ Alveolo-palatal fricatives
ɺ Voiced alveolar lateral flap
ʧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

t͡s k͡p

VOWELS




Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

ˈ Primary stress ˌfoʊnəˈtɪʃən
ˌ Secondary stress
ː Long eː

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ◌̥ | Voiceless | n̥ | d̥ | ◌̤ | Breathy voiced | b̤ | a̤ | ◌̪ | Dental | t̪ | d̪ |
| ◌̬ | Voiced | s̬ | t̬ | ◌̰ | Creaky voiced | b̰ | a̰ | ◌̺ | Apical | t̺ | d̺ |
| ◌ʰ | Aspirated | tʰ | dʰ | ◌̼ | Linguolabial | t̼ | d̼ | ◌̻ | Laminal | t̻ | d̻ |
| ◌̹ | More rounded | ɔ̹ | | ◌ʷ | Labialized | tʷ | dʷ | ◌̃ | Nasalized | | ẽ |
| ◌̜ | Less rounded | ɔ̜ | | ◌ʲ | Palatalized | tʲ | dʲ | ◌ⁿ | Nasal release | | dⁿ |
| ◌̟ | Advanced | u̟ | | ◌ˠ | Velarized | tˠ | dˠ | ◌ˡ | Lateral release | | dˡ |
| ◌̠ | Retracted | e̠ | | ◌ˤ | Pharyngealized | tˤ | dˤ | ◌̚ | No audible release | | d̚ |
| ◌̈ | Centralized | ë | | ~ | Velarized or pharyngealized | ɬ | | | | | |
| ◌̽ | Mid-centralized | e̽ | | ◌̝ | Raised | e̝ | ( ɹ̝ = voiced alveolar fricative) | | | | |
| ◌̩ | Syllabic | n̩ | | ◌̞ | Lowered | e̞ | ( β̞ = voiced bilabial approximant) | | | | |
| ◌̯ | Non-syllabic | e̯ | | ◌̘ | Advanced Tongue Root | e̘ | | | | | |
| ◌˞ | Rhoticity | ə˞ | a˞ | ◌̙ | Retracted Tongue Root | e̙ | | | | | |

ˑ Half-long eˑ

◌̆ Extra-short ĕ

| Minor (foot) group

‖ Major (intonation) group

. Syllable break ɹi.ækt

‿ Linking (absence of a break)

TONES AND WORD ACCENTS

| LEVEL | | | CONTOUR | | |
|---|---|---|---|---|---|
| e̋ or e᷅ | ˥ | Extra high | ě or e᷄ | ˧˥ | Rising |
| é | ˦ | High | ê | ˥˧ | Falling |
| ē | ˧ | Mid | e᷌ | ˦˥ | High rising |
| è | ˨ | Low | e᷈ | ˧˩ | Low rising |
| ȅ | ˩ | Extra low | e᷉ | ˩˦ | Rising-falling |
| ↓ | | Downstep | ↗ | | Global rise |
| ↑ | | Upstep | ↘ | | Global fall |

Typefaces: DejaVu Sans (symbols), Doulos SIL (menutext)

Annex 2)

Taken from Wioland, 1991, in Pierre Léon, *Phonétisme et Prononciations du français*, Paris, Nathan-Fac, 1992 / 3e édition, 1998 / 4e édition, Armand-Colin, 2005 / 5e édition, Armand-Colin

