



PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

FACULTY OF SOCIAL SCIENCES
SCHOOL OF PSYCHOLOGY

**YOUNG INFANTS CAN LEARN OBJECT AND ACTION-
WORDS FROM CONTINUOUS AUDIOVISUAL STREAMS**

BY

M^a CRISTINA JARA GONZÁLEZ

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the
requirements for the degree of Doctor in Psychology

Advisor:

MARCELA PEÑA

December 2018
Santiago, Chile
© 2018, M^a Cristina Jara González



FACULTY OF SOCIAL SCIENCES
SCHOOL OF PSYCHOLOGY

**YOUNG INFANTS CAN LEARN OBJECT AND ACTION-
WORDS FROM CONTINUOUS AUDIOVISUAL STREAMS**

BY

M^a CRISTINA JARA GONZÁLEZ

Members of the Committee:

LUCA L. BONATTI

MARCIA OLHABERRY

MARCELA PEÑA

DOMINGO ROMÁN

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the
requirements for the degree of Doctor in Psychology

December 2018

Santiago, Chile

© 2018, M^a Cristina Jara González

© 2018, M^a Cristina Jara González

Se autoriza la reproducción total o parcial, con fines académicos, por cualquier medio o procedimiento, incluyendo la cita bibliográfica del documento.

To my family and friends

ACKNOWLEDGEMENTS

I would like to thank and dedicate this thesis to my family, my parents M^a Teresa and Samuel, and my brothers Rodrigo and César, for their valuable support and unconditional love through my life. To Piloto and Pastora, for their company and make my days happier. Moreover, I would like to thank Juan Carlos Moenne and Silvia Vargas, for their support and affection through these years.

I would like to express my sincere gratitude to my advisor Prof. Marcela Peña, for giving me the opportunity to learn under her guidance, for her academic support, patience and motivation during my thesis. I will be always grateful for helping and encourage me to improve this project, and to develop my critical and scientific thinking, allowing me to grow as a better researcher.

I would also like to thank my thesis committee. To Prof. Marcia Olhaberry for her valuable comments from the Clinical Psychology. To Prof. Domingo Román, for his support and valuable teaching from the interesting field of Phonetics. Finally, to Prof. Luca Bonatti, for his brilliant comments to improve this project and for being one of the professors who encourage me to continue a scientific career since I started my master. Thank you all for the support, suggestions and valuable comments that made this thesis an amazing project.

All my deepest gratitude to my friends that made this long way much more enjoyable. Special thanks to Valeska Dote, Rodrigo Vergara, Sofía Herrera, Miguel Ibaceta and Ricardo Morales. To my fellows from the Psychology program, especially to Carolina Araya and Joel Álvarez. Thank you all for the time we spent together, for being my friends and for always being there when I needed.

My sincere thanks also go to my fellow lab partners. Especially Diana Arias and Severin Lions. It was a great experience to share the lab with them during the last years.

My gratitude to the people who helped me in the realization of this thesis, especially to Amaya Lorca, Karina Gutiérrez, Carolina Rocha and Catherine Andreu for their assistance in the creation of the stimuli.

Special thanks to Hospital Dr. Sótero del Río, especially Dra. Enrica Pittaluga, Tania Suárez, and Orieta Palacios, who were crucial in the realization of this project. My gratitude also goes to parents and the infants for their valuable interest in participating in this thesis. This project was only possible through to their valuable contribution.

Last but not least, I want to thank my life partner Cristóbal Moenne. I am very grateful for his unconditional and generous support, kindness, and love through these years. His amazing scientific mind and expertise were of great importance during the developing of this project. Meeting him was the greatest thing that happened to me during my doctorate.

This work was supported by the Chilean National Council of Scientific and Technological Research (CONICYT) National PhD grant number 21140640.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	i.
LIST OF FIGURES	viii.
LIST OF TABLES	x.
ABSTRACT	xi.
I. GENERAL INTRODUCTION	1
1.1. Audiovisual mapping during infants’ word learning	1
1.2. Vowels and Consonants and their implications during word learning	5
II. ABOUT THIS DOCTORAL THESIS	7
2.1. Objectives	9
2.2. Hypothesis	9
2.3. General Methodology	10
III. STUDY 1:	
Young infants can learn object or action words from audiovisual continuous streams: A Pilot Study	12
Abstract	13
3.1. Introduction	13
3.2. Methods	16
3.2.1. Participants	16
3.2.2. Apparatus	16
3.2.3. Stimuli	16
3.2.3.1. Familiarization phase	16
3.2.3.2. Test phase	18
3.2.4. Procedure	18
3.2.5. Data acquisition and analysis	19
3.2.5.1. Spatiotemporal regions of interest (ROI)	19
3.2.5.2. Data preprocessing	20
3.2.5.3. Visual variables	20

3.2.5.4. Data analysis	21
3.3. Results	21
3.3.1. Familiarization phase	21
3.3.2. Test phase	21
3.3.2.1. Object-and action-word learning	22
3.3.2.2. Correlations between the data from the familiarization and test phases	24
3.3.2.3. No preference for particular words	25
3.4. Discussion	25
3.4.1. Five-months-old infants can learn object and action word associations from audiovisual streams	25
3.4.2. Infants have high interest to explore audiovisual streams done with speech and faces	25
3.4.3. Five-months-old infants not always learn both type of words ...	26
3.4.4. Five-months-old equally learn object and action words	27
3.5. Conclusion	27
3.6. Author Contributions, Acknowledgments, Declaration of Conflicting Interests, Funding	28
IV. STUDY 2:	
Vowels help 8-month-old infants to learn object and action-words from audiovisual continuous streams	29
Abstract	29
4.1. Introduction	30
4.2. Results	34
4.2.1 Experiment 1. Action-word learning when words conveyed vowel harmony	34
4.2.1.1. Familiarization	34
4.2.1.2. Test	34
4.2.2. Experiment 2. Object-word learning when words conveyed vowel harmony	35
4.2.2.1. Familiarization	35

4.2.2.2. Test	36
4.2.3. Experiment 3. Action-word learning when words conveyed consonant harmony	36
4.2.3.1. Familiarization	36
4.2.3.2. Test	37
4.2.4. Experiment 4. Object-word learning when words conveyed consonant harmony	37
4.2.4.1. Familiarization	37
4.2.4.2. Test	38
4.2.5. Analysis between experiments	39
4.2.5.1. Evaluating sociodemographic differences across experiments	39
4.2.5.2. Familiarization across experiments	39
4.2.5.3. Learning during test across experiments	39
4.2.5.4. Comparing the learning in experiments with vowels and consonant harmony	40
4.2.5.5. Age as a covariable	40
4.3. Discussion	42
4.3.1. Eight-months-old infants learn object and action-words from continuous audiovisual stimuli	42
4.3.2. Faces as facilitator of audiovisual processing	42
4.3.3. Phonological cues: phoneme harmony	43
4.3.4. Face-speech association as inter-sensorial binding	45
4.3.5. Object- versus action-word learning	45
4.4. Conclusions	46
4.5. Materials and methods	47
4.5.1. Experiment 1. Participants	47

4.5.2.	Experiment 2. Participants	47
4.5.3.	Experiment 3. Participants	47
4.5.4.	Experiment 4. Participants	48
4.5.5.	Stimuli	48
4.5.5.1.	Auditory stimuli	48
4.5.5.2.	Visual stimuli	49
4.5.5.3.	Audio-visual stream	49
4.5.6.	Procedure	50
4.5.7.	Analysis Overview	52
4.5.7.1.	Spatiotemporal regions of interest (ROI)	52
4.5.7.2.	Data pre-processing	53
4.5.7.3.	Visual variables	53
4.5.7.4.	Data analysis	54
4.6.	SUPPLEMENTARY MATERIAL	56
V.	EXPERIMENT IN PROGRESS: Generalization of action-words	59
5.1.	Purpose of the experiment	59
5.2.	Participants	59
5.3.	Stimuli and apparatus	60
5.4.	Procedure	62
5.5.	Data acquisition and analysis	63
5.6.	Preliminary results	63
5.6.1.	Familiarization phase	63
5.6.2.	Test phase	63
5.7.	Preliminary Discussion	64

VI. GENERAL DISCUSSION	65
6.1. Significance	67
6.2. Limitations and projections	68
VII. ADDITIONAL OUTPUT DURING THIS DOCTORAL THESIS	
Cognitive Models of Language (Book chapter)	70
VIII. OTHER ADDITIONAL OUTPUT DURING THE DOCTORAL PROGRAM	91
8.1. Publications	91
8.2. Book Chapters	91
8.3. Conference Posters presentations	91
8.4. Participation in other research projects	92
REFERENCES	93

LIST OF FIGURES

Fig. 1.1. Schematic model of the different studies conducted in this doctoral thesis	11
Fig. 3.1. Continuous audiovisual stream used during familiarization phase. Four novel words were concatenated into a continuous stream. Object-words co-occurred with stationary images and action-words with moving images	18
Fig. 3.2. Trial structure for the test phase. We illustrate the timing of the events and the time window for the analysis in the trials evaluating the object-words (A) and action-words (B)	19
Fig. 3.3. In each infant, we plot the mean TTL ratio greater than chance (0.5) for object-word (left side) and action-word (right panel). We draw a black line to illustrate the 4 infants with a mean TTL ratio lower than 0.5	23
Fig. 3.4. The time dedicated to exploring the target during the familiarization increase the probability to learn the image-word associations	24
Fig. 4.1. Infants succeeded to find object-words (Exp.1) and action-words (Exp.2) when words had harmony vowel. We plot the mean proportion for correct responses for longest gaze (LG) and mean number total looking time (TLTa), compared with a proportion expected at chance (0.5). Circles represents each infant for each experiment. (* = <.05; † = <.1)	38
Fig. 4.2. The figure illustrates bivariate correlations between TLTa in vowel (in blue) and consonant experiments (in orange) against the age of evaluation	41
Fig. 4.3. Continuous audiovisual stream used during familiarization phase for all experiments. Four novel words were concatenated into a continuous stream. Each word co-occurred with one of four possible visual stimuli. Audio stimuli for each experiment are presented in the bottom of the illustration	50
Fig. 4.4. Trial structure during the test phase. We illustrate the trial evaluating the learning of the object-words (A) and action-words (B). Time window of analysis is presented in the bottom of the illustration	52

Fig. 5.1. Continuous audiovisual stream used during familiarization phase. Four novel words were concatenated into a continuous stream. Object-words co-occurred with stationary images and action-words with moving images 62

Fig. 5.2. Trial structure during the test phase. We illustrate the trial evaluating the learning of action-words. Time window of analysis is presented in the bottom of the illustration 62

Fig 5.3. Non-significant results were obtained for all eye-tracking variables. We plot the proportion for target looking, for longest gaze (LG), total looking time (TLTa and TLTr), and First gaze (FG) compared with a proportion expected at chance (.5). Circles represents each infant for each experiment 64

LIST OF TABLES

Table 4.1. Infants gender and Mothers' educational level distribution for Study 2	56
Table 4.2. Descriptive demographic variables for Study 2	57
Table 4.3. Infant Behavior Questionnaire – Revised (IBQ-R) (Putnam, Helbig, Gartstein, Rothbart & Leerkes, 2014)	58
Table 5.1. Sound and visual stimuli for Experiment in progress	61

ABSTRACT

This doctoral thesis explored whether and how preverbal infants can discover novel words associated with objects and actions from a continuous audiovisual stream. Specifically, we tested infants on their ability to associate a novel word that systematically co-occurred with a stationary face (for object-words) and a novel word that systematically co-occurred with moving faces (for action-words). Moreover, this doctoral thesis seeks to explore what phonological cues infants rely on and will help to succeed in this learning process and also if infants can generalize the learning of action-words into new visual stimuli.

In study 1 we tested 5-month-old infants on the learning of these two types of words from a continuous audiovisual task. The results showed that in global, most infants learned the image-word associations, however, some of them learned only the object-words, other only the action-words and others succeeded to learn both. We interpreted such pattern of the results as an index of the interindividual variability especially related to the cognitive resources that the infants recruited to solve the task.

In study 2, in a series of 4 experiments, we evaluated older infants i.e. 8-month-old infants and we facilitated the task by adding vowel or consonant harmony to either the object or the action words. The results showed that infants were able to learn both objects and action-words from the continuous audiovisual streams only when words conveyed vowel harmony. We interpreted these results as vowels may have a role in early ages due to the perceptual salience in audiovisual task.

An additional experiment in progress, evaluated whether ~7-month-old infants generalized the learning of novel action-words to novel agents (novel moving faces not presented during familiarization). Our results indicated that infants failed to generalize action words to new agents. In sum, the results of these thesis showed that from very early ages infants can associate what they hear with what they see from continuous audiovisual streams, when audiovisual stimuli are interesting and facilitated the task. Such initial associations might correspond to primary forms of the word-mapping observed later steps of language acquisition.

I. GENERAL INTRODUCTION

During language acquisition, infants are exposed to a natural environment on which words forms are linked to particular visual stimuli. Words are embedded in fluent speech and most of the visual scenes are inserted in complex scenarios. Because of this, infants must develop different skills in order to extract words from continuous speech and map the newly words to their specific visual referents.

Two important type of words that infants must learn are those that refers to objects and those that refers to actions. In this thesis, we tested infants between 5 and 8-months of age on their ability to associate novel words with novel visual stimuli, as stationary faces i.e. object-words, or moving faces, i.e. action-words, in order to contribute with empirical evidence about the initial steps of the word-mapping process.

The different studies composing this doctoral thesis represent a proposal of a mechanism on weather and how preverbal infant learn these two types of words from continuous stimulation.

Study 1 explored the question if 5-month-old infants were able to learn these two types of words from an audiovisual statistical learning task. Study 2, explored in four experiments whether phonological cues (i.e. vowel and consonant harmony) helped 8-month old infants object and/or action word learning. An additional experiment, still uncompleted, explored whether ~7-months-old infants generalized the learning of action-words referred to a person to novel persons.

In the following sections, we present the theoretical framework on which this doctoral thesis is based, follow by the objectives, hypotheses, experimental design and general methodology of the thesis project.

1.1. Audiovisual mapping during infants' word learning

In a natural context, infants are exposed to different types of words embedded in fluent speech, and to objects and actions that are part of complex visual scenes. Infants,

that are exposed to a multisensory input, must advance in encode the concurrent audiovisual information into single audiovisual associations, and to save such associations in memory for further recognition. One important mechanism present in infants from early ages as 6-months, is the multisensory binding on which infants are capable of fusing audio and visual stimuli in order to create a simultaneous percept (Kopp, 2014). Other important mechanism is the analysis of distributional information of the stimuli, presented for instance in speech and complex visual scenes. Indeed, 8-month-old infants are sensitive to the transitional probabilities embedded in speech signal that help them to extract words forms (Saffran, Aslin & Newport, 1996). This study reported that infants were able to extract from an artificial language, chunks of potential words using cues as transitional probabilities presented in the continuous stream. Furthermore, infants from a sequential visual streaming can extract chunks of static shapes that co-occurred with a high transitional probability between them (Kirkham, Slemmer & Johnson, 2002) and also, chunks composed by different motions (Roseberry, Richie, Hirsh-Pasek, Golinkoff & Shipley, 2011).

Extracting words forms from the continuous speech is conceived as an important prerequisite to forming word-object pairings. Some authors have proposed that infants can track the statistical relation between word forms and objects that co-occurred simultaneously, when presented isolated, deducing a form-referent mapping (Smith, Suanda & Yu, 2014, for a review). However, to succeed in the word-mapping learning the extraction of words and visual referents is not enough, but requires a high-level perceptual binding, where the auditory and visual stimuli converge into a single, and may be amodal, representation (Miller & D'Esposito, 2005). Although such high-level inter-sensorial integration are weak during early infancy, starting to develop by the end of the second year (Shaw & Bortfeld, 2015), it seems to relate to low-level multisensory associations that infants handle soon after birth. Low-level multisensorial association would thus uncover the neonate's capacities to detect and process the synchronic properties in space and time of the cross-modal stimuli (e.g. Lewkowicz, 2000).

When infants learn the association between a word form and a visual referent, they must learn how they match each other, for instance, they must learn whether faces and lips information match the speech signal, or whether spoken words associate with specific objects and events. For instance, infants can link specific movements of the visual articulators to an auditory speech signal, during an audiovisual task (Lewkowicz & Pons, 2013).

While infants are very sensitive to face-to-face interaction from birth (Guellaï, Coulon & Streri, 2011), and able to match the phonetic information they hear with faces and lips by the 2nd month of age (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003), it seems interesting to test audiovisual processing using faces as visual stimuli.

For word-object associations, the empirical evidence shows that this ability is present from very early ages. A recent neural study using electroencephalogram (EEG), reported that 3-month-old can associate novel words with novel objects (Friederich & Friederici, 2017). The results of this study found that preverbal infants present an increase in the late negativity component that refers to be sensitive to word comprehension in young children (Conboy & Mills, 2006)

When infants associate novel word forms with objects, they can exploit different cues in order to succeed in this process. For instance, around 6-months of age infants can map a novel word with a visual referent, only when this was aligned with a phrasal prosodic boundary (Shukla, White & Aslin, 2011). Infants at this age were able to segment a statistically defined novel word with a target object, only when this was aligned with an intonational phrase.

Another study found that by 8-months of age, infants can learn word-object associations depending on the type of motion (Matatyaho-Bullaro, Gogate, Mason, Cadavid, Abdel-Mottaleb, 2014). At this age infants associate a specific label with an object, only when this object was accompanied with a familiar motion such as looming or shaking. Thus, the salience of these two familiar motions helped infants to map novel words with a specific object. This result can be found when adults accompanied objects with gestures when communicate with infants, increasing the salience of objects.

Word-action associations also have been reported in early language acquisition when audio and visual component are presented isolated. By 7-months of age, infants can relate vowel sounds with moving object (Gogate & Bahrick, 1998). In different controlled conditions, infants were tested on their ability to relate synchronous vocalizations (e.g. /a/, /i/) with moving objects after a brief exposure. The importance of this study, is that the infants only were able to detect the relations when the moving objects were in temporal synchrony with the vocalization, and not in the moving-asynchronous or when the object was still. Another study found that by 8-months, infants were able to associate a novel word into a specific visual action (Gogate & Maganti, 2017). Infants were habituated to a novel word that represent two specific motions (i.e. looming and shaking), performed by one object. In a switch paradigm, infants were habituated to the presentation of different objects performing two specific motions. After being familiarized, infants increase their looking behavior when the labels were interchanged during testing, concluding a previous learning of this novel words.

All the previous studies, explored whether infants do learn associations between isolated words with isolated images or motions. However, there is one study that explores infants' ability to identified word-objects relations during the presentation of a continuous audiovisual statistical stream. Thiessen (2010) explored this question with 8-month-old infants and adults on three different experiments. Both groups were familiarized with a continuous audiovisual stream created with transitional probabilities between syllables and words, that co-occurred synchronously with visual shapes. This continuous stream, could have been presented in one condition solely as fluent speech (no-video condition); as an audio-visual task were a specific word was presented consistently with a specific visual shape (regular-video condition), or an audio-visual task were words were not presented consistently with the visual shapes (irregular-video condition). This study found that, both adults and infants were able to segment fluent speech using the statistical information embedded in the continuous audio stream, but only adults were able to benefit for the consistent presentation of word-object associations in order to learn these relations. Only adults were able to do the audiovisual mapping process, whereas infants only were

able to segment the continuous speech stream, but not associate the extracted words with the visual referents. Despite the negative results on the infants' experiment, this study is a first approach about measuring continuous audiovisual statistical task in younger infants.

Together the previous reports showed that infants from very early ages are able to create low-level associations between a specific word form with specific visual image. The extraction of image-word elements from continuous audiovisual stimulation is subject of intense research, since it would be at the basis of the word-mapping processes.

1.2. Vowels and Consonants and their implications during word learning.

During language acquisition, infants must learn the different language units. One of the most important language unit are phonemes, which are the basic linguistic element that may bring a change of meaning (Cruttenden, 2014).

Infants must learn the repertory of phonemes of their native language, including vowels and consonants. These sounds are present in all languages around the world and are part of the sound structure of speech, carrying different and specific properties (Ladefoged, 1993).

One of the differences between vowels and consonants is that during the articulation of these two sounds, vowels have a higher opened mouth channel and a greater number of vibrations of the vocal cords per time unit than consonants, giving to vowels an increased sound and a higher tone. During the articulation of consonants, there is a different position and a greater intervention of the articulators, which consequently produce more complex acoustic signals. In the case of vowels, the airflow does not present interruption following the vibration of the vocal cords, producing different tones by the mobility of the mouth channel and the position of the tongue. As per consonants, in addition to the vibration of the vocal cords, that which may be present or not, there is an airflow interruption by resonators (Hidalgo & Quilis, 2012).

Vowels are also more salient than consonants in the speech signal. Mehler, Dupoux, Nazzi, & Dehaene-Lambertz (1996) proposed that the segmentation of utterances from a continuous speech is a result from an early sensitivity to the rhythm class of the

native language, on which vowels have a major role. Specifically, vowels carry prosodic information that allows infants to distinguish different kinds of words.

In the acquisition of consonants and vowels, evidence has shown a difference in the age of acquisition, on which by 6-months of age, infants have already acquired a significant part of the native vowel repertory (Kuhl, Williams, Lacerda, Stevens & Lindblom, 1992) whilst the native consonant repertory is observed after 10-months of age (Werker & Tees, 1984).

Because of the differences that vowels and consonants present, some authors have proposed that these two categories may have different implications during language acquisition. Nespor, Peña & Mehler (2003) following this line of proposal, explain that both categories have a division of labor during the acquisition of language in infants. This theory, called the Consonant-Vowel hypothesis (CV hypothesis) mentions that vowels have implications on the recognition of some linguistics properties like the rhythmic class and specific properties of syntactic structure (rule learning), whereas consonants participate during lexical processing. Indeed, several evidences shows that infants from early ages are capable to differentiate rhythm classes between languages, helping themselves to identify the phonological categories from the native tongue and analyze continuum speech (Mehler, Jusczyk, Lambertz, Halsted, Bertocini & Amiel-Tison, 1988; Moon, Panneton-Cooper & Fifer, 1993). Moreover, the implication of vowels during the learning of rule structure presented in language have been also well documented (Gervain, Macagno, Cogoi, Peña & Mehler, 2008; Pons & Toro, 2010; Hochmann, Benavides-Varela, Nespor & Mehler, 2011).

