Separation and acquisition of two languages in early childhood : a multidisciplinary approach

Maria Julia Carbajal

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'Ecole Normale Supérieure

Separation and acquisition of two languages in early childhood: A multidisciplinary approach

Séparation et acquisition de deux langues chez le jeune enfant : une approche pluridisciplinaire

Ecole doctorale n°158

CERVEAU, COGNITION, COMPORTEMENT

Spécialité SCIENCES COGNITIVES

COMPOSITION DU JURY :

Mme. ADDA-DECKER Martine
Université Sorbonne Nouvelle, France
Rapporteur, Président du jury

Mme. BOSCH Laura
Universitat de Barcelona, Espagne
Rapporteur

Mme. SKORUPPA Katrin
University of Neuchâtel, Suisse
Examinateur

M. DUPOUX Emmanuel
EHESS, France
Co-Directeur de thèse

Mme. PEPERKAMP Sharon
Ecole Normale Supérieure, France
Co-Directeur de thèse

Soutenue par Maria Julia CARBAJAL
le 21 septembre 2018

Dirigée par **Emmanuel DUPOUX**
et Sharon PEPERKAMP

ENS PSL RESEARCH UNIVERSITY PARIS

Ecole Doctorale 158 *Cerveau, Cognition, Comportement*
Université Paris Sciences et Lettres
École Normale Supérieure

Ph.D. Thesis

Separation and acquisition of two languages in early childhood: A multidisciplinary approach

Presented by
Maria Julia Carbajal

Under the direction of
Emmanuel Dupoux and Sharon Peperkamp

September 2018

Jury composition:

Dr. Martine Adda-Decker (Rapporteur & President of jury)
Dr. Laura Bosch (Rapporteur)
Dr. Katrin Skoruppa (Examiner)
Dr. Emmanuel Dupoux (Thesis supervisor)
Dr. Sharon Peperkamp (Thesis supervisor)

Laboratoire de Sciences Cognitives et Psycholinguistique
UMR 8554 CNRS-ENS-EHESS

Abstract

During the first years of life, children rapidly learn to process speech from a continuous acoustic signal, and soon become able to understand and produce the sounds, words and structure of their native language. Children growing up in a bilingual environment face an additional challenge: they must simultaneously discover and separate their bilingual input into individual (yet potentially overlapping) systems, with independent sound units, vocabularies and grammars, without knowing a priori how many languages are spoken in their environment. In spite of this, language acquisition in young bilinguals follows, to an extent, a similar time-line as in monolinguals. Understanding how children come to discover the presence of two languages in their input, and to what extent they are able to keep them apart, are to this day crucial questions to the field of childhood bilingualism. In this thesis we focus on these two questions by exploring how perceptual and environmental properties of the input can help or hinder the discovery and lexical development of two languages, and whether the phonological representations formed by young bilinguals are language-specific. In order to investigate these questions, we take a multidisciplinary approach, using both empirical and computational techniques, which can provide different insights on the task of early language separation.

In the first part of this dissertation we examine the problem of discovering two languages in the input from an acoustic perspective. Based on a large body of research on language discrimination abilities in newborns and infants, and inspired by previous modelling work, we aim to provide a computational account of infant perception of multilingual speech. Borrowing a state-of-the-art system from speech technologies, we conducted a series of computational experiments that can help us understand what kind of representations young infants form when hearing different languages, and how different factors may shape their perception of language distance.

In the second part, we investigate several environmental aspects of bilingual exposure. Previous research on quantitative and qualitative properties of bilingual input had shown strong influences of each language's relative amount of exposure on infants' lexical development, but diverging results were reported regarding the impact of the separation of the two languages in their environment. We used a home diary method to investigate the co-existence of two languages in young bilinguals' input, and explore how this and other environmental factors may influence their vocabulary acquisition.

Finally, in the last part of this dissertation, we consider bilingual preschoolers' perception of language-specific phonological rules. Unlike other properties of young bilinguals' phonological systems, their acquisition and separation of phonological rules has barely been explored, with the only prior evidence coming from production studies. We conducted a behavioral experiment using a touchpad videogame

to test French-English bilinguals' cross-linguistic perception of phonological assimilations.

Overall, this thesis contributes new insights to the question of language separation and acquisition in early bilingualism, with multiple perspectives for future research on this topic.

Résumé

Durant les premières années de leur vie, les enfants apprennent rapidement à traiter la parole à partir d'un signal acoustique continue, et très vite, ils sont capables de comprendre et de produire les sons, les mots et la structure de leur langue maternelle. Les enfants qui grandissent dans un environnement bilingue rencontrent un défi supplémentaire : ils doivent simultanément découvrir et séparer les deux langues en deux systèmes individuels (qui peuvent cependant se chevaucher), avec des unités sonores, des vocabulaires et des grammaires indépendantes, sans savoir a priori combien de langues sont parlées dans leur environnement. Malgré cela, l'acquisition du langage chez les jeunes bilingues suit, dans une large mesure, une chronologie similaire à celle des enfants monolingues. Comprendre comment les enfants arrivent à découvrir la présence de deux langues dans ce qu'ils entendent, et dans quelle mesure ils arrivent à les séparer, sont des questions cruciales pour le domaine de la recherche en bilinguisme chez les enfants. Dans cette thèse, nous nous concentrons sur ces deux questions en explorant comment les propriétés perceptuels et environnementales de ce qu'ils entendent peuvent aider ou ralentir la découverte et le développement lexical des deux langages, et si les représentations phonologiques formées par les jeunes bilingues sont spécifiques à chaque langue. Afin d'étudier ces questions, nous adoptons une approche multidisciplinaire, en utilisant à la fois des techniques empiriques et computationnelles, qui permettent d'apporter différents éclaircissements sur la tâche de la séparation des langues à un jeune âge.