**Liste 2** **Fréquence d'occurrence des phonèmes dans le discours**

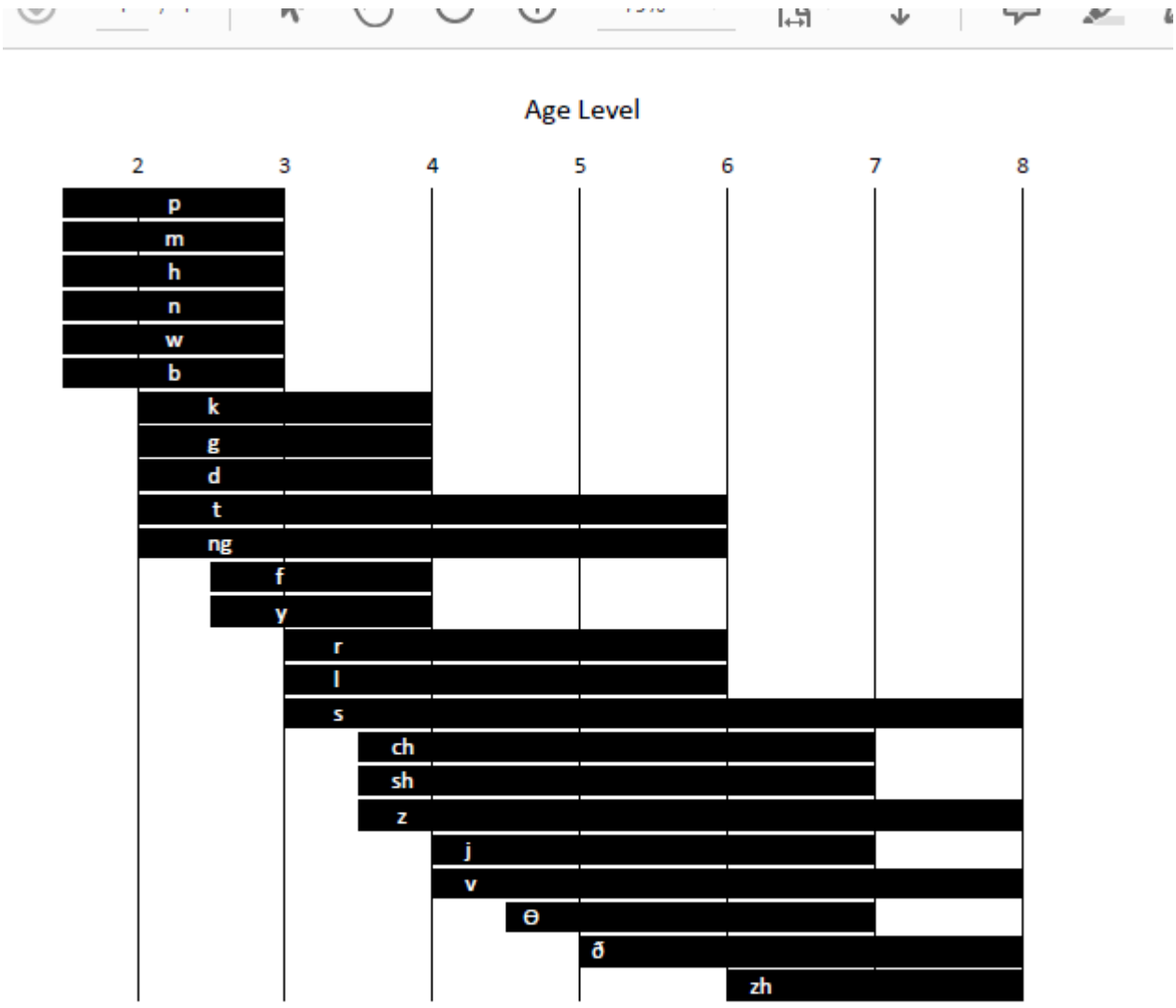
pour les consonnes

| | | | | | |
|---|---|---|---|---|---|
| 1 | – / R / | 7,25 % | 11 | – / j / | 2,00 |
| 2 | – / s / | 6,00 | 12 | – / ʒ / | 1,66 |
| 3 | – / l / | 5,63 | 13 | – / z / | 1,535 |
| 4 | – / t / | 5,335 | 14 | – / f / | 1,40 |
| 5 | – / k / | 4,06 | 15 | – / w / | 1,40 |
| 6 | – / d / | 4,035 | 16 | – / b / | 1,31 |
| 7 | – / m / | 3,845 | 17 | – / ʃ / | 0,535 |
| 8 | – / p / | 3,715 | 18 | – / ɥ / | 0,515 |
| 9 | – / n / | 3,095 | 19 | – / g / | 0,475 |
| 10 | – / v / | 2,755 | | | |
| | | | soit | | 56,55 % |

pour les voyelles

| | | |
|---|---|---|
| 1 | – / E / | 10,60 % |
| 2 | – / a / | 8,55 |
| 3 | – / i / | 5,115 |
| 4 | – / Œ / | 4,31 |
| 5 | – / O / | 3,36 |
| 6 | – / ɑ̃ / | 3,09 |
| 7 | – / u / | 2,425 |
| 8 | – / ɔ / | 2,255 |
| 9 | – / y / | 1,90 |
| 10 | – / ɛ̃ / | 1,845 |
| soit | | 43,45 % |