For consonants, their implication during word learning have also been well documented in older infants (12-month-old, Hochmann, et al. 2011; 11-month-old, Poltrock & Nazzi, 2015) and adults (Toro, Nespor, Mehler & Bonatti, 2008; Toro, Shukla, Nespor & Endress, 2008).

However, recent evidence also suggests that during the first half of the first year of life, infants rely more on vowels than consonants during the word learning process. Indeed, since infants are newborns rely on vowel sound, and not on consonants, in order

to memorize new words (Benavides-Varela, Hochmann, Macagno, Nespor & Mehler, 2012) and differentiate novel words relying primarily on vowels sounds (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy & Mehler, 1988). Furthermore, recent evidence suggests by 8-months of age can segment fluent speech streams relying on a phonological property as vowel harmony (Mintz, Walker, Weldon & Kidd, 2018),

Moreover, Bouchon, Floccia, Fux, Adda-Decker & Nazzi (2015) found that 5-month-old infants rely more on vowels than on consonants during the recognition of their own names. Infants at this age, discriminate the phonemic contrast of the first syllable of the names and pay attention to minimal differences. Infants notice the differences of their own names based on vowel sounds and not on consonants. These authors postulate that at the beginning of word learning, when infants start to segment continuous speech to find word forms, vowels could have a major role in their recognition, despite all the data that consonants participate in lexical processing with older infants. Moreover, Hochmann, Benavides-Varela, Fló, Nespor & Mehler (2017) replicated their previous study with 12-month-old infants (Hochmann et al. 2011), finding that by 6-months of age infants rely on vowels, and not on consonant as their previous study, during the learning of novel words.

Due to this evidence previously exposed, some authors have proposed a vowel bias during the first months of age for word learning processing, changing to a consonant bias by the end of the first year of life (Nazzi, Poltrock & Von Holzen, 2016).

In sum, all the previous evidence suggests that vowels and consonant may play a role during the learning of novel words, whereas the previous evidence suggest that infants may rely on vowels sounds in the first months of life, and on consonants by the end of the first year.

II. ABOUT THIS DOCTORAL THESIS

Taking together the previous evidence, we propose a series of studies directed to explore whether and how preverbal infants can discover novel words associated with objects and actions from a continuous audiovisual stream. Specifically, this thesis comprised several studies that tested infants on their ability to associate a novel word that

systematically co-occurred with a stationary face (for object-words) and a novel word that systematically co-occurred with moving faces (for action-words). Moreover, this doctoral thesis also explored on what phonological cues infants rely in order to succeed in this learning process and also, if infants can generalize the learning of action-words into new visual stimuli.

To our knowledge, only one study explored the learning of words related to static images in 8-month-old infants during the presentation of a continuous audiovisual statistical learning task, on which words were presented in a continuous speech stream, and continuously associated with a specific visual referent (Thiessen, 2010). However, this study showed that infant have difficulties to discover audiovisual associations from such type of audiovisual streams, when visual stimuli are static geometric shapes, indicating that at that age infants would rely on auditory more than on visual stimuli to find words from this type of stimuli.

Based on this previous finding, we first (study 1) tested 5-month-old infants during the learning of novel words that were continuously associated with a specific visual stimulus. For this study, and based on the Thiessen (2010) results, we introduce human faces as visual stimuli, because have been reported to be highly salient for young infants (Gliga, Elsabbagh, Andravizou & Johnson, 2009; Valenza, Simion, Macchi Cassia & Umiltà, 1996). Additionally, during the same task we also measure the learning of novel action-words that were associated with specific visual motions performed by the same female faces. We expected to find in this study, that infants will segment, extract and memorize the novel words that refers to object and action-words from the continuous audiovisual stimulation presented.

In a second study (Study 2) we evaluated if 8-month-old infants were able to learn a novel object or action-word from a continuous audiovisual stream, relying on a specific phonological cue implemented in the fluent speech, such as vowels or consonants. Taking the salience of the visual stimuli (human faces), and the salience presented in both vowels and consonants, our expectations for this study were that infants at this age will be able to

learn both objects or action-words relying mainly on vowels than in consonants sound according to the previous evidence exposed.

The last study presented in this thesis project (Experiment in progress) aimed the question if infants will be able of generalized only the action-words learned during the stimulation of the continuous audiovisual stream, into new visual exemplars.

We believe that all these studies together, will contribute to a better understanding about the exploration of the early steps of intersensory processing and its role on word learning during language acquisition.

2.1. Objectives

The main objective of this doctoral thesis is to experimentally investigate, if pre-verbal infants between 5 and 8-months of age, can learn the association between a novel word with an object or an action visually presented. Moreover, we also seek to evaluate whether phonological cues may facilitate this task, considering that young infants have a bias to exploit vowels for word learning.

Specifically, we were interested in:

- a) Explored if 5 and 8-month old infants can learn object and action-words in a continuous audiovisual statistical learning task. (Study 1, 2 and Experiment in progress)
- b) Studying the role of phonological cues (vowel and consonant harmony) for object and action-words learning (Study 2).
- c) Assessed infants' ability to generalize the learning of novel action-words to new visual exemplars (Experiment in progress).

2.2. Hypothesis

We posit that infants between 5 and 8-months of age would be able to extract the novel words representing visual objects and actions from a continuous audiovisual stream.

Namely, 5 and 8-month-old infants will be able to learn and associate the novel object and action words into stationary faces (visual objects) and moving faces (visual actions). Moreover, we also hypothesize that infants will use mainly vowel harmony as a cue for the learning novel words referring to actions and objects.

Specifically, 5 and 8-month-old infants:

- a) Will segment, extract and learn the novel words associated to objects and actions from the continuous audiovisual speech signal (study 1, study 2 and Experiment in progress)
- b) Will exploit vowel but not consonant harmony as a phonological cue for object and action-word learning (Study 2)
- c) Will generalize the learning of novel action-words associated to a specific visual stimulus (specific head movement), into novel visual exemplars (novel moving faces) (Experiment in progress).

2.3. General methodology

This doctoral thesis project is experimental and cross-sectional, because we evaluated a random sample of infants between 5 and 8-month-old in the ability of learning object and action-words during a specific developmental moment. The experimental procedure evaluated infants in a statistical learning task, and the possible strategies used by infants during learning.

In all the three studies presented in this doctoral thesis we used the same experimental paradigm but with different questions, using a remote Eye-tracking technique to measure infants looking behavior to these tasks.

In study 1 we evaluated a group of 5-month-old infants on their ability to learn objects and actions-words from a continuous audiovisual stream.

This study is the starting point for this doctoral thesis. Taking the results of this study, we design a new study (Study 2), comprising four experiments to evaluate 8-month-old infants on what phonological cues rely during the learning of this two types of words.

Furthermore, an additional experiment in progress was conducted in order to test the generalization of the action-words learned into new visual exemplars.

All the pre-analysis performed in this thesis project were mainly focused on traditional different looking behavior variables taken from eye-tracker previous infants' studies, follow by traditional statistical analysis commonly used by the scientific community on this type of experimental paradigms.

The following schematic model represent the different studies conducted in this doctoral thesis:

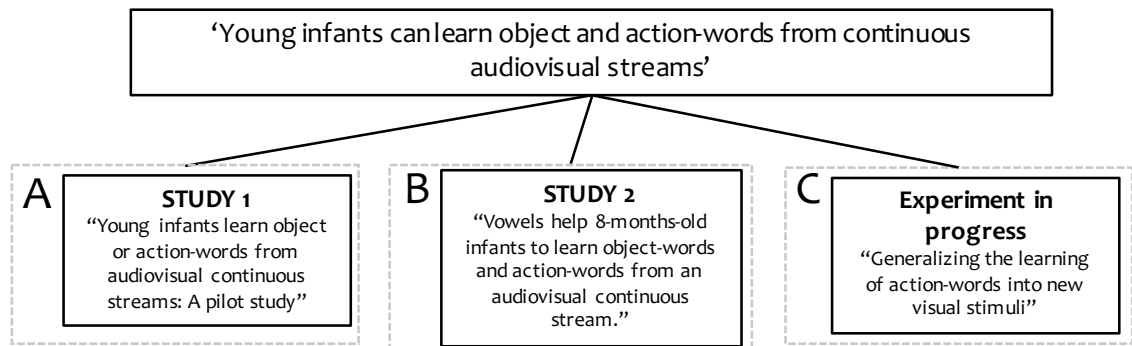


Fig. 1.1. Schematic model of the different studies conducted in this doctoral thesis.

III. STUDY 1

Young infants can learn object or action-words from audiovisual continuous streams: A pilot study.

Cristina Jara^a, Cristóbal Moënné-Loccoz^{b,c} & Marcela Peña^a

^aLaboratorio de Neurociencias Cognitivas, Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile.

^bDepartamento de Neurociencia, Facultad de Medicina, Universidad de Chile, Santiago, Chile.

^cBiomedical Neuroscience Institute, Facultad de Medicina, Universidad de Chile, Santiago, Chile.

This manuscript was submitted to the journal *Mind, Brain and Education*. Currently is under review.

ABSTRACT

We explored whether and how 5-month-old infants discovered novel words associated to objects and actions from a continuous audiovisual stream, where specific words systematically co-occurred with a stationary face (for object-words) or with moving faces (for action-words).

Our results show that some infants learned either object or action words, while others succeeded to learn both. We interpreted such pattern of the results as an index of the interindividual variability especially in the cognitive resources that the infants recruited to solve the task.

In sum, our results showed that from very early ages infants can associate what they hear with what they see from continuous audiovisual streams. Such initial associations would correspond to primary forms of the word-mapping observed later steps of the language acquisition.

Keywords: audio-visual learning; word learning; word-mapping; action-word; object-word

3.1. INTRODUCTION.

By the end of their first year, infants are able to learn, in an apparently highly unsupervised way, that the words they heard map specific objects or events. This knowledge takes place although the environmental stimulation usually involves concurrent multisensory information (e.g. Gibson, 1966), where the words mostly occur embedded on fluent speech and the visual referents are inserted on complex visual scenes. To succeed in such word-mapping learning, infants seem to be endowed with cognitive mechanisms which allow them to segment the speech signal and visual scenes into audio and visual chunks, to encode the concurrent audiovisual information into single audiovisual associations, and to save such associations in memory for further recognition. Two of such mechanisms that the infant mind handles from early development are the

analysis of the distributional information of the stimuli and the capacity to bind multisensory information into a common referent (Kopp, 2014).

Previous studies demonstrated that 8-months-old infants do segment artificial speech streams at the syllable boundaries where the transitional probability (TP) between adjacent syllables drops, grouping syllables with high TP into chunks that infants would conceive as potential words (Saffran, Aslin & Newport, 1996). Similarly, 2-days-old neonates (Bulf, Johnson & Valenza, 2011) and 2,5 and 8-month-old infants also succeed to segment visual sequences into 2-images (Kirkham, Slemmer & Johnson, 2002) or 3-images chunks (Roseberry, Richie, Hirsh-Pasek, Golinkoff & Shipley, 2011), grouping the images with high adjacent TP.

To succeed in the word-mapping learning the extraction of words and visual referents is not enough, but requires a high-level perceptual binding, where the auditory and visual stimuli converge into a single, and may be amodal, representation (Miller & D'Esposito, 2005). Although such high-level inter-sensorial integration are weak during early infancy, starting to develop by the end of the second year (Shaw & Bortfeld, 2015), they appear to relate to low-level multisensory associations that infants handle soon after birth. Low-level multisensorial association would uncover the neonate's capacities to detect and process the synchronic properties in space and time of the cross-modal stimuli (e.g. Lewkowicz, 2000). Early low-level audiovisual associations for words and images might thus provide a primary framework for further word-mapping. Data supporting this view has been provided by brain studies, exploring the object-word learning from isolated audiovisual presentations in 3-months-old infants (Friedrich & Friederici, 2017). Evaluated with electroencephalogram (EEG), infants displayed an increase in the amplitude of the late negativity component, during object-word repetitions, where the late negativity has been reported as sensitive to word comprehension in young children (Conboy & Mills, 2006).

To our knowledge, to date only one study has explored whether infants learn object-word associations from continuous audiovisual artificial streams (Thiessen, 2010). This study showed that 8-months-old infants have difficulties to discover audiovisual

associations from such type of streams, at least when visual stimuli are static geometric shapes.

The current study wanted to contribute to the research on that subject. We evaluated whether and how 5-months-old infants learn to match specific trisyllabic words to either specific object, i.e. human female faces, or to specific actions, i.e. head motions, after been familiarized with a 2-min long continuous audiovisual stream. The speech component of the audiovisual stream was artificial, built by concatenating 4 different trisyllabic non-sense words, played monotonously, and having high transitional probability (TP) within words, TP reductions between words. The visual component was constructed by concatenating eight visual images with a TP equal to .5 between any of them. Four images corresponded to two different stationary pictures of each one of two different female faces (including heads and shoulders), and four images were videos showing the same female face appeared as stationary images but now making a horizontal or a vertical head movement, similar to those used for negation and affirmation in many cultures. The statistical properties of our audiovisual stimuli allowed us to provide a continuous environment from where infants could not only extract words and images because their TPs, but could associate them upon the basis of their co-occurrence and covariance.

We evaluated the learning of what we called object-words (i.e. words associated to stationary faces) and action-words (i.e. those associated to head movements) by using remote eye tracking. We expected that infants who succeeded in object and/or action-word learning would look longer and/or faster to the correct image when they would be called to match an isolated trisyllabic word to one of two stationaries or moving images presented simultaneously. We also anticipated planned comparisons to test if at 5-months of age infants did learn only one type of word.

3.2. METHODS.

3.2.1. Participants.

We evaluated 32 5-months-old infants (Mean = 4.8, 16 girls; age range from 3.5 to 5.7 months). All infants were born as full term (40 ± 2 weeks of gestation) developed in a Spanish monolingual, middle-low socio-economic level environment, from and had not history of any atypical sensorial or neural development or any clinical condition that may influence the cognitive development and confound our results. Thirteen infants were excluded from the analysis because they did not complete the experimental protocol ($n = 7$) or had less than 4 valid trials (see below) evaluating the learning of each type of word ($n = 5$), remaining a final sample of 19 infants. Infants were recruited at public primary health centers, where they regularly attend for preventive control. Parents should sign a written consent form to participate in the study. The study received the approval from the regional Ethics committee.

3.2.2. Apparatus

Visual stimuli were displayed on a 17-inch eye-tracker monitor. In 22 infants, we used the eye tracker Tobii T120, 60 Hz sample rate, screen resolution 1280 x 1024 pixels and, in 10 we recorded with a Tobii 1750, 50 Hz sample rate, screen resolution 1024 x 768 pixels both systems with true 8-bit color depth. Both trackers are remote and recorded the infant's binocular eye fixations.

3.2.3. Stimuli

3.2.3.1. Familiarization phase. We created one speech and one visual stream separated, to then combined them into a continuous audiovisual stream.

Speech stream. We used 12 consonants and 5 vowels to construct 4 nonsense trisyllabic words: *puliso* and *tofanu*, which co-occurred with stationary images (i.e. object-words), and *degova* and *mabeki*, which associated with moving images (i.e. action-words). We

created the artificial continuous speech stream by concatenating the 4 nonsense words with the restriction that object and action-words always alternated, taking care that the TPs between adjacent syllables were equal to 1.0 within words and .5 between words across the stream. Each word was 1368 ms long (i.e. 170 ms per consonant and 286 ms per vowel). We used MBROLA (The MBROLA project, <http://tcts.fpms.ac.be/synthesis/mbrola.html>), an open source text-to-speech software to synthesize the speech stream, using French diphones (i.e. database *fr4*). The full duration of the speech stream was 2.188 min.

Visual stimuli. We constructed the continuous visual sequence by concatenating four different images of the face of two different women. Two face images of each woman appeared as stationary pictures, always synchronized with the object-words, while the other two images were videos where the faces were accompanied by a head movement, which were always displayed aligned with the action-words. The head movements could be horizontal or vertical, mimicking those that we used to say “yes” and “no” in our culture. The continuous stream contained 24 repetitions of each stationary and each moving image, presented always alternating to each other. Given that we restricted the consecutive presentation of the same women to a maximum of two, the TP between any image was equal to 0.5 across the sequence. Each image was 1368 ms long with a duration of 2.188 min for the full sequence. All visual stimuli in this study were presented in color.

Audiovisual stream. We created the audiovisual stream by adding the sound and video tracks, synchronizing the onset of the object-words with the onset of the stationary images and the onset of the action-words to the onset of the moving images (Figure 3.1, and supplementary audiovisual stream).

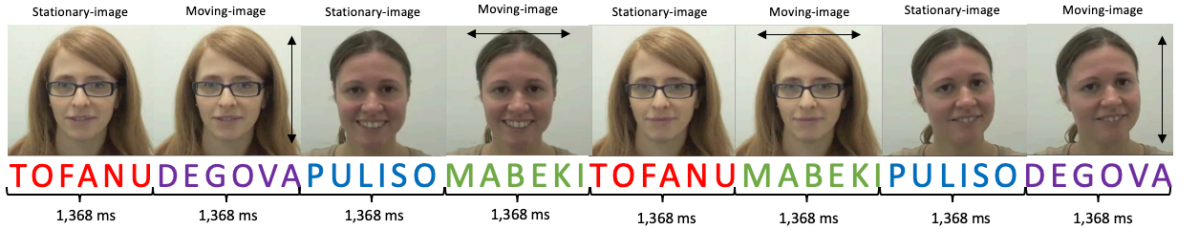


Fig. 3.1. Continuous audiovisual stream used during familiarization phase. Four novel words were concatenated into a continuous stream. Object-words co-occurred with stationary images and action-words with moving images.

3.2.3.2. Test phase. For the test phase, we presented each isolated word associated with pairs of images.

3.2.4. Procedure

The experiment was conducted in a soundproof and dimly lit room. During the evaluation, infants were seated on their caregiver's lap in front of the eye-tracker monitor situated approximately at 60 cm of distance. Parents were visually masked and instructed to do not intervene with the infant behavior during the evaluation.

We first calibrated infant's binocular gaze using fixations longer than 100 ms over one centered and four corner points of the monitor. After gaze calibration, we displayed an audiovisual attention-getter, and then started the familiarization phase by presenting the continuous audiovisual stream. The video was projected over a centered 550 x 580 pixels area of the screen, with black background, subtending a 11.6° x 19.2° of the visual field. The sound was delivered by loudspeakers at ~60 dB.

Once the familiarization phase ended, infants were exposed to 16 test trials resulting from the combination of two different exemplars (image 1 & image 2), by two different women (women 1 & women 2) by two types of words (object-word & action-word) by two sides (left & right). Each test trial begun with an audiovisual attractor to bring the infant's attention to the center of the screen. Then the infants simultaneously heard a word, repeated twice separated by 1000 ms, and saw a visual array composed by two images, one by the left and the other by the right side of the screen. In the trials

evaluating the object-word learning, the words were object-words and the visual array contained two stationary images, one from each woman (Figure 3.2a), while in the trials testing the action-word learning, the words were action-words and the visual stimuli had the video of the same woman moving her head horizontally by one side and vertically by the other (Figure 3.2b). Each image of the visual array was projected over a 380 x 350 pixels area, subtending 11° x 12.7° of the visual field, one the left and the other by the side, equidistant from the center of the screen, with a black background.

The test trials were presented in a pseudo-random order, restricting the repetition of the same word to a maximum of two.

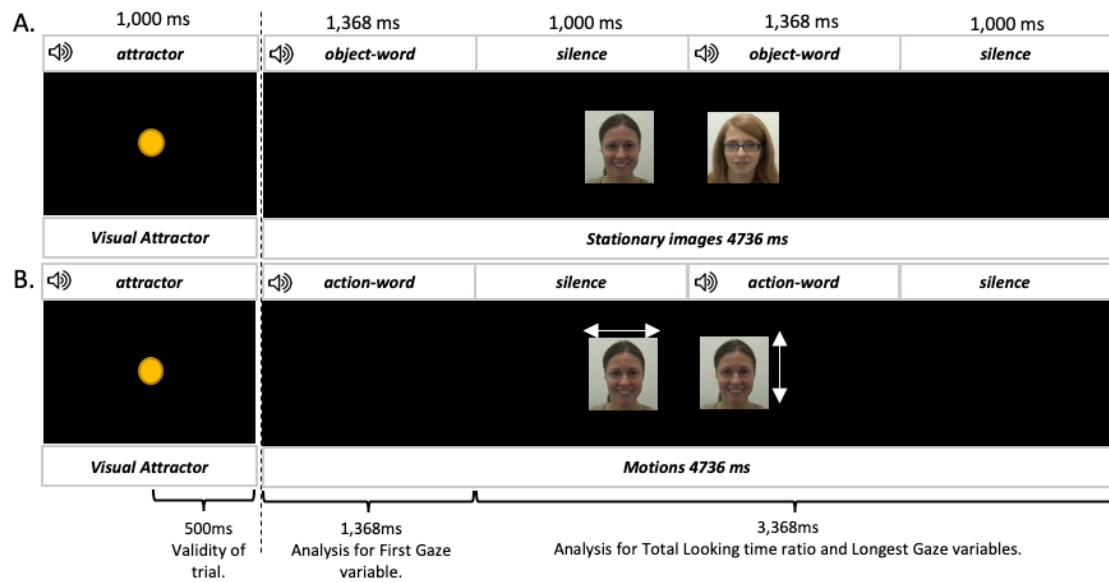


Fig. 3.2. Trial structure for the test phase. We illustrate the timing of the events and the time window for the analysis in the trials evaluating the object-words (A) and action-words (B).

3.2.5. Data acquisition and analysis

3.2.5.1. Spatiotemporal regions of interest (ROI)

For the familiarization phase, we measured the visual behavior over the full centered region of the screen, where the video was displayed, and also over the eyes and mouth regions.

For the test phase, we divided the full screen region in three with equal size (i.e., left, middle and right), and we measured the visual behavior over each one of those regions, from 500 ms before to 4736 ms after the test stimuli onset.

3.2.5.2. Data preprocessing

For the test phase, we first identified the valid trials, defined as the trials when the infants gaze remained on the central attractor the 500 ms before the stimulus onset, and remained over the screen at least 100 ms. We also excluded trials with no eye-tracker data, even if infants looked at the fixation point before the trial started.