Dans la première partie de cette thèse, nous nous intéressons au problème de la découverte de deux langues dans ce qu'entend l'enfant d'un point de vue acoustique. Basé sur des recherches existantes concernant les capacités de discriminations des langues chez les nouveau-nés et les enfants, et inspirés par de précédents travaux de modélisation, nous cherchons à décrire d'un point de vue computationnel la perception de la parole dans plusieurs langues chez l'enfant. En empruntant un système de l'état de l'art dans les technologies de la parole, nous avons mené une série d'expériences qui peuvent servir à comprendre quel type de représentations les jeunes enfants créent quand ils entendent des langues différentes, et comment différents facteurs peuvent influencer leur perception de la distance entre les langues.

Dans la deuxième partie, nous étudions plusieurs aspects environnementaux de l'exposition à deux langues. Des recherches précédentes sur les propriétés quantitatives et qualitatives de l'input bilingue a montré des influences fortes de la quantité relative de l'exposition à chaque langue sur le développement lexical de l'enfant, mais des résultats divergents ont été rapportés quant à l'influence de la séparation des deux langues dans leur environnement. Nous avons utilisé une méthode de journal que les parents tenaient chez eux afin d'étudier la co-existence de deux langues dans ce qu'entendent les jeunes

bilingues, et d'explorer comment cela, ainsi que d'autres facteurs environnementaux, peuvent influencer l'acquisition du vocabulaire.

Finalement, dans la dernière partie de cette thèse, nous examinons la perception de règles phonologiques spécifiques à chaque langue chez les enfants à l'âge de la maternelle. Contrairement à d'autre propriétés des systèmes phonologiques chez les jeunes bilingues, leur acquisition et séparation de différentes règles phonologiques ont été très peu explorées, et les seules données antérieures proviennent d'études sur la production. Nous avons mené une expérience comportementale en utilisant un jeu vidéo, afin de mesurer la perception inter-langue de l'assimilation phonologique chez les bilingues français-anglais.

Dans son ensemble, cette thèse contribue en apportant de nouveaux éclaircissements sur la question de la séparation et l'acquisition des langues dans le bilinguisme précoce, avec de nombreuses perspectives pour des recherches futures sur le sujet.

Publications

Carbajal, M.J., Fér, R. and Dupoux, E. (2016). Modeling language discrimination in infants using i-vector representations. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Philadelphia, USA.

Carbajal, M.J., Dawud, A., Thiollière, R. and Dupoux, E. (2016). The "language filter" hypothesis: A feasibility study of language separation in infancy using unsupervised clustering of I-vectors. In *Proceedings of the 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, Cergy-Pontoise, France, 195-201.

Carbajal, M.J., Chartofylaka, L., Hamilton, M., Fiévet, A.C. and Peperkamp, S. (*in revision*) Compensation for phonological assimilation in bilingual children.

Carbajal, M.J. and Peperkamp, S. (*submitted*) Dual language input and the impact of language separation on early lexical development.

Acknowledgements

This thesis would not have been possible without the help and support from many, many people. I hope in these lines I will be able to transmit how grateful I am for the amazing colleagues, family and friends that have accompanied me on this adventure.

Firstly, I would like to express my gratitude to my supervisors, Sharon and Emmanuel, who have guided me and motivated me in this learning process. Working with you has been a great honor and an enriching experience, both scientifically and personally. Throughout my thesis you have shared with me your immense knowledge, you have encouraged me to develop my own ideas, and you have helped me find a path whenever I was lost. Your support and patience have given me the courage to always give my best whatever the circumstances. I thank you both from the bottom of my heart.

Next, I owe a huge thank-you to the people who made my projects possible: Anne-Caroline Fiévet and the Babylab team (Luce, Audrey and Chase), as well as the most wonderful support team one could ever have: Michel Dutat, Vireack Ul, Radhia Accheb and Isabelle Brunet. Furthermore, my modelling work would not have been possible without the immense help of the CoML team, and in particular the fantastic engineers Rolland Thiollière, Mathieu Bernard, Juan Benjumea and Julien Karadayi. I also thank the team in Brno (Radek, Hynek, Lukas, Honza) for hosting me and helping me give my first steps in i-vector modelling, and of course the numerous families that participated in my studies.

I also thank all the researchers and team members of the LSCP for creating such a friendly and inspiring work atmosphere, with a very special thanks to my tutor Alex Cristia, who offered great advice and extra encouragement when needed during these four years.