Annexe 2) Average age estimates of customary consonant production (1) We suppose that this is better than the next

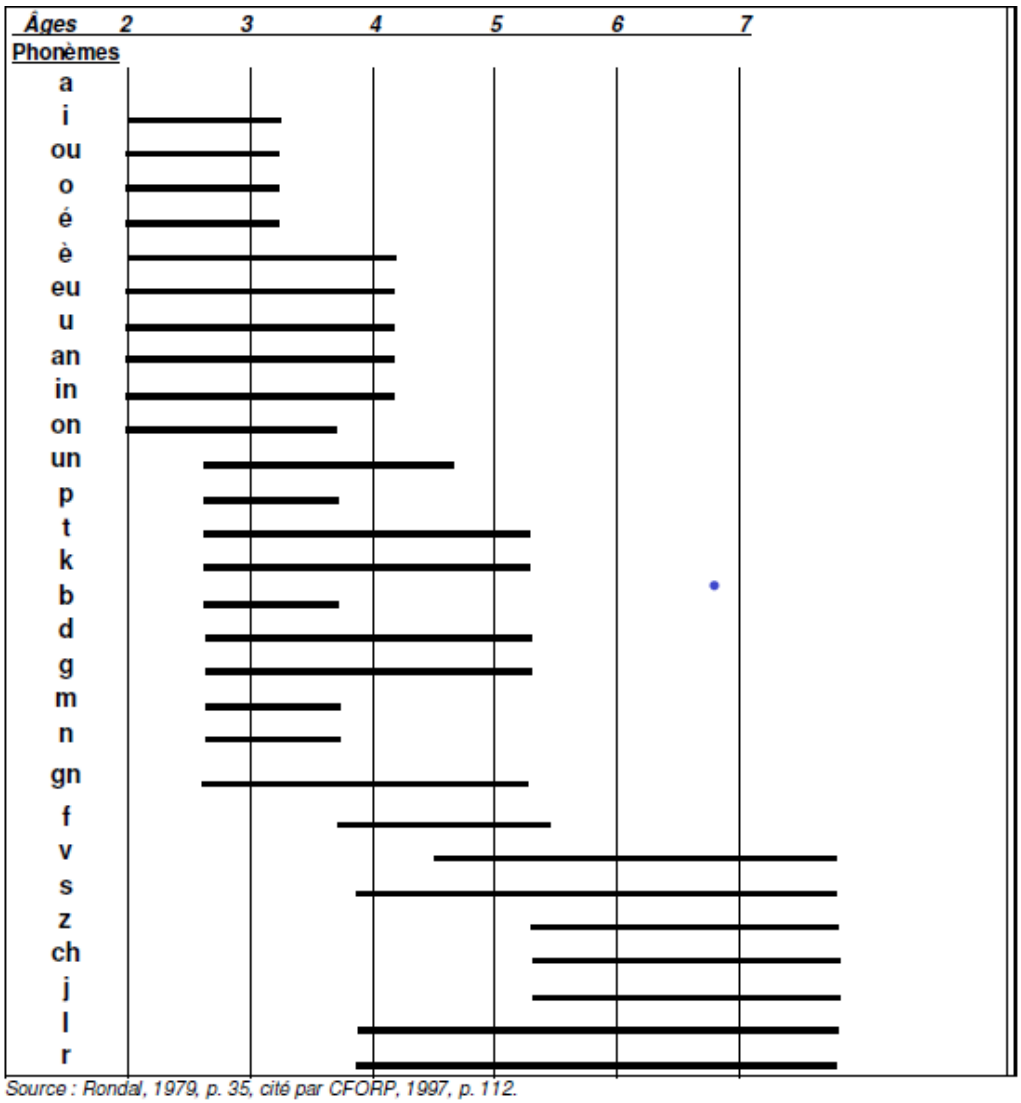



Average age estimates and upper age limits of customary consonant production. The solid bar corresponding to each sound starts at the median age of customary, articulation; it stops at an age level at which 90% of all children are customarily producing the sound (from Templin, 1957; Wellman et al., 1931). Source: Sander © 1972 American Speech-

Annexe 3)

Le tableau suivantindique des points de repèrerelatifs au moment de l'intégrationphonologique

de chacun des phonèmes (point de départ : âgeoù 50% des élèves environ prononcent le son correctement; point d'arrivée : âgeoù la grandemajorité a acquis la bonne prononciation du son).