3.2.5.3. Visual variables

In both phases, we measured in each infant the total looking time (TLT), defined as the total time that the infant's gaze fell over the ROIs. In the test phase, we also evaluated the visual behavior over each lateralized image by measuring the longest gaze (LG), defined as the side where the longest fixation settled, and the first-gaze direction (FG), defined as the side where the first gaze arrived after the stimulus onset.

In each infant, the TLT for correct responses in each test trial was first computed as a ratio by dividing the TTL over the correct side (left or right) by the sum of the TLT over the correct and incorrect sides, to then be averaged across trials. This procedure allowed us to normalize the visual data across infants, who may have a tendency to make longer or shorter gazes in any circumstance.

The FG and LG for each trial were classified as correct every time that they fell over the correct target. The LG was transformed to a binary value (1 for correct and 0 for incorrect) and then averaged across all trials for each infant. Similarly, the FG in each trial was a binary value that was averaged across all the trials in each infant.

3.2.5.4. Data analysis

In the familiarization phase, we compared the infant gaze toward the stationary versus the moving images, to eyes vs mouth in each of type of images by submitting the mean TTL ratio of each infant over the corresponding regions and events to a paired *t*-test (2 tailed, $\alpha = 0.05$). We also explored for eventual preferences to look longer during the presentation of specific words by submitting the mean TTL ratio of each infants to a univariate ANOVA, with word (*puliso*, *tafonu*, *degova* & *mabeki*) as intra-subject factor. During the test phase, we evaluated the object and action words learning, by submitting the TLT, LG and FG averages of each infant in trials evaluating object and action-word learning, to a series of one sample *t*-test (2 tailed, $\alpha = 0.05$) comparisons, against the chance level at 0.5.

3.3. RESULTS

3.3.1. Familiarization phase

We found that the mean TLT over the familiarization ROI was $1.567 \pm .460$ min, which correspond to the 71.532% of the total duration of the audiovisual stimulation, confirming that infant's attention was considerably directed over the audiovisual stimuli during the familiarization phase. We did not find significant differences between the mean TLT over the stationary and moving images ($p = .574$) across the familiarization, and the mean TLT was equally larger over the eyes ($p = .653$) than the directed over the mouth ($p = .554$) in both, object and action-words. Together the familiarization results indicated that the infants were highly engaged to explore the audiovisual stimulation, having the opportunity to learn both type of word-image associations from it.

3.3.2. Test phase. From a maximum of 8 trials, infants contributed in average with 5.737 (Range = 4 to 8, SD = 1.521) and 6.211 (Range = 4 to 8; SD = 1.584) valid trials for the evaluation of object and action-word learning, respectively.

3.3.2.1. *Object-and action-word learning*

When we compared the mean TTL, LG and FG for object and action-word across all infants against a performance at chance (i.e. 0.5), we did not find significant differences in any comparison. Indeed, the mean \pm the standard deviation of the mean TTL, LG and FG were equal to $.492 \pm .114$, $.505 \pm .177$ and $.479 \pm .182$ for object-word, and $.491 \pm .134$, $.448 \pm .171$ and $.473 \pm .129$ for action-word, respectively. These results showed that as a group, the infants did not learn both types of image-words associations.

We explored then if each infant did learn one type of word only. We first coded with 1 each infant who had a mean TTL, LG or FG greater than 0.5, in any type of image-word learning.

We found that the mean TTL for object or action-word learning was greater than 0.5 in 15 from 19 infants. Indeed, 4 infants (21.053 %) succeeded to learn object-words only, 6 infants (31.579 %) learned action-words only, and 5 infants (26.315 %) achieved both, object and action-words. Only 4 infants (21.053%) did not learn any type of word (see Figure 3.3).

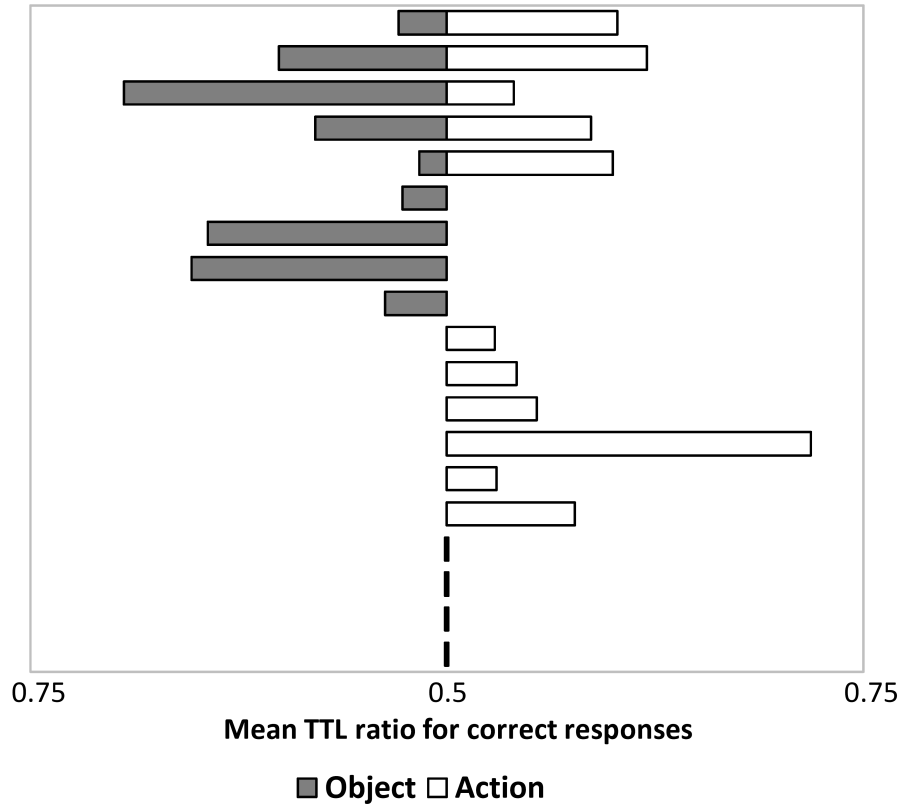


Fig. 3.3. In each infant, we plot the mean TTL ratio greater than chance (0.5) for object-word (left side) and action-word (right panel). We draw a black line to illustrate the 4 infants with a mean TTL ratio lower than 0.5.

Although, the mean LG ratio was larger than chance in 12 from the 19 infants, that difference was not significant ($p = .251$), and the FG over the correct and the incorrect sides did not differ each other ($p = .819$).

Together, the data from the test phase showed that, evaluated with TTL ratio (which is one the most robust variable to evaluate the infant visual behavior; Csibra, Hernik, Mascaró, Tatone & Lengyel, 2016), the 78.947 % of the infants did learn at least one type of word. This proportion was significantly greater than that expected by a performance at chance (Chi-square goodness of fit, $\chi^2_{(1)} = 6.12$, $p = .0134$). We speculated that the failure to succeed in both type of image-word associations in all infants could reveal the interindividual variability that infants exhibited to manage the cognitive requirements imposed by our task.

3.3.2.2. Correlations between the data from the familiarization and test phases

To evaluate whether the infants who looked longer during familiarization got greater learning during the test, we submitted the mean TTL and LG ratio for the correct responses in object or action word learning to a bivariate correlation analysis against the mean TTL directed to the visual target during familiarization. We found that the mean LG positively correlated with the mean TTL dedicated to look the area where the visual track of the audiovisual stimuli was projected during familiarization (Pearson's $r = .669$; $p = .002$; Figure 3.4), which agrees with the idea that the time spent looking toward the target during the familiarization, improved the opportunities to learn.

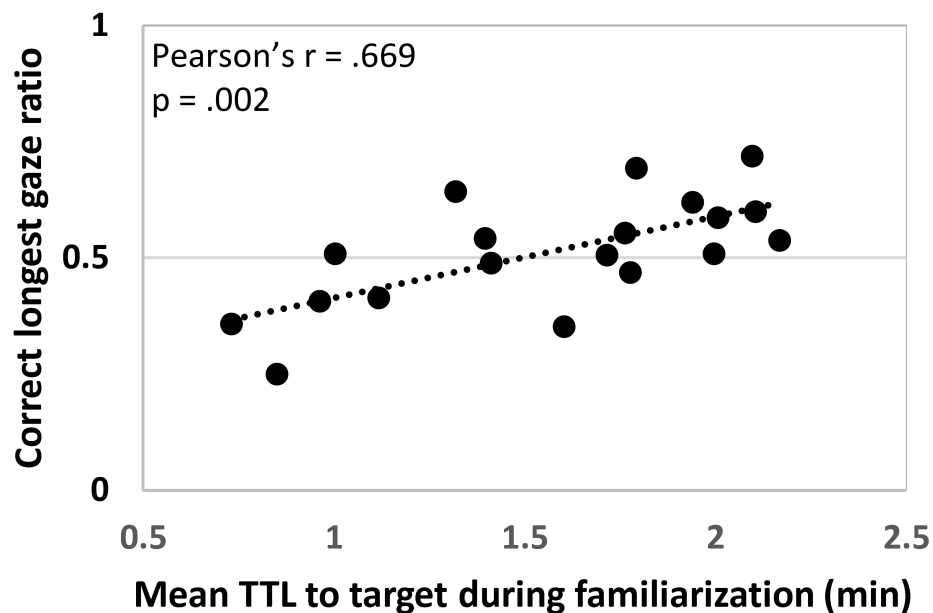


Fig. 3.4. The time dedicated to exploring the target during the familiarization increase the probability to learn the image-word associations.

3.3.2.3. *No preference for particular words.*

Finally, we confirmed that our results could not be explained because infants learned certain particular words. Indeed, the 63% of the infants learned more than one action-word and the 68% learned more than one object-word. We confirmed the absence of such bias by submitting the mean TTL of each infant for the correct responses for each one of the four words to a univariate ANOVA. We did not find significant difference for any word ($p = .152$) or any type of word ($p = .239$).

3.4. DISCUSSION

3.4.1. *Five-months-old infants can learn object and action word associations from audiovisual streams.*

To our knowledge our study is the first to show infants as younger as 5-months of age can segment fluent audiovisual speech streams into chunks that map novel words to stationary and/or moving images. From birth, human infants are familiar with the concurrent stimulation provided by faces and speech (e.g. Kopp, 2014), but it is notable that they can extract such associations from a continuous stream, and use it to later recognize image-word pairs. The repeated exposure to a precise synchrony between speech and face stimulation, may have provide redundant information about that both type of stimuli referred to a single referent. Indeed, in adults is proposed that the multisensory stimulation provides perceivers with redundant information that increases the probability to learn (e.g. Bahrick, Lickliter, & Flom, 2004).

3.4.2. *Infants have high interest to explore audiovisual streams done with speech and faces.*

We also showed that 5-months-old infants were highly attracted by the audiovisual stream composed by speech and faces. By using faces as visual stimuli we wanted to better

capture the infant attention on the visual stimuli, given that speech and human faces are highly preferred by young infants over other type of visual stimuli (Gliga, Elsabbagh, Andravizou & Johnson, 2009; Valenza, Simion, Macchi Cassia & Umiltà, 1996). Such types of stimuli may have facilitated to uncover the infant capacity to discover audiovisual association from continuous streams, which could remain hidden in other experiments using other type of audiovisual stimuli. Indeed, the unique study we know evaluating word-mapping from continuous audiovisual streams in 8-months-old infants, showed difficulties to map words with images when audiovisual streams are composed by four trisyllabic words co-occurring with one geometric shape each (i.e. cross, diamond, heart and hexagon) (Thiessen, 2010). In our study, speech and faces may have played a crucial role for audiovisual integration, while they promoted the infant engagement with the task, somehow supporting the “Human First Hypothesis”, which states that infants possess information about their conspecifics and use it to facilitate their identification and to learn from them (Bonatti, Frot, Zangl & Mehler; 2002).

3.4.3. *Five-months-old infants not always learn both type of words*

The fact that 5-months-old infants had difficulties to learn both type of words, could be because some of our infants had less cognitive resources to deal with our somehow complex task. Indeed, the interindividual variability is the norm for cognitive abilities during development (De Ribaupierre, 2015). For instance, although in average the neonates can memorize novel bisyllabic spoken words after 2-minutes of familiarization, they show a large interindividual variability (Benavides–Varela, Hochmann, Macagno, Nespor & Mehler, 2012). The interindividual variability may have a number of causes that we cannot disentangle here, but may involve factors such as a transient decrease in the attentional engagement with the task to a delay in neural maturation. Indeed, in adults the statistical learning is modulated by attention (e.g. Toro, Sinnett & Soto-Faraco, 2005; Turk-Browne, Jungé & Scholl, 2005), and the activity of

the neural networks recruited for audiovisual associations highly develop during the first 6 months year of age (e.g. Hyde, Jones, Flom & Porter, 2011).

Confronted to our multitask protocol some infants could concentrate their cognitive efforts to learn one of the two types of words, suggesting also possible interference and competition in the processing of the two concurrent type of information, similar to that recently reported for adults (Chen, Gershkoff-Stowe, Wu, Cheung & Yu, 2017).

3.4.4. *Five-months-old equally learn object and action words*

Our results did not show any bias to learn one type of word over the other, suggesting that infant perception may have randomly coupled with one type of audiovisual association, defining thus the unit for further analysis. Indeed, behavioral and neural coupling with the external stimulation has been reported during early infancy in domains such as music (Cirelli, Spinelli, Nozaradan & Trainor, 2016) and speech (Kabdebon, Pena, Buiatti & Dehaene-Lambertz, 2015). However, the mechanisms explaining the individual bias to process stationary or dynamic images need further investigation.

3.5. CONCLUSION

As far as we know, this is the first study showing that 5-month-old infants succeed to learn object and/or action-word from continuous audiovisual streams, which could play a crucial role in word-mapping learning, developed at older ages. We believe that our results might contribute to update the current models about word learning during early infancy, and would provide the scientific community with new methods to evaluate it.

3.6. Author Contributions

M.P. developed the study concept and design. M.P. y C.J. performed the testing and data collection. C.J., M.P. and C. M. performed the data analysis. C.J. and M.P. interpreted the data and drafted the manuscript. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank to Latin American School for Education, Cognitive and Neural Science on their support in this study. We are grateful with Marina Nespor, Jacques Mehler and Alan Langus for their comments about the study design, with Enrica Pittaluga, for their help recruiting participants, and with Alissa Ferry and Marijana Sjekloca for their assistance in the creation of the stimuli. Our gratitude also goes to parents and the infants for their valuable interest in participating in this study.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was supported by CONICYT PhD Grant #21140640 to C.J., FONDECYT # 1110928 to M.P. and BNI Beca Puente ICM P09-015-F to C.M.

IV. STUDY 2

Vowels help 8-month-old infants to learn object and action-words from audiovisual continuous streams.

ABSTRACT

The current study explored whether and how 8-month-old infants may discover novel words that were continuously associated to static or moving images in an audiovisual stream. Specifically, in a series of four experiments, we tested young infants on their ability to learn novel words that systematically co-occurred with a stationary face (for object-words) and a novel word that systematically co-occurred with moving faces (for action-words), and the phonological cues (vowel or consonant harmony) on what infants at this age rely to succeed on this task.

Our study showed that infants were able to learn both object and action-words but when the visual referent were social faces and, only when the speech component had harmony vowel, and not consonant harmony, as a facilitator for the segmentation task. We believe that our results will open a series of questions on this topic, providing relevant evidence to the scientific community about the visual and speech cues on which infants rely during word-mapping process, as a contribution to future word learning studies.

4.1. INTRODUCTION

By the end of their first year, infants learn, apparently with no big effort, that the words they heard are not only associated with specific objects or events, but they point to a common referent, process known as word-mapping. This capacity develops even under natural environments, where the words and referents are not presented isolated, but they appear concurrently as multisensory information (e.g. Gibson, 1966), and embedded on continuous contexts, without clear edges between elements. To succeed in word-mapping, infants should be able to segment the speech signal and visual scenes into audio and visual chunks, to encode the concurrent audiovisual information into common audiovisual associations, and to save such associations in memory for further recognition. Two of such mechanisms that infants handle before their first year, are the analysis of the distributional information that the audio and visual stimuli convey, and the ability to integrate multisensorial information into common representations.

Indeed, the analysis of the distributional information of external stimuli is implemented in the infant mind. Previous studies have demonstrated that 8-months-old infants do segment artificial speech streams at the syllable boundaries where the transitional probability (TP) between adjacent syllables drops, grouping the syllables with high TP into chunks that infants would conceive as potential words (Saffran, Aslin & Newport, 1996). Additionally, infants can also exploit distributional cues to segment visual sequences. After been exposed to a sequence of visual images presented consecutively, 2-days-old neonates (Bulf, Johnson & Valenza, 2011) and 2,5 and 8-month-old infants succeed to segment visual sequences into 2-images (Kirkham, Slemmer & Johnson, 2002) or 3-images chunks (Roseberry, Richie, Hirsh-Pasek, Golinkoff & Shipley, 2011), on the edges where the TP between adjacent images drops.

Moreover, soon after birth neonates can learn novel audiovisual associations, probably exploiting their abilities to detect and process the synchronic properties in space and time of the cross-modal stimuli (e.g. Lewkowicz, 2000). For instance, 2 to 4-month-old infants have remarkable capacities to match faces and lips information with speech

(Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003). Early audiovisual associations would be thus relevant for further word-mapping.

Data supporting this view has been provided by a recent brain study (Friedrich & Friederici, 2017), showing that, after been exposed to the simultaneous presentations of isolated pairs of novel words and novel objects, 3-months-old infants displayed an increase in the amplitude of the late negativity component, during object-word repetitions, an electrophysiological component sensitive to word comprehension in toddlers (Conboy & Mills, 2006).

Nevertheless, to our knowledge, to date, only one study has explored whether infants learn object-word associations from artificial, nonsense monotonous continuous audiovisual streams (Thiessen, 2010). This study showed that 8-months-old infant have difficulties to discover audiovisual associations from such type of audiovisual streams, when visual stimuli are static geometric shapes, indicating that at that age infants would rely on auditory more than on visual stimuli to find words from this type of stimuli.

Our study wanted to contribute to this area of the research by exploring whether and how 8-months-old-infants discovered words associated to objects and to actions from artificial continuous audiovisual streams, when images were faces and words contained phonological cues. A number of studies have demonstrated that faces are highly interesting stimuli to young infants (Gliga, Elsabbagh, Andravizou & Johnson, 2009; Valenza, Simion, Macchi Cassia & Umiltà, 1996) and that the attention to faces, either static or dynamic, rapidly increased during the second half of the first year (Frank, Vul, & Johnson, 2008). Furthermore, we used vowel repetition as phonological cue, which is an extreme version of vowel harmony (Rose & Walker, 2011). Numerous studies have shown that vowel harmony helps infants to learn novel words. For instance, after been familiarized with bi-syllables containing the same vowels, newborns (Benavides-Varela, Hochmann, Macagno, Nespor & Mehler, 2011) and 2-months of age (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy & Mehler, 1988) recognize the changes in the vowel tiers, but not the changes in the consonant one; and a recent study reported that the vowel harmony helps 8-months-old to discover trisyllabic words from artificial fluent speech (Mintz,

Walker, Welday & Kidd, 2018). With an exploratory view, we also evaluated consonant repetition, given that previous studies have shown that infants do exploit the analysis of the consonantal tier for word learning at older age (Hochmann, Benvides–Varela, Nespor & Mehler, 2011)

We expected thus that by combining phonological cues and faces we would provide infants with more opportunities to learn from an artificial, no sense, monotonous continuous audiovisual environment.

In the current study we ran 4 experiments, each one testing a different group of 8-months-old infants familiarized with one of 4 possible versions of a 2.10-min long audiovisual stream. The streams were always composed by the concurrent stimulation with 4 trisyllabic novel words with 4 faces images from two different women (total 8 images). The speech component of the audiovisual stream was artificial, built by concatenating 4 different trisyllabic non-sense words, played monotonously, and having a transitional probability (TP) equal to 1 within words and .5 between words. The visual component was constructed by concatenating eight visual events with a TP equal to .5 between any of them: 4 images were static (2 per woman) and included head and shoulders, and 4 images were videos (2 per woman) and showed the same faces but now making a horizontal or vertical head movement, similar to those used for negation and affirmation in many cultures. In the continuous stream the presentation of the words was synchronized with the presentation of either the static faces (thereafter object-words) or with the moving ones (thereafter action-words). The statistical properties of our audiovisual stimuli provided infants with a continuous environment from where they can extract words because their phonological cues and TPs, but they can also match those words with a visual referent upon the basis of their co-occurrence and covariance, where each object-word always co-occurred with a specific woman, and each action-word always appeared associated with a specific head motion, made by any of the two women (Figure 4.3). In each familiarization stream we counterbalanced the words, faces and motion across infants. During the familiarization, we measured the total looking time (TLT) over the full stream presentation, to estimate the engagement of the infants with the

task. However, to identify particular patterns of visual behaviors that the infants could develop during familiarization, we also separately measured and compared the TLT over the static versus the moving images and by the eyes versus the mouth area by applying two paired samples *t*-test, and for each one of the 4 words, by applying a multivariate ANOVA (see methods).

In all experiments, we evaluated the learning of the object and action-words by using remote eye tracking. After the familiarization with the continuous stream, we expected that infants who succeeded in object and/or action-word learning would look longer and/or faster to the correct image, when they heard one word and simultaneously saw two stationary or moving images. In each trial, the words were presented twice, separated by 1 second of silence, while images remained displayed on the screen from the onset of the first presentation of the word until 1 second after the offset of the second presentation of the word. We specifically estimated the infant's accuracy over such time window, by measuring the following variables: a) total looking time ratio (thereafter TLTr) computed as the time spent over the correct image divided by the time sighted over the correct and incorrect images; b) the accuracy of the total looking time (thereafter TLTa), the longest gaze (thereafter LG) and the first gaze (thereafter FG) computed as the mean of the times that the TLTa, LG and FG fell over the correct images. Those variables were computed using standard procedure for infant's eye behavior analysis (Csibra, Hernik, Mascaró, Tatone & Lengyel, 2016) and are sensitive to the infant visual preference in a two alternatives choice paradigm. For statistical analysis, each one of such variables were submitted to one sample *t*-test (2 tailed, $\alpha = .05$) against chance at level 0.5. We also submitted those data to a series of separated two paired samples *t*-test to evaluate eventual preferences for particular words.