I would like to express my gratitude to the Ecole des Neurosciences de Paris–Ile-de-France (ENP) for funding this thesis. I also wish to sincerely thank the members of my jury for accepting to read and critique my work: Martine Adda-Decker, Laura Bosch and Katrin Skoruppa.

Last but not least, I would like to thank my family and friends, without whom I would not have made it this far:

Mamá, Diana, Pedro y Mario: A ustedes les debo todo. Gracias por creer en mí, por bancar mis llantos y quejas, por alentarme a seguir cuando quería bajar los brazos, por hacer el esfuerzo de venir a visitarme todos los años. Los quiero tanto! Gracias miles.

John: I don't have enough words to thank you for all your love and support all these years, I couldn't have made it without you! Thank you for believing in me and supporting me through good and bad.

Adriana, Naomi and Page: You have been the most amazing friends I could ever ask for, and kept me sane through this thesis. I thank you all for your friendship, for your support, for your amazing advice, for your hugs and smiles, for your encouragement, for everything.

And many many other friends, family and colleagues that have accompanied me during these four years and made my PhD life more enjoyable: Alex & Alex, Ahmad, Andrea, Antu, Camila, Christina, Eric, Ewan, Gerda, Graciela, Hannah, Hayat, Hernán, Horacio, Laia, Lamprini, Mireille, Mollie, Monica, Mora, Pame, Parvaneh, Pau, Poli, Rory, Rahma, Sho, Susana, Tommy... Thank you all!

Contents

Chapter 1

Introduction

Human spoken language is a highly complex cognitive ability that develops at a remarkable speed. During the first years of life, children rapidly learn to process speech from a continuous acoustic signal, and soon become able to understand and produce the sounds, words and structure of their native language. Children growing up in a bilingual environment face an additional challenge: they must simultaneously discover and separate their bilingual input into individual (yet potentially overlapping) systems, with independent sound units, vocabularies and grammars, without knowing a priori how many languages are spoken in their environment. In spite of this, language acquisition in young bilinguals follows, to an extent, a similar time-line as in monolinguals (Werker, Byers-Heinlein, & Fennell, 2009). Understanding how children come to discover the presence of two languages in their input, and to what extent they are able to keep them apart, are to this day crucial questions to the field of childhood bilingualism. In this thesis we focus on these two questions by exploring, on the one hand, how perceptual and environmental properties of the input can help or hinder the discovery and acquisition of two languages, and, on the other hand, whether the phonological representations formed by young bilinguals are language-specific.

In this first chapter we will frame the problem of language separation in early childhood, i.e., from birth through pre-school years. We will first describe early bilingualism and provide an overview of the challenges young bilingual children face. Next, we will review previous research that has aimed at understanding how children discriminate and learn two languages, with a focus on their phonological and lexical development. Finally, we will outline the research questions of this thesis.

1.1 Defining bilingualism in early childhood

When describing bilingualism in adults, multiple definitions can be adopted, usually focusing on the degree to which both languages are mastered and actively used by the individual. In early childhood, and particularly during the first two years of life, classifying children as bilinguals or monolinguals using similar criteria is not possible, as their language abilities are still under development and their capacities in one or both languages may not be evident until later in life. Definitions of early childhood bilingualism thus put more weight on the input than on the output. Two main factors defining bilingualism in the first few years of life are the age of acquisition and the amount of exposure to each language. In pre-schoolers, however, some researchers may opt to classify children based on whether they produce words or sentences in both languages. In this thesis, we will focus principally on bilinguals' perceptual systems, and thus will base our classification on their input and not on their productive output.

Regarding the amount of exposure, no consensus exists on how much input in each language is required for a child to be bilingual. However, many studies used criteria based on the proportion of exposure in each language, often defining a minimum of about 20% to 30% and a maximum of 70% to 80% of the child's total input (see for example Fennell, Byers-Heinlein & Werker, 2007, Pearson et al., 1997). In general, children hearing a second language comprising less than 10% - 20% of their input are unlikely to develop communicative skills in that language, and are thus typically classified as monolinguals (Pearson et al., 1997). These percentages are usually obtained through parental questionnaires, either directly asking parents to estimate how much of each language the child has heard, or by asking detailed questions regarding the amount of hours per week that the children heard each language throughout their life.

Moreover, age of acquisition can be used to make a distinction between *simultaneous* bilinguals, that is, those who began their dual language exposure early in life (also often referred to as *bilingual first language –BFL– learners*), and *sequential* bilinguals, i.e., those for whom exposure to a second language began some time after the first language. However, where the dividing threshold lies is not clear. For instance, De Houwer (2009) defined Bilingual First Language Acquisition (BFLA) as "the development of language in young children who hear two languages spoken to them from birth", while other researchers proposed later thresholds, some as late as 2-3 years old (e.g., McLaughlin 1978). In this thesis, we will take an intermediate position, by considering BFLA as dual language exposure

that began within the first 6 months of life, that is, while infants' phonological systems are still very much under development. However, this definition is not motivated by a strict developmental cut-off. Indeed, language development is a continuum and children starting their bilingual exposure after the age of 6 months (and even beyond their first birthday) may achieve good language skills in both languages. The reason for this cut-off is simply to focus on the challenges that children face when encountering two languages from the earliest stages of development. In this context, we will refer to BFL learners as bilinguals, even if they do not (yet) talk in one or both of their languages.