Source : Rondal, 1979, p. 35, cité par CFORP, 1997, p. 112.

Annexe 4)

Some examples showing how to transcribe French orthographic norm into IPA symbols.

Differently from Italian, French phonetic system is composed by 11 oral vowels and 4 nasal vowels

As French is a quite complex language with an extraordinarily arbitrary relationship between sounds and graphic symbols : 36 phonetic units can be written in more than 500 different ways. Differently from Italian and Spanish orthographs (to cite two sister languages) where the relation between simbols and sounds is defined as to be « clear », in French this relation – due to its ambiguity – it is defined to be « opaque ». Thus, IPA alphabet is extremely useful for second language learners because of its univocity : it assigns a unique symbol to a given sound, giving a way to avoid ambiguity in reading and writing skills and, above all, allowing adult learners to know how to spell a word without keeping in memory a thousand of exceptions to the rule that quite often cause a cognitive overload (as well as some kind of irony on the notion of convention in itself that has been smartly represented iby the authors of the video)

A funny video to explore this subject by two belgian professors is available in this TED Talk : https://www.youtube.com/watch?v=5YO7Vg1ByA8 “La faute de l’orthographe » A. Hoedt ; J. Piron

| VOYELLES ORALES | | CONSONNES | |
|---|---|---|---|
| [ a ] | bal,roi,noyer | [ p ] | paquet, api, attrape |
| [ e ] | été,je plongeai | [ b ] | béret, abîme, snob |
| [ ɛ ] | lait, je plongeais, pêche | [ d ] | dire, Adèle, odeur |
| [ ɪ ] | mille, cygne, île | [ t ] | tas, attelage, vite |
| [ o ] | rôdé, seau, pot | [ k ] | cou, barque, chœur |
| [ ɔ ] | bol, Paul, pomme | [ g ] | goût, agapes, aguerri |
| [ u ] | loup, roux, cour | [ f ] | fou, affreux, effacer |
| [ y ] | pur, lune, but, il eut | [ v ] | vent, avenir, vert |
| [ ø ] | feu, nœud, jeûne | [ s ] | saut, essai, laisse |
| [ œ ] | beurre, fauteuil, œil, accueil | [ z ] | zouave, roseau, raser |
| [ ə ] | le, belette, lever | [ ʒ ] | je, joli, âge |
| | | [ ʃ ] | chat, lâcher, bêche |
| | | [ l ] | lire, délavé, vélo |
| | | [ ʀ ] | rire, hériter, arracher |
| | | [ m ] | mot, âme, lime |
| | | [ n ] | nous, année, panne |
| | | [ ɳ ] | oignon, cigogne, lorgner |
| | | [ ŋ ] | camping, parking |

| VOYELLES NASALES | | SEMI-CONSONNES | |
|---|---|---|---|
| [ ɛ̃ ] | simple, examen, bain | [ j ] | œil, yeux, paille, lier |
| [ ɑ̃ ] | lent, paon, chant | [ ɥ ] | puits, éternuer, nuit |
| [ ɔ̃ ] | songe, plomb | [ w ] | ouest, oui, toit |
| [ œ̃ ] | un, emprunt, parfum | | |

Annex 5)

This is a non exhaustive list of French phonetic units ordered according to their degree of articulatory effort in a bottom-up decreasing way. This is a rough schema that has been conceived in an « hand-craft » manner in order to be able to analyse in a pre-determined way infants' occurrencies that we exported from CoLaJE.