In Experiment 1 we used vowel harmony in action-words and we only tested the visual preference for this type of word; in Experiment 2 we used the same stimuli of the Experiment 1, but this time the object-words had vowel harmony, and we tested the preference for this type of word only. The Experiment 3 and 4 were identical to Experiments 1 and 2, except that they used words with consonant harmony instead of the

vowel one. Finally, we compared the results across the 4 experiments to evaluate the contribution of each word learning under our experimental paradigm.

4.2. RESULTS

The results for the 4 experiments are illustrated in Figure 4.1.

4.2.1. *Experiment 1. Action-word learning when words conveyed vowel harmony.*

4.2.1.1. Familiarization. We computed the analysis over twenty 8-months-old infants ($M = 7.96$; $SD = .453$). We observed that the mean TLT over the central region over which the continuous stream was displayed, was $1.467 \pm .361$ which correspond to the 67.04% of the full stream duration, confirming that, in average, the infants were interested to explore the visual component of the stream. We did not find significant differences between the mean TLT directed over the static versus the moving faces ($p = .311$). In both images, the mean TLT was significantly greater for eyes versus mouth across familiarization for action-words ($t(19) = 8.899$, $p < .001$, Cohen's $D = 3.011$) and object-words ($t(19) = 9.022$, $p < .001$, Cohen's $D = 2.928$), and those preference for eyes was equally during the presentation for moving and static faces ($p = .061$) and also for mouth ($p = .937$). All the results together indicate that during familiarization phase infants looked to the static images versus the moving faces, that co-occurred with object and action-words respectively, in a similar probability to learn both word-images associations.

4.2.1.2. Test. From a maximum of 16 trials (8 per each action-word), infants contributed in average with 9.3 valid trials (see methods; range 5 to 16; $SD = 3.715$). We evaluated the learning of action-word only, those conveying the vowel harmony cue and associated to moving faces.

As we predicted, 7-8-months-old infants had a significantly greater than chance mean number TLTa ($M = .593$, $SD = .171$, $t_{(19)} = 2.437$, $p = .025$, Cohen's $D = .77$) and mean

number of LG ($M = .585$, $SD = .169$, $t_{(19)} = 2.261$, $p = .036$, Cohen's $D = .711$) over the correct image.

Consistent with those results we also found a strong tendency for mean TLTr, corresponded to the proportion of the TLT directed to the correct image over the TLT directed toward both images, that was greater than chance ($M = .539$, $SD = .093$, $t_{(19)} = 1.9$, $p = .073$, Cohen's $D = .593$).

In contrast, the mean FG over correct response was not significantly different from the chance level of .5 ($p = .648$). These results can be explained because the 69% of infants' first gaze were directed over the right side of the screen, indicating that in most of the valid trials, infants started their visual exploration from this side of the visual field.

Across infants, we did not find preferences for particular words. Indeed, 18 infants learned one or both action-words when associated to either horizontal or vertical motion.

Together the results of Experiment 1 showed that infants can learn and memorize the action-words conveying vowel harmony associated with a specific motion.

Importantly, given that infants generalized the action-words to both women we can sustain that they discovered the label were associated to the motion and not to a particular woman. Further studies are however necessary to explore whether infants can generalize this knowledge to novel agents.

4.2.2. Experiment 2. Object-word learning when words conveyed vowel harmony.

4.2.2.1. Familiarization. The final analysis involved twenty 8-month-old infants ($M = 8.01$; $SD = .241$). We found that the mean TLT over the central region was $1.49 \pm .529$ which correspond to the 68.09% of the of the audiovisual stream. We did not find significant differences between the mean TLT for static versus moving faces ($p = .473$), and also for eye region and mouth region for static ($p = .824$) and moving faces ($p = .726$). We also found that infants looked longer to the eye region compared to the mouth region for object ($t(19) = 7.102$, $p < .001$, Cohen's $D = 2.409$) and action-words ($t(19) = 7.558$, $p < .001$, Cohen's $D = 1.759$). Similar to Experiment 1, infants explored the stream a considerable amount of time, having the opportunity to learn word-images associations.

4.2.2.2. Test. Infants contributed in average with 9.4 valid trials (range = 5 to 15; SD = 2.88). We evaluated the learning of object-word only, those conveying the vowel harmony cue and associated to stationary faces.

We found that the mean LG for correct response was significantly greater than chance ($M = .573$, $SD = .140$, $t_{(19)} = 2.350$, $p = .030$, Cohen's $D = .737$), and the mean TLTa toward the correct image tended to be significantly higher than chance ($M = .563$, $SD = .155$, $t_{(19)} = 1.825$, $p = .084$, Cohen's $D = .574$), with no significant difference that chance for the mean TLTr ($p = .382$) and FG ($p = .734$).

Again, the mean FG was not significantly different from chance ($p = .734$) because in the 60% of the valid trials the FG went to the right side of the visual field.

Across infants, we did not find preferences for particular words. Indeed, 16 infants learned one or both action-words when associated to either horizontal or vertical motion.

In sum, the results of the Experiment 2 showed that infants were able to learn object-words to identify the label associated to particular women, although with weaker statistical power than that observed for the Experiment 1, probably because during the test the static images were less interesting for infants. Indeed, previous studies have shown that when presented isolated, the moving faces better capture the infant attention than the static ones (Otsuka, Konishi, Kanazawa, Yamaguchi, Abdi & O'Toole, 2009). Moreover, if we compared the results with those from the Experiment 1, we may propose that the dynamic nature of the moving stimuli may have facilitated the analysis of the synchrony between moving images and ongoing auditory stimuli. Further studies are necessary to test this possibility.

4.2.3. Experiment 3. Action-word learning when words conveyed consonant harmony.

4.2.3.1. Familiarization. The final analysis was computed over twenty 8-months-old infants ($M = 8$; $SD = .258$).

We found that the mean TLT over the central region for the presentation of the audiovisual stream was $1.64 \pm .316$, which corresponded to the 75.03% of the familiarization. We did

not find significant differences between the mean TLT for stationary faces versus moving faces ($p=.882$), or between stationary and moving faces for eyes region ($p = .538$) and mouth region ($p=.225$). We also found that infants looked longer to the eye region compared to the mouth region for object ($t(19) = 12.170, p<.001$, Cohen's $D = 4.178$) and action-words ($t(19) = 12.545, p<.001$, Cohen's $D = 4.260$). Similar to Exp. 1 and 2, here the infants had the opportunity to learn the word-images associations.

4.2.3.2. Test. Infants contributed in average with 10.3 valid trials (range = 5 to 14; SD = 3.431).

We did not find significant differences from chance in any visual variable, i.e. TLTa and mean TLTr ($p = .615$ and $p = .384$, respectively), mean LG ($p = .551$) and mean FG ($p = .588$).

Although the infants did not succeed in the task, they also showed a bias to direct their first gaze to the right visual field in the a 61% of the valid trials.

In global the results indicate that infants could not exploit consonant harmony to discover action-word from the continuous audiovisual stream.

4.2.4. Experiment 4. Object-word learning when words conveyed consonant harmony.

4.2.4.1. Familiarization. We computed the final analysis over seventeen 8-month-old infants ($M = 7.89$; $SD = .185$).

We found that the mean TLT over the central projection area was $1.65 \pm .338$, which corresponded to the 75.54% of the total familiarization. We found that the mean TLT toward the moving faces was longer than those directed to the static ones ($M = 50.3s$ and $M = 48.9s$) with a significant difference between this two times ($t_{(16)} = 3.292$; $p = .005$; Cohen's $D = 0.138$). Infants also looked longer to the eyes region for moving faces over stationary ones ($t_{(16)} = 3.553, p = .003$; Cohen's $D = 0.116$), but looked equally for mouth region ($p=.148$). Again, we also found that infants looked longer to the eye region compared to the mouth region for action ($t(19) = 7.603, p < .001$, Cohen's $D = 3.044$) and object-words ($t(19) = 7.746, p < .001$, Cohen's $D = 2.994$).

4.2.4.2. Test. In average infants contributed with 9 valid trials (range = 5 to 16; SD = 3.968).

In contrast to previous experiments, we found that infants preferred the incorrect response. Indeed, the mean TLTr ($M = .466$, $SD = .052$; $t(16) = -2.636$, $p = .018$, Cohen's $D = .924$), and mean FG ($M = .385$, $SD = .145$; $t(16) = -3.25$, $p = .005$, Cohen's $D = 1.121$) were below the chance level. We did not find significant results for TLTa ($p = .436$) and LG ($p = .058$). Moreover, we found that the mean FG equally fell over then right and left sides of the visual field.

Together these results showed that although infants were engaged in the familiarization, during the test seem to avoid the correct response. Further studies are necessary to explain that behavior, however, given that infants sighted longer to moving faces during the familiarization, they may have allocated their cognitive resources in the learning of the action words, items that we did not tested. As in Exp. 3, the consonant repetitions did not seem to have helped infants to discover image-word associations

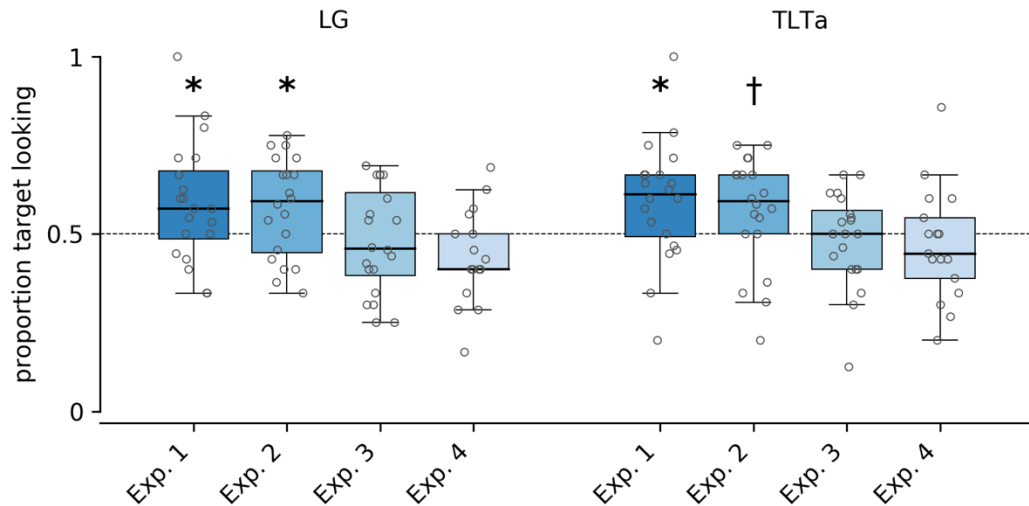


Fig. 4.1. Infants succeeded to find object-words (Exp.1) and action-words (Exp.2) when words had harmony vowel. We plot the mean proportion for correct responses for longest gaze (LG) and mean number total looking time (TLTa), compared with a proportion expected at chance (0.5). Circles represents each infant for each experiment. (* = $<.05$; † = $<.1$).

4.2.5. Analysis between experiments

4.2.5.1. Evaluating sociodemographic differences across experiments.

To quantify the effects of vowel versus consonant repetitions we demonstrated that the infants from each experiment did not differ in their sociodemographic data by submitting to a one-way multivariate analysis of variance (MANOVA) with Experiment (Exp. 1, Exp. 2, Exp. 3 and Exp. 4) as a between-subjects factor the variables age of evaluation ($F_{(3,73)} = .63, p = .598$), gestational age at birth ($F_{(3,73)} = .47, p = .704$), APGAR at 5 min at birth ($F_{(3,71)} = .284, p = .837$), months of breastfeeding ($F_{(3,72)} = .941, p = .426$), and mothers' level of education (Pearson's Chi-square, $\chi^2=14.736, p = .471$).

4.2.5.2. Familiarization across experiments.

To quantify the TLT dedicated to exploring the audiovisual sequence during the familiarization across the experiments, we submitted to the mean TLT over the central region to a one-way ANOVA with Experiment (Exp. 1, Exp. 2, Exp. 3 and Exp. 4) as between-subject factor. We did not find a significant effect of Experiment ($p = .328$) indicating that in each experiment the infants had similar engagement and opportunities to learn from the audiovisual stimuli.

4.2.5.3. Learning during test across experiments.

To quantify the differences in the audiovisual sequence during the test across the experiments we submitted the TLTr, and the mean of the TLTa, LG, and FG to a one-way multivariate analysis of variance (MANOVA) with Experiment (Exp.1, Exp.2, Exp. 3 and Exp. 4) as inter-subject factor. We found a main effect of Experiment on TLTr ($F_{(3,73)} = 3.274, p = .026$; $\eta^2p = .119$) and LG ($F_{(3,73)} = 4.501, p = .006$; $\eta^2p = .156$). Post hoc comparison showed that the mean LG was significantly greater in the Experiment 1 and Experiment 2 than in the Experiment 4 ($p = .018$ and $p = .036$ Bonferroni corrected

respectively). Moreover, the TLTr was significantly greater during the Experiment 1 than in Experiment 4 ($p = .036$ Bonferroni corrected).

Together, the results showed that the greater performance was observed in both experiments providing vowel harmony, either they evaluated object or action-words over the experiment providing consonant harmony to evaluate the learning of object-words. These data support the proposal that before the first year of age, infants exploit vowels over consonants to learn words (Hochmann et al., 2011, 2017).

4.2.5.4. Comparing the learning in experiments with vowels and consonant harmony.

To quantify the learning between the experiments providing vowel versus consonant harmony as phonologic clues, we submitted the TLTr, TLTa, LG and FG of each infants evaluated in the experiments 1 and 2 (thereafter vowel experiments) versus those obtained in the infants evaluated in the experiments 3 and 4 (thereafter consonant experiments) to a one-way multivariate ANOVA with Phoneme (vowel & consonant) as between-subject factor. We found that as compared to consonant experiments, the vowel experiments had greater TLTr ($F_{(1, 75)} = 8.369, p = .005, n2p = .100$), TLTa ($F_{(1, 75)} = 8.262, p = .005, n2p = .099$) and LG ($F_{(1, 75)} = 12.798, p = .001, n2p = .146$). Together then results suggested that vowel harmony was a phonological cue that facilitate word-image associations, independently that the words referred to static or moving images.

4.2.5.5. Age as a covariable

Previous studies have shown that from birth infants are sensitive to vowels to learn words presented isolated, while consonants would be exploited with the same goal by the end of their first year (Hochmann et al., 2011, 2017). We thus looked for an eventual effect of age in the image-word learning in both, vowel or consonant experiments. We submitted to a Pearson's bivariate correlation the TLTr, TLTa and LG against the infants' age of evaluation, which ranged from 7 months 15 days to 8 months 25 days of age across experiments. Interestingly, we found that the TLTa positively correlated with the age of

evaluation of the infants in the consonant experiments (Pearson's correlation $r = .339$, $p = .040$) but not in the vowels ones ($p = .901$) (Fig. 4.2), with a similar tendency in the TLTr (Pearson's correlation $r = .297$, $p = .074$) in the consonant experiments, but not in the vowels ones (Pearson's correlation $r = -.105$, $p = .518$).

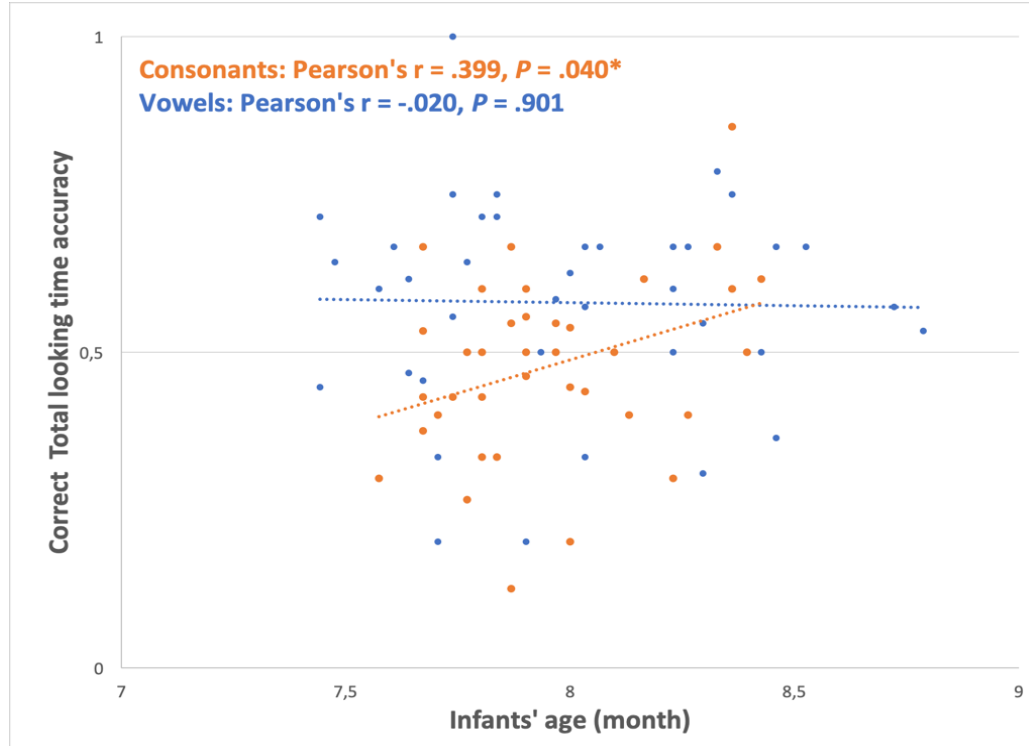


Fig. 4.2. The figure illustrates bivariate correlations between TLTa in vowel (in blue) and consonant experiments (in orange) against the age of evaluation.

Moreover, when we submitted the TLTa and TLTr to a one-way multivariate ANCOVA, with Phoneme (Vowel & Consonant) as between-subject factor, and age in months as covariable, we found that for both variables the performance covariate as a function of the age of evaluation in the consonant experiments only ($F_{(2, 74)} = 4.279$, $p = .017$, $n2p = .104$ and $F_{(2, 74)} = 4.670$, $p = .012$, $n2p = .112$, for TLTa and TLTr respectively).

Together, these results subtly suggest that infants may have not been able to exploit

the information from consonant harmony to find image-word associations, because this ability was at different level of development across the infants. Intra-subject and longitudinal studies are necessary to explore this possibility.

4.3. DISCUSSION

4.3.1. Eight-months-old infants learn object and action-words from continuous audiovisual stimuli.

We showed that 8-months-old infants succeeded to find object and action-words from continuous audiovisual streams when the visual stimuli were faces and words conveyed harmony vowel.

Our results contrasted with previous showing that, contrary to adults, 8-month-old infants do not benefit from the concurrent presentation of images and words to learn novel word-object associations, when words were trisyllables and objects geometric shapes (Thiessen, 2010). The authors attributed this failure to the infants' difficulties to detect the relations between words and images, but not in the segmentation themselves. We may have facilitated such detection by exposing infants to some audiovisual stimuli which provided socially salient visual stimuli, i.e. static and moving faces instead of geometric shapes, and cues to facilitate the segmentation of the continuous streams with visual, i.e. alternating static and moving images, and speech, i.e. vowel harmony, cues.

4.3.2. Faces as facilitator of audiovisual processing

- ***Social cues.*** Given that the faces we used were smiling and directing the gaze to the infants, they provided powerful social cues that may helped infants to capture their attention on the stimulus target and concentrate their cognitive efforts over the audiovisual stimuli processing. A smiling face making eye-contact is a powerful cue to inform infants that the stimulation is directed to them and seems to enhance the learning of the attended stimuli (e.g. Csibra and Gergely, 2009). For instance, eye-contact is necessary for gaze shifts to successfully orient attention in 4- and 6-months-old infants (e.g. Senju and Csibra,

2008), and enhances the predictions about the multimodal events (Wu & Kirkham, 2010).

- ***Static and moving faces alteration.*** The sole alternate between static and moving faces added a physical cue to segment both, the speech and the visual tracks, and may have facilitated the extraction of audiovisual chunks, instead of unintegrated audio and visual elements. Indeed, previous studies have shown that 6-months-old infants associate novel words embedded on short sentences to objects when the objects synchronically moved at each occurrence of the word (Shukla, White and Aslin, 2011).

- ***Familiar head movements.*** The familiarity with the horizontal and vertical head movements may have also contributed to facilitate the learning of the task. A recent study showed that 8-months-old infants learn more the label associated to novel words when the objects make familiar movements such as shaking and looming than when they move with unusual movements such as slow up-down and left-right motions (Matatyaho-Bullaro, Gogate, Mason, Cadavid, Abdel-Mottaleb, 2014). In sum, the faces we used as visual stimuli may have facilitated not only the segmentation of the visual targets and the words but also their integration into a common referent.

4.3.3. Phonological cues: phoneme harmony.

Our speech stimuli here also provided efficient cues to segment the fluent speech streams. Indeed, a recent study reported that vowel harmony improved the discovery of novel words from artificial non-sense monotonous continuous speech stream (Mintz et al., 2018). Vowel and consonant harmony refers to long-distance phonological assimilations, in which phoneme sounds become more like the nearby sounds (Rose and Walker, 2011). The assimilation of phonemes at the local level, i.e. inside a word, helps adults to extract words from continuous surrounding contexts. In the current study, by using vowel harmony, we provided infants with locality restrictions that helped them not only to extract words from the fluent stream but also to succeed in the discovery of types of words, i.e. those associated to static faces and other to the moving ones. Indeed, locality cues benefit

categorization and memory in adults. For instance, adjacent rather than distant objects tend to be identified as members of the same category (Rakison & Yermolayeva, 2010), and memory for parts of sequences usually chunks adjacent elements, such as saving a phone number as a first and last half of the number instead of trying to memorize odd and even numbers, which may not be adjacent (Baddeley, Gathercode & Papagno, 1998)

Why consonant harmony did not help then? A corpus of data supports the idea that before their first birthday, infants possess greater abilities to exploit vowels over consonants to learn words (Hochmann et al., 2011).

For instance, newborn infants memorize newly learned words helped by vowel sounds (Benavides-Varela et al., 2011) and differentiate novel words relying primarily on vowels sounds (Bertoncini et al., 1988). Moreover, by 5-months of age infants rely more on vowels than on consonants during the recognition of their own names (Bouchon, Floccia, Fux, Adda-Decker & Nazzi, 2015).

Consonant would also informative for word learning but it would play a role after 11 months of age (Hochmann et al., 2011; Poltrock & Nazzi, 2015).