1.1.1 The challenge of bilingual first language acquisition

What does growing up in a dual language environment entail? The first and most obvious challenge is that bilingual learners have twice the amount of knowledge to acquire, while their input in each language is likely to be smaller than what a typical monolingual receives. But the problem is not restricted to a simple ratio between available input and target output. Indeed, speech is an inherently variable signal, and acquiring one's native language involves learning which dimensions of variability encode meaning, and which do not. Infants growing up in a bilingual environment will inevitably encounter additional variability across all linguistic (and extra-linguistic) dimensions, with varying degrees of overlap between their two languages. To understand the difficulty of the task, let's look at some examples at two linguistic levels, the acoustic-phonetic and lexical-semantic levels:

At the *acoustic-phonetic* level, young bilinguals must deal with a complex combination of differences and similarities between the phonological properties of each of their two languages. A given property A that exists in one language may not be present in the other language. For example, a child learning English and Mandarin will have to learn that tone is not a lexical cue for the former, but it is for the latter. Alternatively, a property α may exist in the other language, which could be fully or partially overlapping with property A, or even lie in between two categories, A and B. A well known example of this is the case of the vowel contrast /e/ - /ɛ/, which exists in Catalan but not in Spanish, whose single /e/ category lies somewhere in between. Thus, Spanish-Catalan bilinguals will encounter conflicting information regarding these phonemic categories. Another difficulty that bilingual infants may find at the acoustic-phonetic level is the exposure to accented speech, which may contain a transfer of phonological properties of one language to the other, or simply add variability to the input.

At the *lexical-semantic* level, bilingual infants will encounter several challenges. For instance, they

must deal with the presence of multiple labels for the same referent. This means that some word learning strategies, such as mutual exclusivity, may be inefficient in a bilingual context. Furthermore, depending on the pair of languages they are exposed to, they may encounter a varying number of cognates, that is, words across languages that share a common root and thus may sound similar. For example, the English word *flower* [flaʊə], has cognates in several languages, such as French (*fleur* [flœʁ]), Spanish (*flor* [floɾ]) and Scottish Gaelic (*flùr* [flʸuːr]). The presence of cognates may facilitate word learning, but they could potentially result in underspecified word forms. Finally, bilinguals may encounter cases in which the semantic scope of a word in one language does not completely overlap with that of its translation equivalent. For example, the common English word *cup* may be used to refer to a porcelain tea cup with a handle, as well as a plastic cup with no handle, among many other things. Other languages may use different labels for these objects. Such is the case of French, where the former would be called a *tasse* while the latter would be called a *gobelet*. These partially overlapping word-referent pairings may create ambiguities which could pose a problem for lexical development.

Beyond linguistic features, bilingual infants and children are also exposed to a variety of contextual scenarios for each language. For instance, some children grow up in bilingual communities, where both languages can be found frequently in every daily-life situation, while others grow up hearing a minority language in a largely monolingual community. Moreover, both languages may be spoken at home, or otherwise one single language may be spoken at home, while the other language is heard elsewhere. Finally, some bilingual families may choose to adopt a one-person one-language approach, while others may choose to speak both languages freely. The contexts in which each language is encountered are thus an additional level of complexity that young bilinguals will have to deal with.

In summary, the combination of two languages will result in an intricate puzzle for the bilingual child to decipher. In order to succeed in this challenge, young bilinguals must eventually *separate* their two languages, correctly linking together the phonological, lexical, and syntactic elements that belong to one or the other language. When and how children achieve this feat, and to what extent their language systems are kept separated, have long been debated (for a review on the "one language or two" debate in early bilinguals see Byers-Heinlein, 2014). In the following sections we will review previous research on the emergence of language discrimination in young bilinguals, and present an overview of their phonological and lexical development.

1.2 Early language discrimination

It has been argued that early language discrimination is an essential step to successful bilingual language acquisition (Bosch & Sebastián-Gallés, 1997; Nazzi, Bertoncini, & Mehler, 1998; Werker et al., 2009). Indeed, if throughout their development infants failed to ever notice the presence of two languages in their input, they could end up learning a hybrid language composed of a mix of their linguistic properties. For a long time, this was thought to be the case, leading some researchers to propose that young bilinguals had a unified language system for the first two or three years of life, and that languages were only gradually discriminated later on (Leopold, 1970; Redlinger & Park, 1980; Volterra & Taeschner, 1978). However, an increasing body of evidence has emerged over the past 30 years, drawing a different picture; while young bilingual children may sometimes mix both languages in their production, they are generally capable of discerning them and learning distinct properties of each language (Genesee, 1989; Hammer et al., 2014; Werker & Byers-Heinlein, 2008). Exactly how young bilinguals discover their two languages is not yet fully understood, but experimental evidence suggests that some perceptual sensitivities available from early on may help them to solve this task.