| | | | |
|---|---|---|---|
| Voyelles | /a/ /ɑ/ /i/ /u/ /ɛ/ /ɔ/ /œ/ /ə/ /ɑ̃/ /ɔ̃/ /ɛ̃/ /œ̃/ /ø/ /o//ɔ/ /y/ | Voyelles + relâchées /a/ /ɑ/ /ɛ/ /ɔ/ /œ/ | |
| | | Voyelles + tendues /i/ /e/ /o/ /ə/ /o/ /y/ /u/ | |
| | | Voyelles nasales /ɑ̃/ /ɔ̃/ /ɛ̃/ /œ̃/ | |
| Consonnes | /b/ /p/ /m/ /d/ /t/ /k/ /g/ /f/ /v/ /n/ /s/ /z/ /l/ /ʁ/ /ʒ/ /ʃ/ /ɲ/ | Occlusives bilabiales /b/ /p/ /m/ | |
| | | Autres Occlusives ● /d/ /t/ /k/ /g/ /n/ /ɲ/ | |
| | | Constrictives avec un point d'articulation précis /f/ /v/ /s/ /z/ | |
| | | Liquides /l/ /ʁ/ | |
| | | Constrictives avec un point d'articulation large /ʒ/ /ʃ/ | |
| Groupes consonantiques | /tʁ/ /kʁ/ /dʁ/ /gʁ/ | Oppositions particulières : /tʁ/ ~ /kʁ/ /dʁ/ ~ /gʁ/ | |
| Semi-consonnes | /j/ /ɥ/ /w/ | Facile : /j/ | |
| | | Oppositions difficiles : /ɥ/ ~ /w/ | |

Annex 6)

## CHILDES Informed Consent Template

Study Title: MyStudy

Principal Investigator: Name Address

phone:

email:

---

Purpose of this Study:

The purpose of the study is to gather data to be placed in a computerized data bank for the study of language and communication in (specify topic area here). Researchers will be able to access these data over the Internet.

Procedures:

(Describe and list procedures here). The session will be videotaped (or audiotaped) for later transcription and analysis. In addition, you will be asked to provide relevant demographic information. The research will be take approximately (specify duration)

of your time and will be done at (specify location).

Participant Characteristics:

Participants in this study should be (specify conditions for participation,

such as normal hearing, children learning English, or children aged 2-4)

Risks:

The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or during normal conversation, which may include frustration, fatigue, and/or boredom.

Benefits:

There is no clear personal benefit for you or your child for participation in this study.

Compensation and Costs:

You will receive $40 as well as free parking during testing. There will be no cost to you for participating in this study.

Confidentiality: To maintain confidentiality, your data and consent form will be kept separate. The consent form will be kept in a locked file. All references to your last name or address will be removed from the transcripts and recordings of the session. Your name, address, contact information and other direct personal identifiers in your consent form will not be mentioned in any such publication or dissemination of the research data.

By participating, you understand and agree that the data and information gathered during this study may be used by qualified researchers. (The following is only necessary, if password protection is required:) Access to the data will be limited by passwords that are only provided to qualified researchers.

Rights:

Your participation is voluntary. You are free to stop participation at any point. Refusal to participate or withdrawal of your consent or discontinued participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. You also have the right at any time to have your data removed from the database.

Right to Ask Questions and Contact Information:

If you have any questions about this study, you should feel free to ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact the Principle Investigator by mail, phone or e-mail in accordance with the contact information listed on the first page of this consent.

If you have questions pertaining to your rights as a research participant; or to report objections to this study, you should contact the Research Regulatory Compliance Office at Carnegie Mellon University. Email: HYPERLINK "mailto:irb-review@andrew.cmu.edu" irb-review@andrew.cmu.edu . Phone: 412-268-1901 or 412-268-5460.

Voluntary Consent:

By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You understand that you may ask questions about any aspect of this research study during the course of the study and in the future. By signing this form, you agree to participate in this research study.

Signature _____ Date: _____

Annex 7)

Consonant and semi-consonants of French. Taken from Yamaguchi, Phd thesis, available on hal.fr . This table is organised as follows: columns represent the *place of articulation* of a consonant while rows represent the *articulation mode*. This table is to be considered specific to french language