Interestingly, our results showed that although at the chance level, the infant's age of evaluation positively correlated with the performances to find image-words when the speech stream conveyed consonant harmony. Further studies with older infants with this task are necessary to clarify this relationship

Other factor that may explain the difficulties to exploit consonant harmony is that, because articulatory constraints, in natural languages, consonant harmony is more likely to find at non-local level than vowel harmony, that is it depend more on non-adjacent than on adjacent relationships across a large number of syllables (Finley, 2011). With this information, we may speculate that the repetitions of consonants inside a word would be naturally less necessary than vowel harmony to inform infants about what is a word. Further studies are necessary to evaluate such conjecture.

4.3.4. Face-speech association as inter-sensorial binding

Since infants and newborns can perceive and discriminate among faces under different conditions (Nelson, 2001, for a review), and are excellent perceivers of human voice and speech since they are in the womb (DeCasper, Lecanuet, Busnel, Granier-Deferre & Maugeais, 1994; DeCasper & Spence, 1986). Moreover, infants have remarkable capacities to match faces and lips information with speech since are newborns (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999, 2003).

As faces and speech occurred together in natural environment, and are presented in a amodal temporal structure, infants developed skills for perceiving and remember face-speech relations since younger ages (Bahrick, Hernandez-Reif & Flom, 2005).

In our study, infants were exposed to an audiovisual sequence, on which faces co-occurred with speech in synchronous and redundant way. The synchrony between audio and visual stimuli may have facilitated the intersensory integration, while previous studies have shown that temporal synchrony is a powerful binding for multisensory perception (Bahrick, Lickliter & Flom, 2015 for a review). Thus, the integration of face and speech in a simultaneous and redundant audiovisual representation may have served as cue for infants to learn both types of words.

4.3.5. Object- versus action-word learning

In our study, we conceived object-words as intrinsically different to action-words, while the former associated with persons, who were labeled with a single word, and the latter associated with a feature generalizable across them, i.e. the motions (e.g. Gogate and Hollich, 2016). From this point of view, the learning of object and action-words somehow resembles the learning of nouns and verbs observed later in life. A number of studies support the idea that nouns are learned earlier and faster than verbs (Waxman, Fu, Arunachalam, Leddon, Geraghty & Song, 2013; Golinkoff & Hirsh-Pasek, 2008; also see Gogate and Hollich, 2016). Our results did not show significant differences to learn any of these types of words. At least two possible explanations can be proposed for that results. First, the learning of the object and action-word associations at this age is not related to

word-mapping observed later, or, second, the object and action-word learning reveal rudimentary forms of such process. According to previous evidence, during early word learning, infants may solve the problem to find the right word-referent pair in a noise context, where the quality of the input (or their referential transparency) and the co-occurrence of stimuli, will increase the possibility to learn new words (Cartmill, Armstrong, Gleitman, Goldin-Meadow, Medina & Trueswell, 2013; Smith, Suanda & Yu, 2014). In our study, both synchronous co-occurrence and immediate word-referent are presented during the audiovisual stream, may have given to infants the necessary cues to associate both types of stimuli in a more specific and clear context. Further studies are necessary to explore those possibilities.

4.4. CONCLUSIONS

In natural environment, adult-infant interactions provide infants with simultaneous multisensory information such as social cues (e.g. visual contact, smiling faces, infant directed speech) associated with facial and body gestures (Gogate, Bahrick, & Watson, 2000). Interestingly, 6 to 8-months-old infants do learn word-objects associations from such complex environments (e.g., Gogate & Hollich, 2010; Gogate, Bolzani & Betancourt, 2006).

Our study showed that infants can learn object and action-words only when the visual referent were social faces and the speech component had harmony vowel as a facilitator for the segmentation task. Importantly, our results also demonstrate that infants were only able to do this audiovisual mapping process when the novel words were created with a specific phonological cue as a vowel harmony.

We believe that our results will open a series of questions, providing relevant evidence about the visual and speech cues on which infants rely during word-mapping process, as a contribution to future word learning studies.

4.5. MATERIALS AND METHODS.

Ethical approval for all experiments were obtained from Comité Ético Científico of Ciencias Sociales, Artes y Humanidades from Pontificia Universidad Católica de Chile. Before running the study, we asked to the parents to read, fill and sign a parental consent form. We recruited only healthy infants without any known neurological, visual or auditory difficulty. Infants were recruited at public primary health centers, were infants regularly attend for preventive control. The invitation to participate in all the experiments included information about the researchers who conducted the evaluation, the purpose of the study and an explanation about the procedure.

4.5.1. Experiment 1. Participants. 31 full-term ($M = 38.9$ wGA, $SD = 0.98$; APGAR 5-min = 9.03) 8-month-old infants ($M = 8.02$ -month-old, $SD = 0.43$; age range 7 months 15 days to 8 months 25 days) were tested. From the complete sample, 11 infants were excluded because disinterest/fussiness ($n = 5$), crying ($n = 3$) and eye-tracking (ET) missing data ($n = 3$), remaining a final sample of 20 infants.

4.5.2. Experiment 2. Participants. 30 full-term ($M = 38.8$ wGA, $SD = 1.1$; APGAR 5-min = 9) 8-month-old infants ($M = 7.98$ -month-old, $SD = 0.24$; age range 7 months 23 days to 8 months 15 days) were tested. We excluded 10 infants from the final sample because disinterest/fussiness ($n = 6$), crying ($n = 1$), ET missing data ($n = 2$) and technical problems ($n = 1$), remaining a final sample of 20 infants.

4.5.3. Experiment 3. Participants. 22 full-term ($M = 38.8$ wGA, $SD = 1.02$) 8-month-old infants ($M = 8$ -month-old, $SD = 0.26$; age range 7 months 17 days to 8 months 14 days) were tested. We excluded 2 infants from the final sample because disinterest/fussiness ($n = 1$) and ET missing data ($n = 1$), remaining a final sample of 20 infants.

4.5.4. Experiment 4. Participants. 29 full-term ($M = 39.1$ wGA, $SD = 1$) 8-month-old infants ($M = 7.93$ -month-old, $SD = 0.26$; age range 7 months 18 days to 8 months 15 days). We excluded 12 infants from the final sample because disinterest/fussiness ($n = 7$), ET missing data ($n = 4$) and technical problems ($n = 1$), remaining a final sample of 17 infants.

4.5.5. Stimuli.

4.5.5.1. Auditory stimuli. Different consonants and vowels in all experiments were used to create nonsense trisyllabic words with the software MBROLA speech artificial synthesizer (The MBROLA project, <http://tcts.fpms.ac.be/synthesis/mbrola.html>) using French voice (database *fr4*). Each word had a duration of 1368 ms with a mean pitch of 200 Hz, where consonants and vowels had a duration of 170 ms and 286 ms respectively. No silent pauses between syllables were placed.

Experiment 1. Nonsense words created to represent action-words were *dagava* and *mibiki* (i.e. random consonants – vocalic harmony), and for object-words, we used *puliso* and *tofanu* (i.e. random consonants and vowels).

Experiment 2. We used the same audio stimuli for experiment 2, but we switched in order to represent object-word with a vowel harmony. Namely, to represent object-words we used *mibiki* and *dagava* (i.e. vowel harmony - random consonants), and for action-words we used *puliso* and *tofanu* (i.e. random consonant and vowels).

Experiment 3. Nonsense words to represent action-words in this experiment were *nonina* and *lalelo* (i.e. consonant harmony - random vowels), and to represent object-words *digave* and *meboki* (i.e. random vowels and consonants).

Experiment 4. As experiment 2, we switched the audio stimuli of experiment 3 to represent in this case object-words with consonant repetitions. Namely, in this case for

object-words we used *nonina* and *lalelo* (i.e. consonant harmony - random vowels), and to represent action-words *digave* and *meboki* (i.e. random vowels and consonants).

4.5.5.2. Visual stimuli. For all experiments, we used the same visual stimuli for familiarization and testing. Namely, visual stimuli were created with two female faces from two different women. The object-words were represented visually by a frontal stationary picture of each female face; for action-words, we used two different head movements made by the two female faces recorded. One of the movements was a vertical movement similar to a “yes”, were the other was a horizontal head movement, similar to a “no” in our culture. All images and videos used in this study were in color.

4.5.5.3. Audio-visual stream. During familiarization phase, we presented a continuous audiovisual stream with the same features in all experiment. 2-min and 10-sec of a continuous stream with and artificial speech that co-occur with a visual representation, was constructed by the pseudorandom concatenation of the 4 trisyllabic nonsense words with stationary pictures and motions. In this continuous stream, object-words co-occurred with stationary female faces, and action-words co-occurred with moving faces.

The nonsense words were presented with a restriction that always the object-word was followed by the action-word. Transitional probabilities (TP) between adjacent syllables on each word were maintained constant equal to 1, were TP between words was maintained constant equal to .5. For visual stimuli, always the stationary face of one of the female faces, was follow by a moving face of the same female face. As always one stationary face can be follow by one of the two motions performed by the same female face, the TP for each visual stimulus was equal to .5. Fig. 4.3. Shows the continuous audiovisual stream projected.

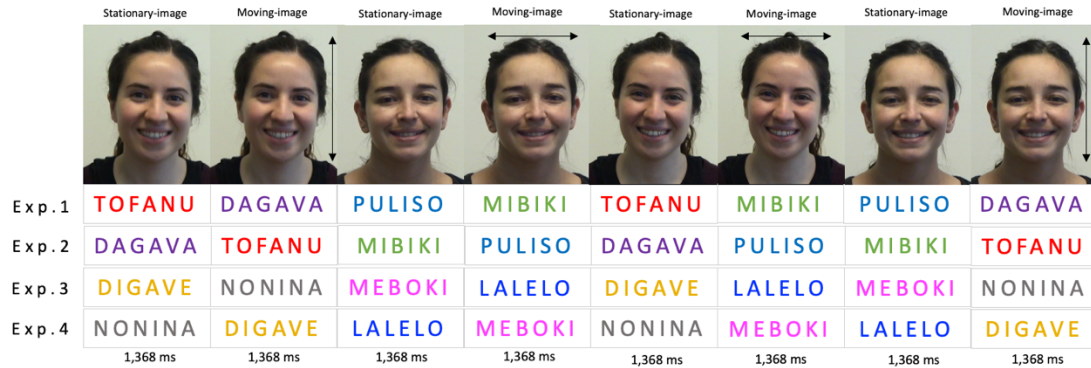


Fig. 4.3. Continuous audiovisual stream used during familiarization phase for all experiments. Four novel words were concatenated into a continuous stream. Each word co-occurred with one of four possible visual stimuli. Audio stimuli for each experiment are presented in the bottom of the illustration.

4.5.6. Procedure.

All visual stimuli were displayed on a 17-in. eye-tracker monitor (Tobii 1705; Tobii Technology, Stockholm, Sweden) with a screen size of 1.024 x 768 pixels and 16-bit color depth. The tracker automatically recorded each infant's binocular eye fixations at a sampling rate of 50 Hz every 20 ms.

Audio stimuli were played via external speaker situated in the center and under the eye-tracker monitor, to a 60 dB of amplitude.

The study took place at the '*Laboratorio de Neurociencias Cognitivas*', Escuela de Psicología in Pontificia Universidad Católica de Chile.

Testing was conducted in a soundproof and dimly lit room without any distraction. During all the experiments, infants were tested on their caregiver's lap in front of the eye-tracker monitor.

Before the study started, we explained to the parents that they cannot not intervene or influence the behavior of the infant during the study because the natural reaction of the infant to the stimuli presented was very important.

During all the experiments, parents wore dark sunglasses so they could not see the stimuli.

We first calibrated infant's binocular gaze using fixations longer than 100 ms on five centered points and on the four corners of the monitor.

We measure infants' looking behavior with a remote eye-tracking. The experiment consisted of two phases, familiarization and testing (pre-test and test). At the beginning of the study, a blinking yellow dot with a jiggling sound appeared in the center of monitor in order to attract infant's attention. After this, the familiarization phase began. In this phase, infants were habituated to the continuous audiovisual stream projected at the center of the Tobii eye-tracker monitor.

Once the familiarization phase ended, the experimenter began the testing phase. Eight trials were presented in a pseudo-random order for all experiments. The 8 trials were repeated twice (16 trials in total). Each trial consisted in two parts: pre-test and test.

At the beginning of pre-test, a simultaneous blinking yellow dot with a sound appeared at the center of the screen in order to attract infant's attention to this specific position, and two empty squares symmetrically positioned on each part of the screen (left and right) were shown. The empty squared and the yellow dot were situated on a grey background. After the presentation of the attractor-getter, infants listen to one of the type of nonsense words (either object or action-word, depending of the experiment) followed by a silence of 1000 ms. The aim of this part was attracting infants' attention to the center of the screen while they were listen to one of the nonsense word, and thus start the exploration of the following options from the center of the screen. Moreover, by adding this part to the procedure we wanted to introduce a sort of priming effect (Arias-Trejo & Plunkett, 2009) that could allow infants to prepare them to the following visual options.

Once the pre-test was finished, the testing part began. Two stationary faces or the two motions performed by the same face were presented on each side of the screen. These visual stimuli co-occurred with the repetition of the same nonsense word previously shown. We repeated the nonsense words twice separated by an inter-stimulus of 1000 ms of silence. Each video of the head movement was repeated four times. Figure 4.4. shows the structure trial for objects (A) and action-words (B).

The side of the presentation and the audio/visual stimuli were counterbalanced across the experiment. All the trials were presented in a pseudo-random order. We run the experiments using a Macbook Pro computer with PsyScope X (<http://psy.cns.sissa.it>). We recorded infants' gaze using a Sony Vaio computer running software ClearView (v.2.7, Tobii Technology AB).

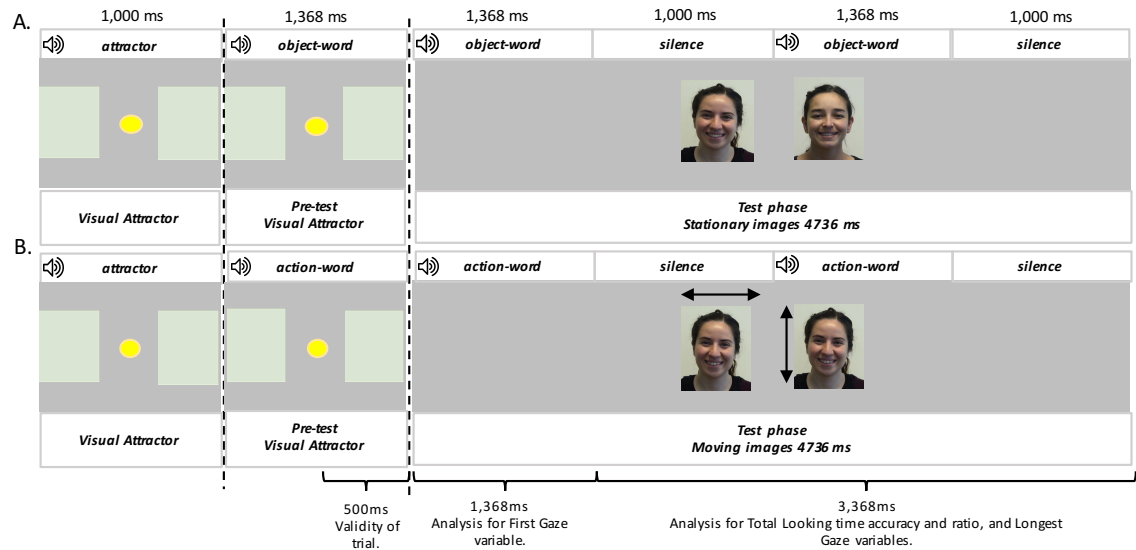


Fig. 4.4. Trial structure during the test phase. We illustrate the trial evaluating the learning of the object-words (A) and action-words (B). Time window of analysis is presented in the bottom of the illustration.

4.5.7. Analysis Overview.

4.5.7.1. Spatiotemporal regions of interest (ROI).

In the familiarization phase, we focused our analysis on the center region of the screen where the video was projected. Inside this region, we also analyzed of eyes and mouth regions of the female faces.

Videos and pictures of the two female faces were displayed over the central region of eye-tracking monitor with a size of 550 x 580 pixels of resolution on a grey background.

For the testing phase, we divided the screen into three equal parts (i.e., left, middle and right), and we measured the visual behavior over each one of those regions, from 500 ms before to 4736 ms after the test stimuli onset.

On each trial, videos and pictures were presented symmetrically on left and right regions with a size of 380 x 350 pixels delimited by light-yellow areas. Infants were positioned at a distance of 60 cm from the monitor.

At this distance, the central region for the facial stimuli subtended a $11.6^\circ \times 19.2^\circ$ area, and the left and right regions for the testing phase each subtended an $11^\circ \times 12.7^\circ$ area. A central yellow attractor used during the experiment, had a $2.9^\circ \times 2.9^\circ$ area.

4.5.7.2. Data pre-processing.

For the test phase, we first identified the valid trials, those where the infants gaze remained on the central attractor for 500 ms before the stimulus onset, and remained over the target for at least 100 ms. We also excluded trials with no eye-tracker data during testing part, even if infants looked at the fixation getter before the trial started.

4.5.7.3. Visual variables.

For the familiarization phase, we measured the Total looking time (TLT) that infants spent over the central ROI, eyes and mouth region during a time window starting from the onset of the continuous audiovisual stream until the end of this presentation. We also measured the TLT that infants expend during the presentation of each stationary faces and moving faces.

In the test phase, we measured the TLT defined as the total time of the infant's gaze fell over the ROIs. Also, we measured the longest gaze (LG), defined as the side where the longest fixation settled. And, we measure the first-gaze direction (FG), defined as the side where the first gaze arrived after the stimulus onset.

In each infant, the TLT for correct responses in each test trial was first computed as a ratio by dividing the TTL over the correct side (left or right) by the sum of the TLT over the correct and incorrect sides, to then be averaged across trials. This procedure allowed us to

normalize the visual data across infants, who may have a tendency to make longer or shorter gazes in any circumstance.

We also computed a TLT accuracy measure, on which every time the TLT the correct side was greater than the incorrect side, we transformed to a binary value (1 for correct and 0 for incorrect), and then averaged across all trials for each infant.

The FG and LG for each trial were classified as correct when they fell over the correct target. The LG was transformed to a binary value (1 for correct and 0 for incorrect) and then averaged across all trials for each infant. Similarly, the FG in each trial was a binary value that was averaged across all the trials in each infant.

4.5.7.4. Data analysis

We computed one-way ANOVA and chi-squared analysis in order to see if the groups were different in demographic variables.

In the familiarization phase, we compared the infant gaze toward the stationary versus the moving images, to eyes versus mouth in each of type of images by submitting the mean TTL ratio of each infant over the corresponding regions and events to a paired t-test (2 tailed, $\alpha = .05$).

We also performed a one-way ANOVA analysis between experiments in order to see if there were differences in the time infants looked to the center region where the audiovisual stream was projected.

During the test phase, we evaluated the object and action-words learning, by submitting the TLT, LG and FG averages of each infant in trials evaluating object and action-word learning, to a series of one sample t-test (2 tailed, $\alpha = .05$) comparisons, against the chance level at .5.

We also computed a MANOVA analysis in order to see difference between the four experiments for object/action-words, and two main groups for vowels and consonants.

Finally, we computed a univariate ANCOVA analysis in order to see if infants' age of evaluation had an effect in infants' performances in task where the vowels and consonants were implemented.

4.6. SUPPLEMENTARY MATERIAL

Table 4.1. Infants gender and Mothers' educational level distribution for Study 2

Non-significant differences for gender between experiments (Pearson's Chi-square, $\chi^2 = .717$, $p = .869$) and mothers' level of education (Pearson's Chi-square, $\chi^2 = 14.736$, $p = .471$).

		Experiment 1 N= 31		Experiment 2 N=30		Experiment 3 N=22		Experiment 4 N=29	
		Total	Percent	Total	Percent	Total	Percent	Total	Percent
Gender	Male	14	45%	14	47%	11	50%	13	45%
	Female	17	55%	16	53%	11	50%	16	55%
Mothers' educational levels	Elementary and middle school, incomplete	0	0%	1	3%	0	0%	0	0%
	Elementary and middle school, complete	2	6%	0	0%	0	0%	2	7%
	High school or equivalent, incomplete	2	6%	6	20%	3	14%	3	10%
	High school complete	11	35%	12	40%	10	45%	15	52%
	College incomplete, undergraduate studies	6	19%	2	7%	3	14%	1	3%
	Bachelor degree	10	32%	8	27%	6	27%	8	28%
	Master and/or doctoral degree	0	0%	0	0%	0	0%	0	0%
	Not informed	0	0%	1	3%	0	0%	0	0%

Table 4.2. Descriptive demographic variables for Study 2

Non-significant differences for demographic variables between experiments ($F_{(3,70)} = .646$, $p = .901$, $n_2p = .085$)

	Experiment 1 N= 31		Experiment 2 N=30		Experiment 3 N=22		Experiment 4 N=29	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Infants' age (months)	8.02	0.43	7.98	0.24	8	0.26	7.93	0.26
Gestational Age (weeks)	38.9	0.98	38.8	1.1	38.8	1.02	39.1	1
Mothers' age (years)	27.84	6.45	29.07	7.53	27.91	5.44	28.48	6.74
Weight at birth (gr)	3356.65	429.72	3352.07	437.33	3397.05	379.86	3270.69	491.53
Length at birth(cm)	49.84	1.59	49.22	2.22	49.72	1.76	48.97	2.18
APGAR 5min	9.03	0.5	8.97	0.18	8.91	0.29	9.03	0.19
Age of babbling (months)	3.73	1.14	4.03	1.59	4.26	1.51	4.62	1.18
Social Smiling (months)	2.77	1.28	3.07	1.2	3.32	1.17	3.48	1.57
Age Breastfeed (months)	5.52	2.47	5	2.82	6.36	3.15	6.22	2.92

Table 4.3. Infant Behavior Questionnaire – Revised (IBQ-R) (Putnam, Helbig, Gartstein, Rothbart & Leerkes, 2014)

Mean for each IBQ-R subscale, for each experiment.