Humans are born with great sensitivity to acoustic properties of speech (for a review, see Gervain and Werker, 2008). Mehler et al. (1988) were the first to investigate whether some of these early perceptual abilities would allow newborn infants to detect differences between languages. Using a high-amplitude sucking procedure (Jusczyk, 1985), they habituated French newborns to either French or Russian phrases spoken by a proficient bilingual. After habituation, half of the infants heard new utterances spoken in the same language (control condition), and the other half heard utterances spoken by the same speaker in the other language (switch condition). Their results show an increase in newborns' attention during the test phase (as indicated by a rise in their sucking rate) which was significantly larger after a language switch than in the control condition, suggesting that they successfully detected the language change. Additionally, they observed an asymmetry in their dishabituation patterns indicating a preference for their native language, French. In a subsequent experiment, the researchers exposed newborns to a low-pass filtered version of the same utterances, showing again evidence of discrimination (Mehler et al., 1988). These results suggested for the first time that prosodic information, such as rhythm and intonation, are salient properties that newborn infants can use to discriminate between languages.

A number of follow-up studies using various language pairs, experimental paradigms and speech ma-

nipulations (e.g., natural, filtered, or resynthesized speech) have since replicated and extended these original results (Byers-Heinlein, Burns, & Werker, 2010; Mehler & Christophe, 1995; Mehler et al., 1988; Moon, Cooper, & Fifer, 1993; Nazzi et al., 1998; Ramus, 2002b). A summary of these studies is shown in Table 1.1. Taken together, the available evidence indicates that newborns, regardless of the language heard in the womb, are generally able to discriminate language pairs that differ in their rhythmic properties, such as French and Russian, while they fail at discriminating close language pairs, such as English and Dutch, or Spanish and Catalan. Furthermore, the discrimination of distant language pairs remains intact when using filtered speech or resynthesized speech in which only prosody and some phonotactic properties are preserved. This ability has also been observed in newborns whose mothers spoke two languages during gestation (Byers-Heinlein et al., 2010).

Table 1.1: Summary of studies on language discrimination in newborns.

Study	Native lang.	Test pair	R. Contrast	Stimuli type	Test type	Results
Mehler et al. (1988)	French	French-Russian	Across	Natural	Habituation	Discrim.
Mehler et al. (1988)	Mixed	French-Russian	Across	Natural	Habituation	No discrim.
Mehler et al. (1988)	French	English-Italian	Across	Natural	Habituation	Discrim.(1)
Ramus (2002)	French	Dutch-Japanese	Across	Natural	Habituation	No discrim.
Moon et al. (1993)	English, Spanish	English-Spanish	Across	Natural	Listening pref.	Native pref.
Mehler et al. (1988)	French	French-Russian	Across	Filtered	Habituation	Discrim.
Nazzi et al. (1998)	French	English-Japanese	Across	Filtered	Habituation	Discrim.
Byers-Heinlein et al. (2010)	English	English-Tagalog	Across	Filtered	Habituation	Discrim.
Byers-Heinlein et al. (2010)	English-Tagalog	English-Tagalog	Across	Filtered	Habituation	Discrim.
Byers-Heinlein et al. (2010)	English	English-Tagalog	Across	Filtered	Listening pref.	Native pref.
Byers-Heinlein et al. (2010)	English-Tagalog	English-Tagalog	Across	Filtered	Listening pref.	No pref.
Byers-Heinlein et al. (2010)	English-Chinese	English-Tagalog	Across	Filtered	Listening pref.	No pref.
Ramus (2002)	French	Dutch-Japanese	Across	Resynthesized(2)	Habituation	Discrim.
Ramus (2002)	French	Dutch-Japanese	Across	Resynthesized(3)	Habituation	No discrim.
Ramus (2002)	French	Dutch-Japanese	Across	Resynthesized(4)	Habituation	Discrim.
Mehler et al. (1988)	French	French-Russian	Across	Inverted	Habituation	No discrim.
Nazzi et al. (1998)	French	English-Dutch	Within	Filtered	Habituation	No discrim.
Ramus (unpublished)(5)	French	Spanish-Catalan	Within	Resynthesized	Habituation	No discrim.
Nazzi et al. (1998)	French	English+Dutch vs. Italian+Spanish	Mixed Across	Filtered	Habituation	Discrim.
Nazzi et al. (1998)	French	English+Italian vs. Dutch+Spanish	Mixed Within	Filtered	Habituation	No discrim.

Abbreviations: *Native lang.*: Language spoken by the mother during gestation. *R. Contrast*: Rhythmic contrast according to the rhythmic class hypothesis. *Discrim.*: Discrimination. *Native pref.*: Preference for native language. *No pref.*: Equal attention to both languages.

Notes: (1) As reanalysed by Mehler and Christophe (1995). (2) Saltanaj speech with original prosody. (3) Sasasa speech with artificial prosody. (4) Saltanaj speech with artificial prosody. (5) As reported in Ramus, Nespor & Mehler (1999).