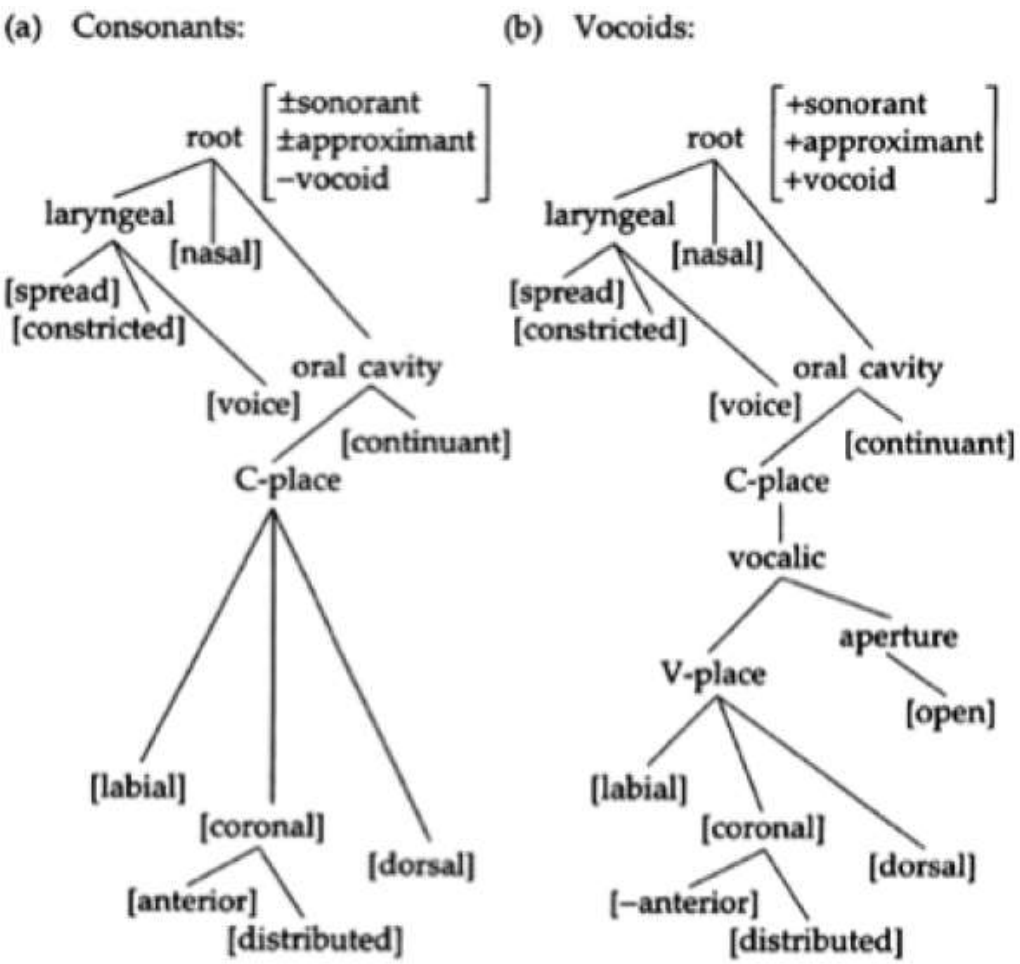
| | Bilabiale | Labio-Dentale | Alvéolaire | Post-Alvéolaire | Palatale | Vélaire | Uvulaire |
|---|---|---|---|---|---|---|---|
| Occlusives | p b | | t d | | | k g | |
| Fricatives | | f v | s z | ʃ ʒ | | | ʁ |
| Nasales | m | | n | | ɲ | | |
| Approx. | | | | | j | | |
| Approx. lat. | | | l | | | | |

Annex 8)

In this table the distinctive traits of French according to « Clements & Hume, 1995 » are listed. Taken form Yamaguchi's thesis, 2012

| **Traits** | p | b | t | d | k | g | f | v | s | z | ʃ | ʒ | m | n | ɲ | l | ʁ | j |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [±sonant] | − | − | − | − | − | − | − | − | − | − | − | − | + | + | + | + | + | + |
| [±approx.] | − | − | − | − | − | − | − | − | − | − | − | − | − | − | − | + | + | + |
| [±continu] | − | − | − | − | − | − | + | + | + | + | + | + | | | | | | |
| [±voisé] | − | + | − | + | − | + | − | + | − | + | − | + | | | | | | |
| [LABIAL] | ✓ | ✓ | | | | | ✓ | ✓ | | | | | ✓ | | | | | |
| [CORONAL] | | | ✓ | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ |
| [±post.] | | | | | | | | | − | − | + | + | | − | + | | | |
| [DORSAL] | | | | | ✓ | ✓ | | | | | | | | | | | ✓ | |
| [±latéral] | | | | | | | | | | | | | | | | + | | − |

Annex 9)




http://advanse.lirmm.fr/EMClustering/