Non-significant difference were found between all infants across experiments and Surgency/Extraversion subscale ($F_{(3, 70)} = 1.158, p = .1332, n2p = .047$), Negative Affective ($F_{(3, 70)} = .368, p = .777, n2p = .016$) and Effortful Control ($F_{(3, 70)} = .942, p = .425, n2p = .039$)

	Expe 1		Expe 2		Expe 3		Expe 4	
PSI	M	SD	M	SD	M	SD	M	SD
Surgency/Extraversion	5.29	.73	5.58	.62	5.30	.45	5.59	.81
Negative Affective	4.05	.91	4.09	1.26	4.24	.98	3.89	.87
Effortful Control	5.44	.67	5.37	.81	5.45	.74	5.76	.66

Non-significant bivariate correlations were found between all infants across the experiments and eye-tracking variables (all the results are for Pearson's correlation)

	Total time ratio (TLTr)		Total Looking time accuracy (TLTa)		Longest Gaze		First Gaze	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
Surgency/Extraversion	.079	.504	.122	.302	-.041	.731	.137	.245
Negative Affective	.068	.567	.066	.577	.048	.685	-.014	.905
Effortful Control	.106	.368	.123	.295	-.126	.286	-.015	.901

V. EXPERIMENT IN PROGRESS

Generalization of action-words.

5.1. Purpose of the experiment.

The purpose of the experiment was to assess if infants of ~7-month-old were able to recognize action-words performed by a novel agent, i.e. the face of unknown women. In the previous studies infants at this age were able to extract, segment and learn novel nonsense words (object and action-words) from continuous audiovisual speech and map this learning into specific visual stimuli (stationary female faces and moving faces). In study 1, most of the infants were able to learn at least one type of word during the study (see section Discussion and Conclusion section of study 1 for more details).

The current experiment modified the procedure of study 1, evaluating the learning of action-words introducing two new visual stimuli during testing phase. In this case we introduce two novel female faces performing the same actions used during familiarization. The aim of this modification was to evaluate infants' early recognition and generalization of the action-words to novel agents. Our hypothesis stated that by introducing this modification, infants would be able to map the previous learned actions-words to new agents. Indeed, in the study 1, the object-word learning may have interfered with that of the action word, for instance for a preference for a particular woman face.

5.2. Participants

The experiment was conducted on a group 18 full-term ($M = 38.6$ wGA, $SD = 1.14$) ~7-month-old infants ($M = 6.93$ -months, $SD = 1.49$; age range 5-months 8-days to 8-months 28-days; 8 girls, 10 boys). From the complete sample, 7 infants were excluded from the analysis for not completing the experimental protocol, remaining a final sample of 11 infants. With this constraint, this study is still in progress.

We recruited only full term healthy infants without any known neurological, visual or auditory problem.

The invitation to participate in our study included information about the researchers who conducted the evaluation, the aim of the study and an explanation about the procedure. Before running the study, we asked to the parents to read, fill and sign a parental consent form.

5.3. Stimuli and apparatus

All visual stimuli were displayed on a 17-in. eye-tracker monitor (Tobii 1705; Tobii Technology, Stockholm, Sweden) with a screen size of 1.024 x 768 pixels and 16-bit color depth. The tracker automatically recorded each infant's binocular eye fixations at a sampling rate of 50 Hz each 20 ms. Audio stimuli were played via speaker (Bose Soundlink mini II) situated at the center and under the eye-tracker monitor, to a 60dB of amplitude.

For the current experiment, we used the same sound stimuli as those for study 1 (see section stimuli in study 1 for more details).

For visual stimuli, we recorded new videos. Four female faces performing the same motions as study 1 were recorded, whereas two of these faces were used during familiarization phase, the other two female faces were used during test phase. Table 5.1. show the visual stimuli used in experiment 2. All the videos were recorded with a video-camera (Sony Handycam DCR-HC85) in a room with a light yellow/grey color background. A post video edition for time duration, brightness control and contrast was made with software Adobe Premier Pro CC (v.1.0.0, 2017) in order to maintain the same features for all the videos and pictures.






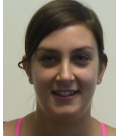
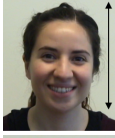



Familiarization phase				Testing phase		
	Sound stimuli	Visual Stimuli		Sound stimuli	Visual Stimuli	
Object-word 1	<i>tofanu</i>	<i>Female stationary face 1</i>		--	--	--
Object-word 2	<i>puliso</i>	<i>Female stationary face 2</i>		--	--	--
Action-word 1	<i>mabeki</i>	<i>"No" like head movement</i>		<i>mabeki</i>	<i>"No" like head movement performed by a new female face</i>	
						
Action-word 2	<i>degova</i>	<i>"Yes" like head movement</i>		<i>degova</i>	<i>"Yes" like head movement performed by a new female face</i>	
						

Table 5.1. Sound and visual stimuli for Experiment in progress.

With the recorded videos, we created a continuous audio-visual stream with the same features as study 1 for familiarization phase (Fig. 5.1) (see stimuli section of study 1 for more details)

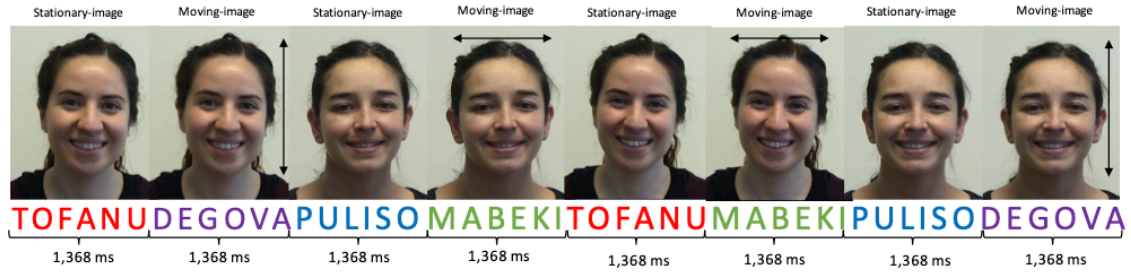


Fig. 5.1. Continuous audiovisual stream used during familiarization phase. Four novel words were concatenated into a continuous stream. Object-words co-occurred with stationary images and action-words with moving images.

5.4. Procedure

Testing place, instructions to parents and experimental procedure were the same as study 2 (see section Procedure in Study 2 for more details).

The experimental paradigm also consisted of two phases, familiarization and testing. However, during testing phase, we only assessed infants' recognition of the novel action-words. For this, we introduce two novel visual stimuli performing the same motions as familiarization (Figure 5.2).

Eight trials were presented in a pseudo-random order during testing phase, which were repeated twice (16 trials in total).

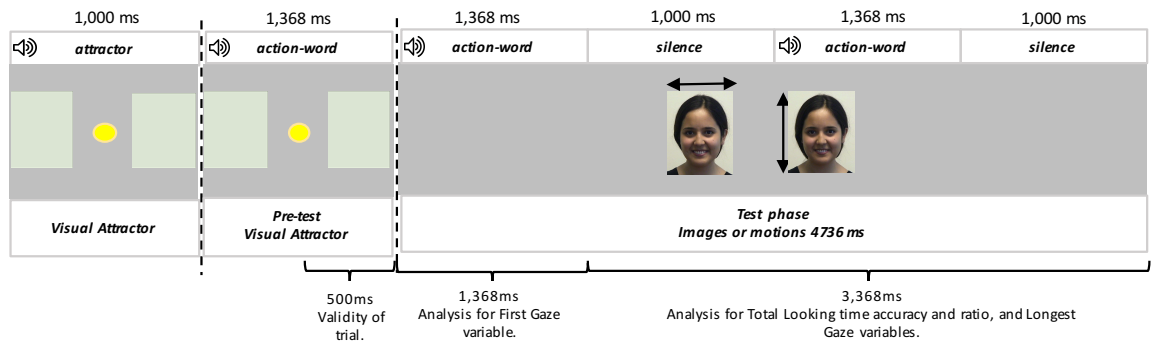


Fig. 5.2. Trial structure during the test phase. We illustrate the trial evaluating the learning of action-words. Time window of analysis is presented in the bottom of the illustration.

5.5. Data acquisition and analysis

Regions of interest, pre-data processing, time window and variables were the same as study 2 (see section data acquisition and analysis for more details).

5.6. Preliminary results

5.6.1. Familiarization phase

Infants looked at the central ROI for approximately $1.72 \text{ sec} \pm 0.325$, corresponding to the 78% of the total time in familiarization phase. We did not find significant differences between the mean TLT for stationary faces versus moving faces ($p=.235$). We find that the mean TLT for the eye region was significantly longer compared to mouth region during actions ($t(10) = 11.211, p<.001$, Cohens' $D = 5.013$) and action-words ($t(10) = 10.643, p<.001$, Cohen's $D = 4.816$). All these results indicate that during familiarization phase infants looked to the stationary images versus the moving faces, that co-occurred with object and action-words respectively, in a similar probability to learn word-images associations.

In order to see if the amount of time infants looked to the central ROI during familiarization phase, is related to infants' performance during test phase, we computed bivariate correlations analysis against the following variables. We found non-significant correlation in TLT accuracy (Pearson's correlation $r = .347, p = .296$), TLT ratio (Pearson's correlation $r = .237, p = .483$), LG (Pearson's correlation $r = .116, p = .735$) and FG (Pearson's correlation $r = -.255, p = .449$).

5.6.2. Test phase

Infants contributed an average of 11.5 valid trials ($SD = 3.3$) of an amount of 16. The average of eye tracker data for all valid trials was of 86% ($SD = 8.5$).

We performed an analysis for TLT, LG and FG according to the same criteria as study 2 (see section Analysis overview in study 2 for more details)

We found no significant differences between all variables against the level of chance .5 for TLTa ($p = .233$), TLTr ($p = .424$), LG ($p=.262$), and FG ($p=.262$).

All these results together suggest that infants at this age were not able to generalize into novel visual stimuli the learning of novel action-words.

Fig. 5.3 represent the proportion of correct looking time for all four variables computed during testing phase.

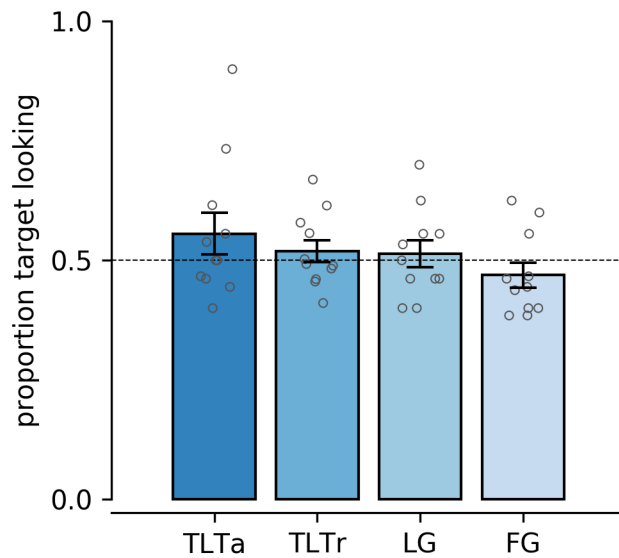


Fig 5.3. Non-significant results were obtained for all eye-tracking variables. We plot the proportion for target looking, for longest gaze (LG), total looking time (TLTa and TLTr), and First gaze (FG) compared with a proportion expected at chance (.5). Circles represents each infant for each experiment.

5.7. Preliminary discussion

The present results demonstrated that ~7-month-old infants were not able to generalize the learning of object-words into new exemplars. All the behavioral measures tested in this study, reflected that infants were not able to generalize action-words in new visual female faces performing the same head movements as the women that were familiarized. One possible explanation might involve that infants are too young to deal with generalization task from continuous stimuli.

VI. GENERAL DISCUSSION

The main objective of this doctoral thesis was to evaluate if pre-verbal infants were able to learn the associations between a novel object/action-word during a continuous audiovisual stream when visual stimuli are static and moving faces and words conveyed vowel harmony. Additionally, this doctoral thesis also pursued to explore whether infants were able to generalize the learning of object-words into novel exemplars.

Given that infants' experience during language acquisition involves multisensory stimulation, from which audio and visual information should be continuously integrated, it becomes relevant for us to explore the mechanisms on how infants become expert audiovisual learners from a multisensorial environment.

In this thesis, infants were first familiarized to a continuous audiovisual stream, on which each trisyllabic nonsense word was associated to a particular object (i.e. a static female face) or a particular action (i.e. a moving female face). During the test phase, we evaluated the word learning by measuring the preferential looking time in a two-alternative choice paradigm, recorded with a remote eye-tracker technique.

The aim of study 1 was to explore if 5-month-old infants were able to learn from an audiovisual statistical learning paradigm, both, object and action-words. Most infants demonstrated that they were able to learn (15 from 19 infants). However, from the pool of infants who succeeded on word learning, only a third of them learned both types of words, while other third learned only object-word and the last third only action-word.

Our results could explain that pre-verbal infants are familiarized with the concurrent stimulation provided by faces and speech (e.g. Kopp, 2014), and the continuous audiovisual stimulation with redundant information might have increased infants' probability to learn (Bahrick et al., 2004). Moreover, our task was highly attractive to infants because we use human faces as a visual stimulus, these have been reported to be preferred by infants compared to other visual stimuli (Gliga et al., 2009; Valenza et al., 1996). This type of stimulus may have facilitated infant's capacity to discover audiovisual associations in continuous streams, compared to other experiments

using another type of audiovisual stimuli. However, our findings did not demonstrate that infants were able to learn both types of words during the same task. Namely, our results indicate that most of the 5-month-old infants were able to learn at least one type of word at once due to a possible high cognitive demand for 5-month-old infants. We hypothesized that confronted to a multitask protocol some infants could have concentrated their cognitive efforts to learn one of the two types of words, suggesting also possible interference and competition in the processing of the two concurrent type of information (e.g. in Chen et al., 2017, recently reported in adults). Moreover, the interindividual variability could be interpreted as related to the young average age of the participants, where some infants may have had difficulty to deal with the learning of the two types of words. We thus in Study 2 facilitated the task and evaluated older infants.

In Study 2 we explored whether 8-months-old infants succeed to learn object and action words from continuous audiovisual stimuli when segmentation was facilitated by adding vowel harmony as a phonological cue, which has been previously reported as a helper cue for such task (Mintz et al., 2018). Specifically, we explored if infants relied on vowels harmony -- and consonant harmony -- in order to learn both object and action-words, and how these sounds may help infants during the learning of this two type of words. Previous evidence suggests that infants relied on vowel sounds in order to memorize and differentiate novel words (Benavides-Varela et al., 2011; Bertoncini et al., 1988), and use vowel harmony to segment fluent speech, even if this are not exposed to a language with vowel harmony properties (Mintz et al., 2018). However, previous evidence also suggests that infants rely on consonants during word learning (Hochmann et al., 2011).

We tested four groups of 8-month-old infants, in four different experiments. In two of them we used novel object and action-words with vowel harmony (vowel repetition) and tested this two types of word in separated experiments (experiment 1 for action-words and experiment 2, for object-word). In the other two experiments, we created novel object and action-words with consonant harmony (consonant repetition), and again tested these

two types of word in separated experiments (experiment 3 for action-words and experiment 4, for object-words).

Together our results demonstrate that 8-month-old infants are able to segment, extract and learn novel object and action-words from a continuous audiovisual stream, and map these four learned words to stationary and moving images, but only when the novel words had vowel harmony, and not when they conveyed consonant harmony. These results can be explained by to the salience of vowels over consonants (Mehler et al., 1996), and a vowel bias presented in younger infants (Nazzi et al., 2016), which would potentiate the infant's preference for human faces (Gliga et al., 2009), help them to succeed in this task.

A third experiment also was conducted in this thesis project (experiment in progress) to evaluate whether preverbal infants were able to generalize the learning of action-words to new visual agents. Our preliminary results ($n = 11$) indicated that infants cannot generalize action-word to new agents.

6.1. SIGNIFICANCE

To our knowledge, this is the first study showing that young infants succeed to learn object and/or action-words from continuous audiovisual streams, when visual stimuli are faces and speech provided vowel harmony as a phonological cue.

These results would serve as an input for current models of the mechanisms explaining word-learning from continuous inter-sensorial stimuli during early development. Specifically, we propose that, increasing the salience of the stimuli and facilitating the speech segmentation benefit both the processing of continuous audiovisual streams as well as the binding, probably at low-level of the audiovisual integration, which is finally saved in memory for further recognition.

Another contribution of our studies regards the methodology. We hope to have provided to the scientific community with new methods to evaluate word-learning during early infancy.

Indirectly, our results might also be relevant to researchers exploring the early steps of the acquisition of grammatical categories of words. Specifically, conceiving our results as rudimentary forms of the object-noun and action-verb associations that infants learn at older ages.

Indeed, the object and noun word learning map referents as tokens, while actions and verb words map types of dynamic events involving relations among object. Although the issue is currently under review, a number of studies suggest that children learn nouns earlier, faster and easier than verbs (Waxman, Fu, Arunachalam, Leddon, Geraghty & Song, 2013; Golinkoff & Hirsh-Pasek, 2008) and our results can contribute to a better understanding of the early steps of the acquisition of these two grammatical categories.

6.2. LIMITATIONS AND PROJECTIONS

We consider the following limitations and projections according to the results obtained in the different studies.

First of all, one limitation presented in the previous experiments has to do with the implementation of only one behavioral technique (remote eye-tracker) in order to measure infants' performances in each task. In order to complement our findings, we could have implemented a co-register with behavioral and neural techniques at the same time. We expect to advance in this challenge to and develop a co-registration protocol including these types of measures in preverbal infants. We believe that such behavioral-neural approach would significantly help us to disentangle the processes underpinning the integration of the visual and speech signal into a common referent, for object and action-words.

Secondly, study 2 evaluated the learning of object and action-words in a between-groups design. Namely, we tested different groups of infants on their ability to learn objects or action-words during a different task, being exposed to a single condition per experiment. Further studies with a within-group design, will be necessary to evaluate if

infants are able to learn both objects and action-words during the same task, implementing vowel harmony as a phonological cue in the audio stream.

Moreover, in order to test our hypothesis that younger infants are not able yet to exploit the information from consonant harmony, further longitudinal studies testing the same infants at older ages, would be necessary for explore this possibility.

Finally, because of time limitations we were unable to get conclusive results about the generalization of action-words, which is crucial to prove that infants learned two types of words, intrinsically different. We thus plan to advance on such subject in future.

VII. ADDITIONAL OUTPUT DURING THIS DOCTORAL THESIS

Book chapter:

Cognitive Models of Language

Jara, C., Gutiérrez, C., Cortés, P. & Peña, M. (2018). Modelos Cognitivos del Lenguaje. In Labos, E., Slachevsky, A., Torralva, T., Fuentes, P., and Manes, F. (Eds.) *Tratado de Neuropsicología Clínica*. 2nd Ed. (pp.187-200). Ciudad Autónoma de Buenos Aires: Librería Akadia Editorial.

Translated to English, original version of this manuscript in Spanish.

COGNITIVE MODELS OF LANGUAGE

10.1. Introduction

Linguistic behavior is governed by a set of conventions that must be shared between listeners and speakers of a given language. These conventions regulate how we organize the units of language without meaning to generate units with meanings. The interest in exploring the rules and models that explain human language dates back to the dawn of humanity. However, the scientific study of human speech and its modeling has been systematized only since 1950, after the pioneering studies of Noam Chomsky. Chomsky (Chomsky, 1957) began the exploration of two basic questions relevant to human speech: What does it mean to know a language? and How is this knowledge acquired? In the following decades, several proposals have been generated in the light of the advances in empirical research from cognitive psychology, neuropsychology, neurosciences, psycholinguistic, and artificial intelligence. To the date, it has not been possible to conceive a general model of language that accounts for the different aspects of this cognitive ability. On the contrary, hundreds of specific models have been proposed, and their discussion goes beyond the bounds of this chapter. However, this chapter succinctly describes relevant aspects of the nature and organization of human language, and the modeling of language perception and production; particularly valid in healthy, normo-listeners, right-handed, and monolinguals.

10.2. What is human language?

The language is a system with symbolic signs (Holdcroft, 1991) that helps us to communicate (Jakobson, 1960) with sounds and/or gestures as perceptual primitives, and to configure our thinking (Cassirer, 1923). Human language is creative, unpredictable, and has specific properties that make it different from all known communication codes in non-

human animals. According to Charles Hockett (Hockett, 1960) the specific properties of human speech are:

1. Duality of patterning: the combination of a sound system (without meaning) and a grammatical system (with meaning).
2. Productivity: the ability to create and understand statements never made before.
3. Arbitrariness: the symbolic signs used as words do not resemble the concepts they represent.
4. Displacement: the ability to refer to non-current events and things that are not present.

A system like this, that allows us to generate infinite statements from a reduced set of sounds, words, and rules has not been found in any species.

10.2.1. What does it mean to know a language?

A language can be defined as the set of correctly constructed sentences according to that language, and its structural descriptions (Chomsky, 1957). In general, knowing a language means having learned its repertoire of sounds, words and grammar rules (Pinker, 1984).

10.2.2. The language, how is it structured?

Sentences are concatenations of words, and words are concatenations of sounds. This linearity describes the dynamics of the processes of perception and production that result from the integration over time of different linguistic units. In general terms, the study of linguistic units includes phonology, lexicon, morphology and syntax. Phonology summarizes the properties of the sounds of a language. This includes distinctive features, phonemes (which in Spanish correspond roughly to the sounds of consonants and vowels) and prosodic variations (rhythm and intonation). The lexicon corresponds to the linguistic units stored in our mental dictionary. Words can convey meanings (content words, e.g.

verbs) or serve mainly as a connection unit to other words (function words, e.g. conjunctions). Morphology refers to the possibility of changing the form of words according to their relationship with the words that generate it. For example, in Spanish it is possible to modify the word /niño/ to /niña/ changing the suffix according to gender. Finally, the syntax describes the order of the words within the statements following grammatical rules. For example, in Spanish, the sentences follow a subject-verb-object order.

According to the tradition of generative grammar, only syntax and morphology are productive, while phonology and semantics are interpretive. Syntax and morphology are conceived in terms of a set of rules that only generate well-formed sentences that allow us to understand the relationships between the elements of the sentence. These rules regulate the combinatorial functions of words, and make possible the generation of a potentially infinite number of phrases and sentences, using a finite number of words. For example, the sentences: "He told me today that he will not be able to come" and "He told me that today he will not be able to come", evoke different meanings in relation to when someone cannot come. This effect is caused by the fact that grammatical rules allow recursive cycles in the construction of sentences. It is worth noting that in a statement a rule can contain itself. This loop of recursive functions is observed in all known languages.