These results have been interpreted as evidence for an innate rhythmic-based discrimination of languages in agreement with some theories of language rhythm typology (Ladefoged, 1975; Ramus, Nespor, & Mehler, 1999). According to the rhythmic class hypothesis, languages can be classified into three

main groups depending on their characteristic rhythmic patterns: syllable-timed (including many Romance languages, such as French, Spanish and Italian), stress-timed (including Germanic languages such as English, Dutch and German) and mora-timed (including languages such as Japanese and Tamil). Indeed, experimental data from newborns seems to reflect a rhythmic classification. However, it is not entirely clear whether we can conclude from these experiments that infants use rhythm to classify the languages in their input, for several reasons. First, (to the best of our knowledge) the only evidence showing that newborns fail at discriminating within-class language pairs comes from studies that have used filtered or resynthesized speech (indeed, these studies were meant to test newborns' sensitivity to rhythmic properties); it thus remains to be proven that they cannot discriminate them when hearing natural, unfiltered speech. Second, the fact that infants can discriminate certain language pairs when hearing filtered speech does not necessarily mean that rhythm is sufficient for language discrimination. Ramus (2002b) made an interesting observation when testing newborns on their discrimination of Dutch and Japanese, two languages belonging to distinct rhythmic classes. When presented with unfiltered speech from multiple monolingual speakers, infants failed to detect the language change, while they succeeded when presented with resynthesized speech. Ramus argued that the presence of multiple speakers - in contrast with a single bilingual speaker as in Mehler et al.'s (1988) study - introduced additional variability to the signal, hiding the relevant contrast. This experiment suggests that variability in other acoustic dimensions besides prosody may compete for infants' attention, eventually blurring otherwise salient distinctions.

This leads us to a third point, which is the fact that most experiments so far have controlled in one form or another the amount of variability in the experimental stimuli, either by using a bilingual speaker, by filtering or resynthesizing the signal, or by simply selecting utterances that were matched in length and number of syllables. These types of manipulations are common in infant experimental research, and are necessary to isolate the factors of interest. However, these cases might end up being far removed from the real-world situations in which infants are immersed, where speech comes from several speakers, some of which may be monolinguals and some not, where phrases come in all shapes and sizes, and where speakers' mood, speech register and even ambient noise may affect the way language sounds to the infant's ear. Further research is thus needed to understand how language and speaker distance, as well as other sources of variability, are reflected in early perceptual representations of speech. All that being said, the studies mentioned above have shown strong evidence that newborns possess sensitivity to rhythmic properties of languages, which might be used as a stepping stone for language discrimination.

Further research with infants aged 2- to 5-months-old has shown that, as infants grow up and become tuned to their native language, discrimination becomes more sensitive but at the same time dependent on infants' familiarity with one (or both) of the languages (Bahrick & Pickens, 1988; Bosch & Sebastián-Gallés, 1997, 2001; Christophe & Morton, 1998; Dehaene-Lambertz & Houston, 1998; Mehler et al., 1988; Molnar, Gervain, & Carreiras, 2014; Nazzi, Jusczyk, & Johnson, 2000). That is, infants may fail at discriminating a language pair, even across rhythmic class, if they are unfamiliar with both languages. For instance, American 2-month-olds fail at discriminating French from Russian, but succeed at discriminating English from Italian (Mehler et al., 1988). On the other hand, infants newly developed sensitivity to their native language allows them to discriminate it from other languages within the same rhythmic class, an ability that was not observed in newborns (Bosch & Sebastián-Gallés, 2001; Nazzi et al., 2000). For example, Bosch and Sebastián-Gallés (2001) showed that Spanish and Catalan 4-month-old monolinguals could discriminate these two languages. Most interestingly, they found that Spanish-Catalan bilinguals also succeed in discriminating their two languages. Thus, before the end of their first semester, infants seem to have accumulated sufficient knowledge about their native language(s) to allow discrimination, with bilinguals being no exception. This distinction might emerge naturally as young bilinguals learn the regularities of their input. As pointed out by Byers-Heinlein (2014), "English words are likely to be adjacent to other English words. They are composed of English sounds, follow English phonotactic rules, take English morphological endings, and are heard in sentences with English prosody". These regularities could thus eventually produce clusters in the infants' mental representations. This idea is supported by PRIMIR (Curtin, Byers-Heinlein, & Werker, 2011; Werker & Curtin, 2005), a theoretical framework that describes how infants (both monolinguals and bilinguals) learn language from their input using a combination of perceptual biases, general learning mechanisms and directed attention. However, no explicit account of how this process may unfold has yet been proposed.