10.3. Cognitive models of the acquisition of native tongue

It is amazing how children learn their mother tongue without the need of directed instruction and even in the most varied conditions of upbringing. The modeling of acquisition of maternal language(s) has been one of the most explored. A few hours after birth, and during the first months, infants shows remarkable abilities to treat linguistic stimuli: it discriminates syllables and sounds extracted from both its mother tongue and foreign languages (Kuhl, 20014), prefers the human voice to other vocalizations (DeCasper & Spence, 1986; DeCasper & Fifer, 1980). Apparently, infants are born with abilities and structures that are particularly sensitive to linguistic stimuli (Chomsky, 1957;

Lenneberg, 1967). However, infants do not benefit from this general knowledge, but instead they must learn the specifications of at least one natural language. To do this, they must be stimulated systematically on the language practiced in their environment, so, infants only learn the languages to which they are exposed and begin to create new phrases around 3-years-old. Language acquisition models try to explain the acquisition of phonology, lexicon, and grammar; and can be grouped into deductive and inductive models.

10.3.1. The symbolic-deductive models

These models emphasize the role of the ease with which humans acquire speech, even when the speech is a "poor" stimulus. The linguistic stimulus reaches children with an indeterminate and degraded form. Indeed, what characterizes a 'well done' phrase in a particular language is not systematically reflected in any physical property of speech (indeterminacy). Moreover, speech is full of omissions and complex intonational or pronunciation variations (degradation).

Towards 1960, there was a defense on the hypothesis to the phenomenon that the stimulus given is so inconsistent and the speech so complex, humans could be born endowed with an innate biological predisposition that facilitates this learning. By 1980, it was further suggested that the human mind is equipped with an innate cognitive apparatus, operationalized in terms of symbolic and categorically defined principles and rules. Both indeterminacy and degradation could be resolved through symbolic transformations.

From a neurocognitive perspective, these systems could be implemented in neural modules. Cognitive modules can be understood as groups of neurons with similar functional properties, located in discrete regions with sharp edges and involved in the preferential processing of a type of stimulus.

In the beginning, modular models propose incremental models. This means that cognitive functions are preferably accessed serially. For example, "syntax-first" theories suggest that listeners access the syntactic structure of statements based on the information

in the grammatical category of the words perceived, regardless of the semantic information they contain. The extraction of meaning would occur in a second stage (Mehler & Dupoux, 1990).

The vulnerability of the symbolic proposals underlies on the lack of empirical evidence of a symbolic system. So far no one has traced the course of a symbolic representation in the brain.

10.3.2. The symbolic-inductive models

An alternative view on the ability to learn speech is proposed by the connectionism, supported particularly by the principles of artificial intelligence and the evidence obtained from computer simulation. Computer simulations of connectionism are scientific assessments between theoretical and experimental science that assume a distributed map of cognitive functions. A map can be understood as a group of distributed neurons with similar functional properties, characterized by a gradual progression in the preference to process certain stimuli in the brain. This perspective proposes a subsymbolic-inductive representation. Connectionist models propose a distributed knowledge of language which does not directly encode symbolic information. There are fixed interconnected elements ('neurons') that by themselves do not represent anything, but together can generate a linguistic representation. The interactions between the fixed elements can be of highly complex, and the observed behaviors varied according to what they learned. Optimally, the systems must have the ability to learn, modulating the weight of their interactions, and continually self-organize. Distributed interactive models assume that the listener uses different types of information immediately in an interactive way (Fodor, 1983; Bates & MacWhinney, 1987; MacDonald, Pearlmutter & Seidenberg, 1994; Marslen-Wilson & Teyler, 1980).

Connectionist models of language can solve the problems of deterioration of the linguistic signal. Indeed, the acoustic signal can be improved by simulating the biological constants of the human ear. However, the problems of indeterminacy are less possible to

explain without recognizing that simulate symbolic representations of linguistic stimuli. In general, connectionist models question the need for rules and linguistic categories, and propose models that result from the exploitation of keys observable in the stimulus and from the relationships that result from the connections of those keys (Taraban & McClelland, 1988; Rumelhart & McClelland, 1986). One of the problems with these models is that they cannot satisfactorily explain linguistic facts based on infrequent rules, such as adding a /-s/ in the German plural 18 (Marcus, Brinkmann, Clahsen, Wiese & Pinker, 1995). In addition, the major limitation of these models is the excess of information. Indeed, given its excessive statistical power, computer simulations find the correct answer in the midst of a huge number of events never observed in a natural language, having great difficulty in generalizing linguistic aspects observable in more than one language. The brain does not have the architecture of a computer, for example, it does not have destination memory (Squire, 1987), so that computer simulations cannot always find what the human mind and brain find. However, they help to understand linguistic processes, particularly those of a statistical nature.

In general, both deductive and inductive cognitive models that have best explained the acquisition of language are those that simulate the architecture and biological principles observed in the human mind and brain. For example, those that include variations in the size of the working memory.

10.4. Models of language perception

For most healthy adults, understanding speech is an easy task. However, the understanding of speech reveals great enigmas such as linearity and the absence of invariance between the acoustic and the linguistic, which have largely guided the studies and models of perception.

Linearity is illustrated by the fact that although listeners perceive speech as a concatenation of discrete sounds (i.e. phonemes, syllables, words), there is no correlate of

discrete units in the physical signal of speech. Sounds and speech gestures occur continuously overlapping each other (Fowler, 1995). Until today, there is no known physical evidence that systematically indicates where a linguistic event ends and where the adjacent begins. In various stages, speech understanding results from the extraction, manipulation, memorization and mapping of these discrete units into linguistic information.

The lack of invariance describes the fact that although the sounds that are made during speech differ greatly in their physical properties, the linguistic mapping does not vary. The same word spoken slowly by an adult man differs, among others, in its fundamental frequency, in its formants, and in its speed of realization of that quickly told by a young woman. However, in the same context, humans map the same linguistic information. Indeed, the linguistic representation of the phonemes does not change even when their physical realization varies significantly depending on the neighboring phonemes (Lisker & Abramson, 1967), the size of the place of articulation (Nygaard & Pisoni, 1995), the identity of the speaker, or the speed of speech (Hillenbran, Getty, Clark & Wheeler, 1995). Listeners perform a series of perceptual adaptations to filter out irrelevant variations and identify the occurrence of phonemes. It is thought that listeners adjust their perceptual system to the specific characteristics of the speaker, for example, the proportion of formants (Syrdal & Gopal, 1986; Strange, 1999; Johnson, 2005), and the speed of speech. This ability is probably acquired early in childhood.

One effect of the adjustments on phoneme perception is categorical perception. Adults perceive phonemes as discrete units, since the limits of perception of a phoneme coincide with narrow windows in their physical changes. For example, the difference between the /b/ and /p/ is partly because the time difference with which the vocal cords begin to vibrate when pronouncing them. Thus, the word /bar/ begins with a rapid vibration of the vocal cords, which makes it feel voiced, while the word /pair/ presents a small delay in the beginning of this vibration, which makes it feel mute. In English, the perception of a /b/ occurs when the delay in sonority is from 0 to 20 ms, and in the case of a /p/ if it is greater than 25 ms. Only in the short window of 20 to 25 ms, adults have

great difficulty in differentiating these phonemes. In this way humans perceive without effort or difficulty the differences between phonemes presented in isolation or within sentences. On the contrary, we are bad at finding differences between different pronunciations of the same phoneme. This ability is known as categorical perception (Liberman, Harris, Hoffman & Griffith, 1957), and we do not observe it systematically in the processing of non-linguistic auditory stimuli (Kewly-Port & Luce, 1984). Although not all phonemes show categorical perception, this cognitive phenomenon has been very useful to illustrate the complexity of speech perception and to model how the mind works when it makes language. In the following, some models of speech perception are briefly presented.

10.4.1. Motor theory

The motor theory of speech perception postulates that speech is perceived according to what is pronounced (Liberman et al., 1957). When speaking, we listen to what we say and learn about the articulatory properties of the pronunciation of sounds. Articulatory gestures, such as bring lips together, are units of perception that directly provide the listener with phonetic information that better explains the absence of invariance observed in perception. Some aspects of speech articulation are visually available for those who observe the speech of another person. The clearest example of this is the McGurk effect, which reflects the modification of perception when we integrate auditory and visual speech information. In effect, if we listen to the syllable /ba/ and simultaneously we see a video of a person pronouncing the syllable /ga/, the adults listen /da/ (McGurk & MacDonald, 1976). This happens because we have a vast experience in listening and seeing how people speak, so that the mental representation of phonemes would be the product of creating a multimodal representation of the way they are performed. In effect, untrained representations of phonemes are not stored in memory and are not considered in their representation. For example, the match between the sound /a/

and the written letter /a/ does not occur naturally even in literate people (Fowler & Dekle, 1991).

10.4.2. TRACE model

The TRACE model of phoneme perception and access to words (McClelland & Elman, 1986) is a connectionist model that describes speech as a process in which linguistic units are accessed and processed at different levels incrementally. TRACE proposes that words are represented on three levels: distinctive (or acoustic) features, phonemes, and words. Processing at each level occurs in units called *nodes* that interact with each other and skew the processing to the solution most strongly represented in memory. The perception of speech in time would occur as follows: initially the distinctive features of the first phoneme are accessed, then the entire phoneme is activated and the word is discovered incrementally. At the beginning the activation is bottom-up, but as moving forward in the discovery of the possible word, a top-down activation is triggered that ends with the access to the word. For example, when we perceive the vibration of the vocal cords at the beginning of the word, we activate the words that we have stored in our mental dictionary that begin with a consonant-sound, then in tens of milliseconds we perceive that the initial phoneme is /b/ and we restrict our expectations of the word to those of the mental dictionary that begin with /b/. Then, as we move forward in listening to the word phonemes, we continue to refine our word expectations to reach a point where we know the word, even before it has been finished pronouncing. For example, if we listened to /bana/ it is most likely that the word /banana/ is activated even before we finish listening to the whole word, since a top-down effect makes us prefer it. Neighboring words like /banality/ are more distant competitors, since they are used less frequently and would be used as second alternatives. Numerous models of speech perception include context as a crucial variable. The effect of the context and the ability of the human cognitive system to anticipate the closure of events is evidenced in some studies that show that, if the

context facilitates it, the detection of a target phoneme occurs even when the phoneme in has been totally removed (Warren, 1970).

10.4.3. Fuzzy logic model

The perception model 'fuzzy logic theory' proposes that the human approaches their perceptions of the acoustic and phonemic reality to the context, and to the experience and expectations of the listener (Massaro, 1989). In this model, linguistic units would be mentally represented in prototypes, corresponding to 'well-made' forms. However, humans often perform less than 'well-made' forms. The decision to evaluate whether a not 'well-made' form corresponds to a given phoneme or word depends on the observer, this is done based on other multiple keys or information sources that are not necessarily linguistic. In this way, the perception of speech is the result of the interpretation of the integration of information from different sources. The same sound can be interpreted in one way by a listener while it is interpreted in another way by another listener. It is common to see that two speakers comment on their differences. For example, in Spanish, when listening /doceymedia/ a speaker can listen to '2 y media' (two and a half) while another listen to '12 y media' (twelve and a half).

10.4.4. Acoustic analysis of distinctive features

This model proposes that human cognition is sensitive to the acoustic properties of linguistic sounds. In this way, the perception of certain acoustic properties allows the identification of a phoneme, the notion of absence of invariance is not necessary. In this model what is perceived are groups of distinctive features that differ from one phoneme to another in a quantum manner, not continuous. The passage from one phoneme to another involves exceeding the thresholds of perception of the set of distinctive features. The acoustic indicators also contain information about the articulatory gestures that produce them (Stevens, 2002).

10.4.5. Theory of the exemplar

The exemplar model refers to the fact that the recognition of the word occurs separately from that of the speaker. In this way, the variation through the speakers or the speed of speech is a noise that needs to be filtered. The model proposes that listeners separately store information about each utterance pronounced by each speaker, achieving a more efficient representation of the words spoken by familiar speakers (Johnson, 2005). This model requires large memory requirements and assumes that the speaker recognizes their articulatory gestures separately from access to words.

10.4.6. Theory of analysis by synthesis

The perception of speech is based on a system of comparison of the analysis of the acoustic stimulus and the models generated by a set of abstract rules common to perception and production. This model needs a set of rules that have been defined during the learning of a language. The comparison is made between the articulatory correlate and the acoustic representation of the speech (Fromkin, 1971).

10.5. Language production models

Normally, adults produce four syllables and about ten to twelve phonemes per second. The words are chosen from an extensive reservoir of words, the mental or lexical dictionary, which contains about 50-100 thousand units. The study and modeling of speech production try to explain how the human being constructs a spoken statement based on a mental concept. The study of speech production has two main limitations: the impossibility of directly evaluating the processes that generate the concepts, and the infinite possible combinations that can be articulated to generate the same message. In particular, the way in which thought is carried out through a linguistic format and how it involves particular structures of the nervous system is still a mystery. These limitations

have led, even under great controversy, to support many of the models in the study of pronunciation errors. For example, agrammatical speech is extremely rare (2% of human population). However, in a natural conversation, doubts, pauses, and repetitions occur more frequently, that is, one every seven or eight words. Errors in production can involve phonemes, syllables or complete words and, in most cases, occur with certain regularities that allow their systematic study, accounting for the hierarchical nature of the speech organization. In effect, phoneme and word errors occur mainly between units in the same category. For example, the vowels are frequently interchanged among them, as well as the consonants, but the exchange of vowels by consonants is practically never observed or vice versa. Also, verbs are exchanged with verbs and nouns with nouns but not between them. In addition, errors occur mostly within an intonational phrase. The occurrence of these errors suggests that they occur after defining the phonological properties of the words. The errors would be due to an incorrect selection of phonemes or words once the appropriate word was chosen. Alternatively, inadequately distributed pauses within sentences are more frequent in syntactically more complex sentences, indicating that they coincide with an active effort of organization in the speech program. In the following subsections some models of speech production are presented.

10.5.1. The statement generator model

Developed by Fromkin (Fromkin, 1971), it proposes a top-down generator that distinguishes six stages of sentence representation. The stages are:

1. The generation of the meaning to be transmitted.
2. The mapping of meaning in a syntactic structure.
3. The generation of an intonational contour.
4. The selection of words from mental dictionary.
5. The selection of phonological aspects of the unit to be pronounced.
6. The generation of the motor speech program.

10.5.2. The Garrett model (Garret, 1975)

It makes certain explicit aspects of Fromkin's model implicit. This model differentiates a conceptual level from a functional level, and a positional level of the units within the linguistic units to be produced.

10.5.3. The Dell model

It is a connectionist model that also includes the concept of disseminated activation. Dell's two-stage interactive model (Dell, 1986) is a neural network that is activated in two stages, one at the semantic level and the second at the phonological level. The semantic node disperses its activation towards the lexical node and this towards the phonological node. The activation cascade is interactive because all connections are bidirectional. The functionality of the bi-directional connections would make possible the fluency in the lexical selection and explain some symptoms observed in aphasic patients who have good access to the word but poor access to the phoneme or vice versa.

10.5.4. The Levelt model (Levelt, 1989)

It resorts to several stages that are produced in series. Each stage must be completed to start the next one. The first stage is the activation of the lexical concept. The lexical concept enables the intermediate form called lemma. The lemma contains syntactic information of the word. Once the lemma has been activated and selected, the phonological form of the word can be retrieved. The phonological form includes the morphology of the word (for example, the adhesion of suffixes), the metric structure (for example, the accent), and the list of phonemes. Segments, syllables and morphemes are included in the metric form one by one to build a larger linguistic unit. It assumes that phonemes and prosody are planned independently, first phonemes and then prosody,

which does not explain the influence of prosody on the articulation of phonemes. Once the phonological form of the word is defined, the preparation of the articulation of the word begins. The model does not specify how the articulation takes place, but proposes that there are certain syllables of frequent use that are accessed more easily than others. Finally, an auditory self-monitoring system checks speech production errors. Speech errors of the type tip-of-the-tongue are explained by difficulties from the transition of the concept to the selection of the word. The main objection to this model is that it proposes serial processing. The WEAVER model (Levelt, Roelofs & Meyer, 1999; Roelofs, 1997) is a connectionist implementation of the Levelt proposal.

10.5.5. DIVA (directions into velocities of articulators)

DIVA (Guenther & Perkell, 2004) uses what we hear and feel to guide production, is based on a connectionist model and focuses on the characterization of the articulatory processes themselves. DIVA learns to build auditory maps from articulatory gestures like what happens in babbling. It assumes that there are several ways to produce a perceptually acceptable speech that is refined with learning. The neural correlate of DIVA involves the motor, auditory, and orosensory areas of the brain and cerebellum. This model adds to others, such as Levelt for example, the description of auditory feedback and the precise description of the programming of articulatory gestures during speech production. From this perspective, this model considers the influence of prosody that surrounds phrases or words. The articulation of the phonemes is more intense in the phonemes that appear at the beginning with respect to those that appear at the end of the sentences. Similarly, they are larger for phonemes at the beginning of the word than at the end of it. This symmetry would be useful to correct the lower predictability of the phonemes located at the beginning of the sentences (Keating, 2005).

10.6. Neurocognitive models of language

The design of neurocognitive models of language requires a focus on the nervous system, which describes the way in which the activity of a set of organs generates language. Numerous approaches have been proposed from the study of brain injuries. In fact, the classic model of Wernicke-Lichtheim, S. XIX (Wernicke, 1874) has been the basis of modern models of language processing (Hickok & Poeppel, 2007; Price, 2000; Friederici, 2002). However, in recent decades the greatest advances have been based on the use of neuroimaging methods. Particularly from the functional magnetic resonance, which allows to detail the anatomy of the linguistic processes, and the magneto-electroencephalography, which informs about the temporal dynamics of these. In this section, the neural bases of language are exemplified in two themes: cerebral lateralization and ventral-dorsal systems.

10.6.1. Brain lateralization

It has been suggested that, at the level of the cerebral cortex, there would be inter-hemispheric and regional differences in language processing. Recent studies have proposed that the basis of these differences lies in the type of processing developed by the neural networks of both hemispheres. The right hemisphere (RH) would be endowed with neural networks specially dedicated to the processing of speech sounds in long time windows, while the left hemisphere (LH) would have specialized networks in linguistic processing in short temporal windows. A 'long time window' corresponds to a range of hundreds of milliseconds, and corresponds to the time necessary to perform and process linguistic units with meaning, such as a word, a syllable, or a phrase. The processing of the 'short temporal window' would involve the analysis of linguistic units that occur in tens of milliseconds, such as phonemes, and would allow to discriminate events like /la/ de /las/. The global access to the meaning of the phrases requires the optimal integration of both systems operating at different time scales (Goldsmith, 1990; Meyer, 1992).

Current evidence supports at least the proposal that LH processes brief linguistic events. In effect, LH is predominantly activated during tasks in which it is required to categorically process phonemes, while tasks that require the analysis of more extensive events, such as word recognition and intonational aspects of speech, present an activation of both hemispheres.

10.6.2. Ventral and dorsal systems

The neuroanatomy of language remains unclear in many aspects, however, recent data support the idea that in adults there would be ventral and dorsal circuits in both cerebral hemispheres. The ventral circuits would be involved in the perception and understanding of the linguistic object of speech, similar to the ‘what’ path observed in the visual system (Hickok & Poeppel, 2000, 2004; Rauschecker, 1998; Scott, 2005; Scott & Wise, 2004). The dorsal circuits, strongly lateralized to the left, would allow the mapping of the sub-semantic components towards the anterior regions in the frontal lobe, influencing the articulatory circuits that allow the production of speech. The ventral circuits involve the middle temporal gyrus (MTG) and superior temporal gyrus (STG), while the dorsals would correspond to the posterior regions of the frontal lobe, the parietal operculum and the more posterior-dorsal region of the temporal lobe.

The STG in its most anterior regions and the upper temporal sulcus (predominantly to the left) map the phonemic component of the language (Price, Wise, Warburton, Moore, Patterson & Howard, 1996; Liebenthal, Binder, Spitzer, Posing & Medler, 2005; Indefrey & Levelt, 2004). From there, this information is widely distributed to the semantic network in a large part of both hemispheres (Damasio & Damasio, 1994). The syntactic networks would be distributed from the anterior region of the temporal lobe (ATL). The ATL area is more active when people read or listen to stories than an unstructured list of words or sounds (Friederici, Meyer & von Cramon, 2000; Humphries, Willard, Buchsbaum & Hickok, 2001; Humphries, Love, Swinney & Hickok, 2005), and their

injury causes problems in understanding syntactic structures (Dronkers, Wilkins, Van Valin, Redfern & Jaeger, 2004).

The backbone networks would be related to the translation of sound into action. It would be similar to the 'where' path observed in the visual system, and would interface with the motor system (Hickok & Poeppel, 2000, 2004). The auditory-motor integration is crucial for speech. Wernicke's model includes a direct path (Wernicke, 1874), the motor theories of auditory perception assume a direct motor auditory connection (Lieberman & Mattingly, 1985), and recently it has been suggested that the temporoparietal area around the Sylvian fissure (Spt area) would be responsible for the translation from sensory to motor (Hickok, Buchsbaum, Humphries & Muftuler, 2005).

10.6.3. New neuronal proposals of speech representation

Semantic neural maps

Recent studies have shown that when normal adults access the meaning of words, similar brain regions are activated in different people. A recent study (Huth, de Heer, Griffiths, Theunissen & Gallant, 2016) that uses the technique of functional magnetic resonance imaging (fMRI), reports that different people listening to words with social content (e.g. friend), that correspond to categories (e.g. colors), indicating body parts (e.g. the word fingers), and also corresponding to numerical symbols or codes (e.g. four), activate areas of the lateral parietal cortex (LPC), supporting the idea of the existence of a common semantic map at neural level. Specifically, the semantic category that refers to words with social valence, activates a central area of the LPC, while the other types of words mentioned activate the peripheral zones of the LPC. The highly specific nature of the semantic representation of words in the LPC exemplifies an advance in the contribution of this type of study to neuropsychology. In fact, the existence of common semantic neural maps offers new techniques to face the challenge of modern neuropsychology to establish specific correlations between brain and meaning.

New areas of language

A recent study using a multimodal approach identified 97 new cortical areas, of which area 55b is involved in language comprehension tasks (Glasser, Coalson, Robinson, Hacker, Harrell, Jacob, Ugurbil, Anderson, Beckmann, Jenkinson, Smith & Van Essen, 2016) (Figure 10.1.).