Besides acoustic and linguistic cues to language discrimination, infants may additionally be sensitive to other sources of information available in their environment. One such cue is visual information (Sebastián-Gallés, Albareda-Castellot, Weikum, & Werker, 2012; Weikum et al., 2007). Weikum et al. (2007) studied infants' ability to discriminate languages by the facial gestures produced while talking. Young monolingual English and bilingual French-English infants were exposed to silent videos of bilingual speakers talking in one of these two languages. After habituation, infants saw new videos of the same speakers either talking in the same language as in habituation, or talking in the other language. They found that while monolinguals discriminated the languages at 4 and 6 months old,

they failed at 8 months old. On the other hand, bilingual infants still succeeded in the task at the latest age, indicating that they had retained their sensitivity to this cue for longer than their monolingual peers. Similar results were observed in Spanish, Catalan and Spanish-Catalan 8-month-olds tested on the same French and English stimuli (Sebastián-Gallés et al., 2012), suggesting that bilingual' ability to discriminate languages visually did not depend exclusively on familiarity with one of the two languages. Like facial gestures, bilingual infants might be able to detect and exploit many other sources of information that happen to co-occur with each language, such as contextual and social cues. For instance, if speakers in the bilingual's environment adopt a one-person-one-language approach, infants may be able to detect this regularity. We will further discuss the role of environmental cues in bilingual language acquisition in Chapters 2 and 3.

1.3 Phonological development

It is by now well established that babies are born with a language-general sensitivity to speech sounds. Evidence from a vast amount of studies on monolingual infants shows that, throughout their first year of life, infants' perception becomes tuned to the sounds of their native language. That is, they gradually perfect their ability to recognise the phonemes (i.e., the sound categories) that are used to distinguish words, while they simultaneously lose sensitivity to contrasts that are unused in their language (for a review, see Maurer and Werker, 2014). Vowel contrasts begin to stabilize around 6 months of age, while consonants take slightly longer, with most contrasts being in place by the end of the first year of life. Phonological development not only involves learning the sounds of one's native language, but also the rules that dictate how these sounds can be combined to form syllables and words – i.e., its *phonotactics* – and how the pronunciation of certain sounds or words may change depending on the phonological context. In monolinguals, sensitivity to the phonotactic rules of their native language emerges around the age of 9 months (Jusczyk, Luce, & Charles-Luce, 1994). Other phonological rules may take longer to acquire, such as the French *liaison*, which may take years to master (Chevrot, Dugua, & Fayol, 2009). Research on phonological development in children growing up with two languages is much more scarce, with only a handful of studies exploring phonological perception, principally in infants and toddlers, and another handful focusing on phonological production, mainly in preschoolers. In the following subsections we will review both lines of research.

1.3.1 Phonological perception in bilinguals

Evidence on phonological perception suggests that, like monolinguals, bilingual infants tune to the phonemic categories of their native languages during the first year of life (for a review, see Werker, 2012 and Hammer et al., 2014). However, their developmental trajectories are not yet well understood, with studies showing different patterns depending on the experimental method, the phonological contrast, and the language pair. Such is the case of the previously discussed vowel contrast /e/-/ɛ/. Bosch and Sebastián-Gallés (2003b) first investigated the perception of this contrast in different age groups of Spanish-Catalan bilingual infants, as well as in Spanish and Catalan monolinguals, using a familiarization-preference procedure. Their results confirmed, first of all, that at 4 months all three groups of infants could discriminate this contrast, regardless of the language they were exposed to. Furthermore, by the time they were 8 months old, both monolingual groups had learnt their native categories, that is, Catalan infants maintained their discrimination of /e/-/ɛ/, while Spanish monolinguals (for whom this contrast is not phonemic) had already lost sensitivity to it. Interestingly, bilingual infants who had been exposed to both languages since birth (and hence to both phonological categories) failed at discriminating these vowels at 8 months, but succeeded again at 12-months-old, suggesting that their phonological development follows a "U-shaped" pattern: they go from language-general perception, to a temporary collapse of the two categories, and finally a recovery of the phonemic distinction. A similar U-shaped acquisition has been attested in Spanish-Catalan bilinguals for the consonant contrast /s/-/z/, which also exists in Catalan but not in Spanish (Bosch & Sebastián-Gallés, 2003a), and the /o/-/u/ vowel contrast, which exists in both languages (Bosch & Sebastián-Gallés, 2005), but not in a more distant vowel contrast, /e/-/u/ (Bosch & Sebastián-Gallés, 2005).

A subsequent study revisited the /e/-/ɛ/ contrast with a different experimental paradigm (an anticipatory looking task), and found evidence that Spanish-Catalan bilingual infants can discriminate this contrast at 8-months-old (Albareda-Castellot, Pons, & Sebastián-Gallés, 2011), suggesting that specific task demands may influence their ability to attend to this contrast. Moreover, Sundara and Scutellaro (2011) studied Spanish-English bilinguals on this same contrast, which also exists in English. They showed, even using a similar experimental paradigm to the one used by Bosch and Sebastián-Gallés (2003b), that Spanish-English bilingual 8-month-olds had retained this distinction. The authors suggested that a possible explanation for Spanish-English bilinguals' success (while Spanish-Catalan infants had failed at this age) is that Spanish and English are easier to discriminate due to the fact that they belong to two different rhythmic classes, as previously discussed in Section 1.2. Thus, Spanish-

English infants' ability to sort out their input by language may have facilitated the separation of these phonemic categories.