The multimodal approach integrates information of various kinds. Specifically, it includes data on the microstructural architecture of the cerebral cortex, obtained by high definition magnetic resonance imaging (MRI); of the functional specialization of the cerebral cortex, obtained by functional MRI; of functional connectivity, obtained by functional MRI in rest state; and of the topographical organization of the white matter, obtained from the Hopf mylar architecture map.

The new area 55b is small and elongated (Figure 10.1), and is delimited by the frontal ocular (FOF) and premotor fields (POF), the primary motor cortex (PMC, or Brodmann area 4), the central premotor cortex (CpMC), area 6v based on Brodmann) and the prefrontal areas 8Av and 8C. Area 55b exhibits less myelination with respect to the surrounding areas and is selectively connected to posterior perisylvian zones (PPZ), deeply related to language processing.

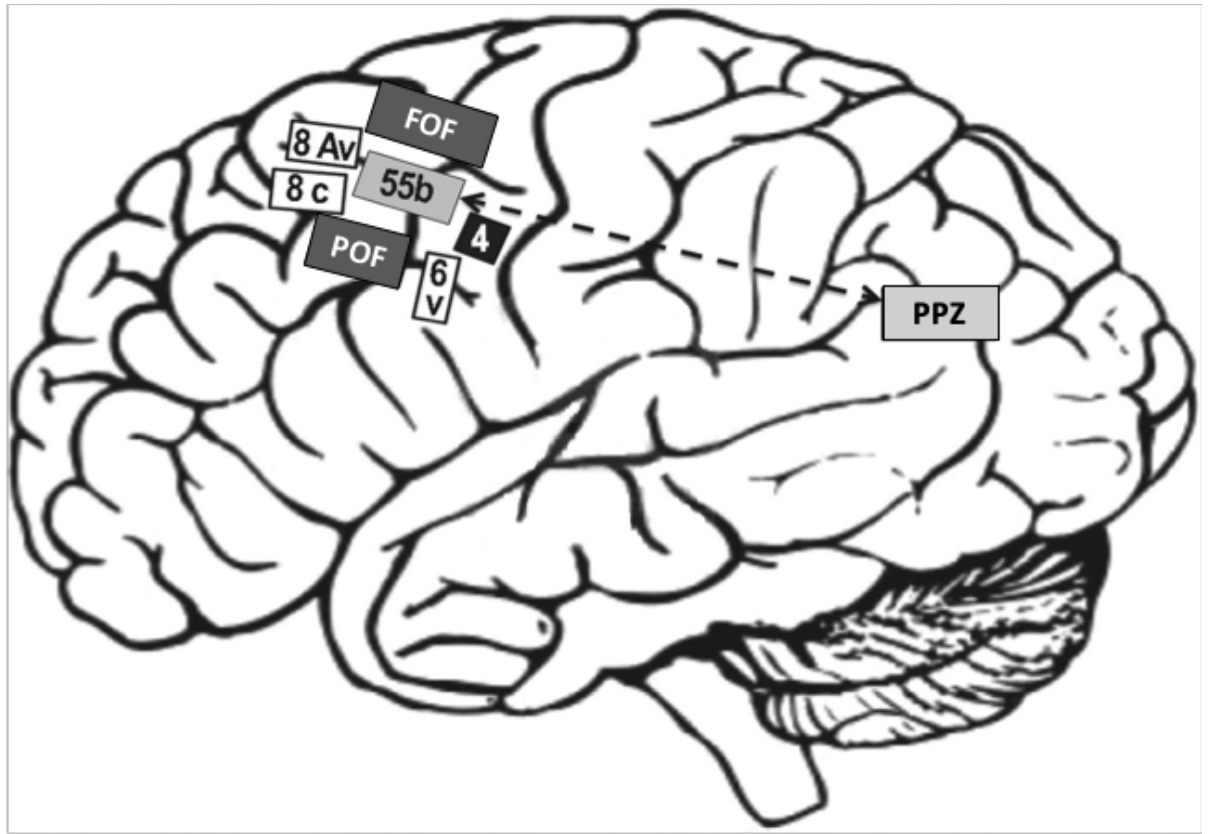


Figure 10.1. Diagram of the cortical area 55b inspired by Glasser et al., 2016.

10.7. Summary

Understanding and modeling language has been one of the main objectives of modern cognitive neurosciences. In recent decades, progress has been made in obtaining empirical data that have helped to model numerous specific aspects of language, such as perception and phoneme production. The current goal is to propose general models that explain in an integrated way the structure of the language in different stages of development, and in normal and pathological conditions, to obtain a morpho-functional maps of language to support neuropsychology.

Acknowledgments

Partially funded by Fondecyt 1060767 to M.P.; Doctoral Scholarship National CONICYT 21140640 to C.J.

VIII. OTHER ADDITIONAL OUTPUT DURING THE DOCTORAL PROGRAM

8.1. Publications

- Flores, J. C., Bohmwald, K., Espinoza, J., Jara, C., Peña, M., Hoyos-Bachilloglu, R., Iturriaga, C., Kalergis, A. M. & Borzutzky, A. (2016). Potenciales consecuencias neurocognitivas de infección por virus respiratorio sincicial humano. *Revista Chilena de infectología*, 33(5), 537-542.
- Peña, M., Jara, C., Flores, J.C., Hoyos-Bachilloglu, R., Iturriaga, C., Medina, M., Carcey, J., Bohmwald, K., Kalergis, A. M. & Borzutzky, A. (2018) Impaired language acquisition after severe respiratory disease caused by human respiratory syncytial virus during infancy (under review).

8.2. Book Chapters

- Cortés, P., Gutiérrez, C., Jara, C. & Peña, M. (2018) Neuroimagen y cognición. In Labos, E., Slachevsky, A., Fuentes, P. & Manes, F. (Eds.) *Tratado de Neuropsicología Clínica*.

8.3. Conference Posters presentations

- Jara, C. & Peña, M. (2016) “*Discovering Noun and Verb precursors during language acquisition*”. 9th *Embodied and Situated Language Processing Conference (ESLP)*. Pucón, Chile.
- Jara, C. & Peña, M. (2017) “*Discovering Noun and Verb precursors during language acquisition*.” *Workshop on Infants Language Development (WILD)*. Bilbao, Spain.

- Jara, C. & Peña, M. (2018) “*Discovering grammatical categories from audio-visual cues during early language acquisition*”. *21st Biennial International Congress of Infant Studies (ICIS)*. Philadelphia, USA.

8.4. Participation in other research projects

- Funded project VRI ‘Neurocognitive evaluation after severe respiratory syncytial virus infection in infants’ (2014 – 2016).
- Funded project VRI Puente ‘Exploring the role of rhythm on predictive reasoning during the first year of life’ (in progress).
- Funded project INTA ‘Efecto del modo de alimentación en el crecimiento y la función cognitiva infantil (2017).

REFERENCES

- Arias-Trejo, N., & Plunkett, K. (2010). The effects of perceptual similarity and category membership on early word-referent identification. *Journal of Experimental Child Psychology*, 105(1-2), 63-80.
- Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological review*, 105(1), 158.
- Bahrack, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, 13(3), 99-102
- Bahrack, L. E., Hernandez-Reif, M., & Flom, R. (2005). The development of infant learning about specific face-voice relations. *Developmental psychology*, 41(3), 541.
- Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum.
- Benavides-Varela, S., Hochmann, J. R., Macagno, F., Nespor, M., & Mehler, J. (2012). Newborn's brain activity signals the origin of word memories. *Proceedings of the National Academy of Sciences*, 201205413.
- Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of experimental psychology: General*, 117(1), 21.

Bonatti, L., Frot, E., Zangl, R., & Mehler, J. (2002). The human first hypothesis: Identification of conspecifics and individuation of objects in the young infant. *Cognitive psychology*, 44(4), 388-426.

Bouchon, C., Floccia, C., Fux, T., Adda-Decker, M., & Nazzi, T. (2015). Call me Alix, not Elix: Vowels are more important than consonants in own-name recognition at 5 months. *Developmental Science*, 18(4), 587-598.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, 121(1), 127-132.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, 110(28), 11278-11283.

Cassirer, B. (1923). *Philosophie der symbolischen Formen: Ernst Cassirer*. Translated as The Philosophy of Symbolic Forms. Vol. 1: Language (1955). New Haven: Yale University Press.

Cirelli, L. K., Spinelli, C., Nozaradan, S., & Trainor, L. J. (2016). Measuring Neural Entrainment to Beat and Meter in Infants: Effects of Music Background. *Frontiers in neuroscience*, 10, 229.

Chen, C. H., Gershkoff-Stowe, L., Wu, C. Y., Cheung, H., & Yu, C. (2017). Tracking multiple statistics: Simultaneous learning of object names and categories in English and Mandarin speakers. *Cognitive science*, 41(6), 1485-1509.

Chomsky, N. (1957). *Syntactic structures*. The Hague, The Netherlands: Mouton.

Cruttenden, A. (2014). *Gimson's pronunciation of English*. London & New York: Routledge.

Conboy, B. T., & Mills, D. L. (2006). Two languages, one developing brain: Event-related potentials to words in bilingual toddlers. *Developmental science*, 9(1), F1-F12.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, 13(4), 148-153.

Csibra, G., Hernik, M., Mascaró, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental psychology*, 52(4), 521.

Damasio, A. R., & Damasio, H. (1994). Cortical systems for retrieval of concrete knowledge: The convergence zone framework. In C. Koch and J.L. Davis (Eds.) *Large-scale neuronal theories of the brain*. (pp. 61-74). Cambridge: MIT Press.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, 208(4448), 1174-1176.

DeCasper, A. J., & Spence, M. J. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant behavior and Development*, 9(2), 133-150.

DeCasper, A. J., Lecanuet, J. P., Busnel, M. C., Granier-Deferre, C., & Maugeais, R. (1994). Fetal reactions to recurrent maternal speech. *Infant behavior and development*, 17(2), 159-164.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3), 283.

De Ribaupierre, A. (2015). Why should cognitive developmental psychology remember that individuals are different?. *Research in Human Development*, 12(3-4), 237-245.

Dronkers, N. F., Wilkins, D. P., Van Valin Jr, R. D., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1-2), 145-177.

Finley, S. (2011). The privileged status of locality in consonant harmony. *Journal of memory and language*, 65(1), 74-83.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, Mass.: MIT Press.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of experimental psychology: Human perception and performance*, 17(3), 816.

Fowler, C. A. (1995). Speech production. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition (2nd ed.)*, Vol. 11. *Speech, language, and communication* (pp. 29-61). San Diego: Academic Press.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, 110(2), 160-170.

Friederici, A. D., Meyer, M., & von Cramon, D. Y. (2000). Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. *Brain and language*, 74(2), 289-300.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, 6(2), 78-84.

Friedrich, M., & Friederici, A. D. (2017). The origins of word learning: Brain responses of 3-month-olds indicate their rapid association of objects and words. *Developmental science*, 20(2), e12357.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, 27-52.

Garrett, M. F. (1975). The analysis of sentence production. *The psychology of learning and motivation*, 9, 133-177.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences*, 105(37), 14222-14227.

Gibson, J.J. (1966). The senses considered as perceptual systems. Oxford, England: Houghton Mifflin.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, Ch., Jenkinson, M., Smith, S. M., & Van Essen, D.C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615), 171-178.

Gliga, T., Elsabbagh, M., Andravizou, A., & Johnson, M. (2009). Faces attract infants' attention in complex displays. *Infancy*, 14(5), 550-562.

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of experimental child psychology*, 69(2), 133-149.

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child development*, 71(4), 878-894.

Gogate, L. J., Bolzani, L. H., & Betancourt, E. A. (2006). Attention to maternal multimodal naming by 6-to 8-month-old infants and learning of word-object relations. *Infancy*, 9(3), 259-288.

Gogate, L. J., & Hollich, G. (2010). Invariance detection within an interactive system: A perceptual gateway to language development. *Psychological review*, 117(2), 496.

Gogate, L., & Hollich, G. (2016). Early verb-action and noun-object mapping across sensory modalities: A neuro-developmental view. *Developmental neuropsychology*, 41(5-8), 293-307.

Gogate, L., & Maganti, M. (2017). The origins of verb learning: Preverbal and postverbal infants' learning of word-action relations. *Journal of Speech, Language, and Hearing Research*, 60(12), 3538-3550.

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology (Vol. 1)*. Colchester, VT: Basil Blackwell.

Golinkoff, R. M., & Hirsh-Pasek, K. (2008). How toddlers begin to learn verbs. *Trends in cognitive sciences*, 12(10), 397-403.

Guellaï, B., Coulon, M., & Streri, A. (2011). The role of motion and speech in face recognition at birth. *Visual Cognition*, 19(9), 1212-1233.

Guenther, F. H., Perkell, J. S., Maassen, B., Kent, R. D., Peters, H. F. M., van Lieshout, P. H. H. M., & Hulstijn, W. (2004). A neural model of speech production and its application to studies of the role of auditory feedback in speech. *Speech motor control in normal and disordered speech*, 29-49.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences*, 4(4), 131-138.

Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory-motor interaction revealed by fMRI: speech, music, and working memory in area Spt. *Journal of cognitive neuroscience*, 15(5), 673-682.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2), 67-99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5), 393.

Hidalgo, A., & Quilis, M. (2012). La voz del lenguaje: Fonética y Fonología del español. Valencia: Tirant Humanidades.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, 97(5), 3099-3111.

Hockett, C. F., & Hockett, C. D. (1960). The origin of speech. *Scientific American*, 203(3), 88-97.

Hochmann, J. R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Consonants and vowels: different roles in early language acquisition. *Developmental science*, 14(6), 1445-1458.

Hochmann, J. R., Benavides-Varela, S., Fló, A., Nespor, M., & Mehler, J. (2017). Bias for Vocalic Over Consonantal Information in 6-Month-Olds. *Infancy*, 23(1), 136-151

Holdcroft, D. (1991) *Saussure: Signs, System, and Arbitrariness*. Cambridge, Mass.: Cambridge University Press.

Humphries, C., Willard, K., Buchsbaum, B., & Hickok, G. (2001). Role of anterior temporal cortex in auditory sentence comprehension: an fMRI study. *Neuroreport*, 12(8), 1749-1752.

Humphries, C., Love, T., Swinney, D., & Hickok, G. (2005). Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. *Human brain mapping*, 26(2), 128-138.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600), 453.

Hyde, D.C., Jones, B.L., Flom, R., Porter, C.L., 2011. Neural signatures of face-voice synchrony in 5-month-old human infants. *Dev. Psychobiol.* 53, 359–370.

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1-2), 101-144.

Jakobson, R. (1960). *Style in language*. In T.A. Sebeok (Ed.). Cambridge: MIT Press.

Johnson, K. (2005). Speaker Normalization in Speech Perception. In D.B. Pisoni and R.E. Remez (Eds.) *The Handbook of Speech Perception* (pp.363). Oxford: Blackwell Publishers.

Kabdebon C, Pena M, Buiatti M, Dehaene-Lambertz G. (2015). Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. *Brain Lang.* 148:25-36. doi: 10.1016/j.bandl.2015.03.005.

Keating, P. A. (2006). Phonetic encoding of prosodic structure. *Speech production: Models, phonetic processes, and techniques*, 167-186.

Kewley-Port, D., & Luce, P. A. (1984). Time-varying features of initial stop consonants in auditory running spectra: A first report. *Perception & psychophysics*, 35(4), 353-360.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, 83(2), B35-B42.

Kopp, F. (2014). Audiovisual temporal fusion in 6-month-old infants. *Developmental cognitive neuroscience*, 9, 56-67.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, 218(4577), 1138-1141.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044), 606-608.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11), 831.

Ladefoged, P. (1993). *A course in phonetics* (4th edition). New York: Hartcourt Brace Jovanovich.

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

Levelt, W. J. (1989). *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1), 1-38.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychological bulletin*, 126(2), 281.

Lewkowicz, D. J., & Pons, F. (2013). Recognition of amodal language identity emerges in infancy. *International Journal of Behavioral Development*, 37(2), 90-94.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5), 358.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1-36.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral cortex*, 15(10), 1621-1631.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and speech*, 10(1), 1-28.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4), 676.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3), 189-256.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8(1), 1-71.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive psychology*, 21(3), 398-421.

Matatyaho-Bullaro, D. J., Gogate, L., Mason, Z., Cadavid, S., & Abdel-Mottaleb, M. (2014). Type of object motion facilitates word mapping by preverbal infants. *Journal of experimental child psychology*, 118, 27-40

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, 18(1), 1-86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746.

Mehler, J. y Dupoux, E. (1990). *Naitre Humain*. Paris: Editions Odile Jacob.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J.L. Morgan & K. Demuth (Eds), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, (pp. 101-116) Lawrence Erlbaum Associates, Inc.

- Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29(2), 143- 178.
- Meyer, A. S. (1992). Investigation of phonological encoding through speech error analyses: Achievements, limitations, and alternatives. *Cognition*, 42(1-3), 181-211.
- Miller, B. T., & D'Esposito, M. (2005). Searching for “the top” in top-down control. *Neuron*, 48(4), 535-538.
- Mintz, T. H., Walker, R. L., Welday, A., & Kidd, C. (2018). Infants' sensitivity to vowel harmony and its role in segmenting speech. *Cognition*, 171, 95-107.
- Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant behavior and development*, 16(4), 495-500.
- Nazzi, T., Poltrock, S., & Von Holzen, K. (2016). The developmental origins of the consonant bias in lexical processing. *Current Directions in Psychological Science*, 25(4), 291-296.
- Nelson, C. A. (2001). The development and neural bases of face recognition. *Infant and Child Development: An International Journal of Research and Practice*, 10(1-2), 3-18.
- Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio*, 2(2), 203-230.
- Nygaard, L. C., & Pisoni, D. B. (1995). Speech perception: New directions in research and theory. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition*

(2nd ed.), Vol. 11. *Speech, language, and communication* (pp. 63-96). San Diego, CA, US: Academic Press.

Otsuka, Y., Konishi, Y., Kanazawa, S., Yamaguchi, M. K., Abdi, H., & O'Toole, A. J. (2009). Recognition of moving and static faces by young infants. *Child development*, 80(4), 1259-1271.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, 22(2), 237-247.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, 6(2), 191-196.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, Mass.: Harvard University Press.

Pons, F., & Toro, J. M. (2010). Structural generalizations over consonants and vowels in 11-month-old infants. *Cognition*, 116(3), 361-367.

Poltrock, S., & Nazzi, T. (2015). Consonant/vowel asymmetry in early word form recognition. *Journal of experimental child psychology*, 131, 135-148.

Price, C. J. (2000). The anatomy of language: contributions from functional neuroimaging. *Journal of anatomy*, 197(3), 335-359.

Price, C. J., Wise, R. J. S., Warburton, E. A., Moore, C. J., Howard, D., Patterson, K., Frackowiak, R.S.J., & Friston, K. J. (1996). Hearing and saying: The functional neuro-anatomy of auditory word processing. *Brain*, 119(3), 919-931.

- Putnam, S. P., Helbig, A. L., Gartstein, M. A., Rothbart, M. K., & Leerkes, E. (2014). Development and assessment of short and very short forms of the Infant Behavior Questionnaire–Revised. *Journal of personality assessment*, 96(4), 445-458.
- Rakison, D. H., & Yermolayeva, Y. (2010). Infant categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6), 894-905.
- Rauschecker, J. P. (1998). Cortical processing of complex sounds. *Current opinion in neurobiology*, 8(4), 516-521.
- Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, 64(3), 249-284.
- Rose, S., & Walker, R. (2011). Harmony systems. In J. Goldsmith, J. Riggle, & A. C. L. Yu (Eds.), *The handbook of phonological theory* (2nd ed., pp. 240–290). Malden, Massachusetts: Blackwell.
- Roseberry, S., Richie, R., Hirsh-Pasek, K., Golinkoff, R. M., & Shipley, T. F. (2011). Babies catch a break: 7-to 9-month-olds track statistical probabilities in continuous dynamic events. *Psychological Science*, 22(11), 1422-1424.
- Rumelhart, D. E., McClelland, J. L. (1986). *Parallel and Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1. *Foundations*. Cambridge, Mass.: MIT Press.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926-1928.

Scott, S. K., & Wise, R. J. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, 92(1-2), 13-45.

Scott, S. K. (2005). Auditory processing—speech, space and auditory objects. *Current opinion in neurobiology*, 15(2), 197-201.

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology*, 18(9), 668-671.

Shaw, K. E., & Bortfeld, H. (2015). Sources of confusion in infant audiovisual speech perception research. *Frontiers in psychology*, 6, 1844.

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*, 108(15), 6038-6043

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word–referent learning. *Trends in cognitive sciences*, 18(5), 251-258.

Squire, L. R. (1987). *Memory and brain*. New York: Oxford University Press.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4), 1872-1891.

Strange, W. (1999). Perception of vowels: Dynamic constancy. In J.M. Pickett, *The acoustics of speech communication, fundamentals, speech perception theory, and technology* (pp. 153-165). Needham Heights (MA): Allyn & Bacon.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, 79(4), 1086-1100.

Taraban, R., & McClelland, J. L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of memory and language*, 27(6), 597-632.

Thiessen, E. D. (2010). Effects of visual information on adults' and infants' auditory statistical learning. *Cognitive Science*, 34(6), 1093-1106.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, 97, B25-B34.

Toro, J. M., Nespor, M., Mehler, J., & Bonatti, L. L. (2008). Finding words and rules in a speech stream functional differences between vowels and consonants. *Psychological Science*, 19(2), 137-144.

Toro, J. M., Shukla, M., Nespor, M., & Endress, A. D. (2008). The quest for generalizations over consonants: Asymmetries between consonants and vowels are not the by-product of acoustic differences. *Perception & psychophysics*, 70(8), 1515-1525.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, 134, 552-564. doi:10.1037/0096-3445.134.4.552

Valenza, E., Simion, F., Cassia, V. M., & Umiltà, C. (1996). Face preference at birth. *Journal of experimental psychology: Human Perception and Performance*, 22(4), 892.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167(3917), 392-393.

Waxman, S., Fu, X., Arunachalam, S., Leddon, E., Geraghty, K., & Song, H. J. (2013). Are nouns learned before verbs? Infants provide insight into a long-standing debate. *Child development perspectives*, 7(3), 155-159.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1), 49-63.

Wernicke, C. (1874). *Der aphasische Symptomencomplex*. Breslau (Poland): Cohen and Weigert.

Wu, R., & Kirkham, N. Z. (2010). No two cues are alike: Depth of learning during infancy is dependent on what orients attention. *Journal of experimental child psychology*, 107(2), 118-136.