Other contrasts seem to be less problematic. For instance, French-English bilingual infants retain their discrimination of a dental vs alveolar contrast distinguishing the French and English realisations of /d/, while monolingual French and English infants lose it by 10 months of age (Sundara, Polka, & Molnar, 2008). In another study with French-English bilinguals, Burns, Yoshida, Hill, and Werker (2007) explored infants' ability to discriminate /b/ from /p/. An interesting property of this contrast is that the voice onset time (VOT) boundary separating these two categories is different in French and English, thus leaving an intermediate range of VOT values in which French adults usually perceive a /p/, while English adults hear a /b/. Three age groups (6-8 months, 10-12 months and 14-20 months) of French-English bilingual and English monolingual infants were tested on their discrimination of this intermediate category against samples from the two extremes of the /b/-/p/ spectrum. Their results showed that, while both monolinguals and bilinguals behaved similarly in the youngest age group, by 10 to 12-months-old each group had tuned to their native language(s). That is, monolingual English infants treated the intermediate category as a /b/ (i.e., they discriminated it only against /p/), while bilinguals discriminated both contrasts, thus showing phonological knowledge of both of their languages. Bilinguals' success in these contrasts may have been due to the fact that their target languages (French and English) belong to different rhythmic classes, but more evidence of similar contrasts across different language pairs would be needed before we could understand the role of language distance on phonological development.

Some recent studies have begun to explore the effects of the specific language pair in a more systematic way, but are still scarce (Havy, Bouchon, & Nazzi, 2016; Liu & Kager, 2015). For instance, Liu and Kager (2015) explored the perception of this same /b/-/p/ intermediate contrast in bilingual infants learning Dutch and an additional language which either shares the same VOT values as Dutch for these two consonants (French, Spanish), or has a different VOT boundary (Chinese, English, German). Their results showed that the perception of these contrasts depended on the language pair (with those sharing similar realisations of these consonants showing a stable contrast discrimination at 11-months-old, although with noisy results at 8-9 months), as well as on their language dominance, affecting principally those infants learning a language pair with differing realisations of these sounds. More studies of this kind are needed in order to form a full picture of the developmental trajectory of phonemic categories in young bilinguals.

Beyond the development of sound categories, and past the first two years of life, not much is known about young bilinguals' phonological perception. Indeed, a language's phonological system is not limited to its composing sounds. Languages may differ in the possible combinations of phonemes within a syllable - that is, their phonotactics - as well as on phonological rules that may alter the surface form of a word when pronounced within a sentence. Sebastián-Gallés and Bosch (2002) explored Spanish-Catalan bilinguals' perception of Catalan phonotactic rules. Ten-month-old bilinguals, as well as Spanish and Catalan monolinguals, were tested on their discrimination of legal versus illegal consonant clusters. Their results showed that Catalan-dominant bilinguals, as Catalan monolinguals, succeeded in discriminating these two types of clusters, while Spanish-dominant bilinguals and Spanish monolinguals failed. To the best of our knowledge, no other study has explored the perception of phonological rules in young bilinguals, leaving a big gap in our understanding of their phonological development.

1.3.2 Phonological production in bilinguals

Many studies on phonological production have focused on preschoolers (Fabiano-Smith & Barlow, 2010; Fabiano-Smith & Goldstein, 2010; Fabiano-Smith, Oglivie, Maiefski, & Schertz, 2015; Goldstein, Fabiano, & Washington, 2005; Goldstein & Washington, 2001; MacLeod & Fabiano-Smith, 2015; Munro, Ball, Müller, Duckworth, & Lyddy, 2005; Nicoladis & Paradis, 2011), while some others have looked at infants and toddlers, although most often through case studies (Deuchar & Clark, 1996; C. E. Johnson & Lancaster, 1998; Kehoe, 2002; Kehoe, Lleó, & Rakow, 2004; Maneva & Genesee, 2002; Paradis, 2001; Schnitzer & Krasinski, 1994, 1996). In a study of a French-English bilingual infant, Maneva and Genesee (2002) found evidence of some language-specific phonological features in the infant's babbling, depending on the language of the parent that the child was interacting with. For instance, the child produced more stop + vowel syllables when interacting with his English-speaking parent, and more approximant + vowel syllables when interacting with his French-speaking parent. This suggests that language differentiation may take place from the onset of speech production.

Once children start producing words, most speech production studies have focused on the development of phonemic categories. Some have conducted longitudinal studies on a small number of children, analysing their natural productions (e.g., C. E. Johnson and Lancaster, 1998; Kehoe, 2002; Kehoe et al., 2004; Schnitzer and Krasinski, 1994, 1996). For instance, Schnitzer and Krasinski (1994, 1996) analysed the productions of two children acquiring Spanish and English, and found evidence of a

merged consonant system in one of them, but not in the other. Others have used elicited speech or picture naming tasks to collect production data in specific age groups (e.g., Fabiano-Smith and Goldstein, 2010; Fabiano-Smith et al., 2015; Goldstein et al., 2005; Goldstein and Washington, 2001). For example, Goldstein and Washington (2001) used a picture naming task to assess the phonological production of Spanish-English bilingual preschoolers. They observed that half of the children had a complete consonant repertoire in at least one of their two languages, and all of them produced the full vowel repertoire of both languages. Furthermore, they found some evidence (although rare) of cross-linguistic effects reflected in consonant substitutions. In a similar study with 3 to 4-year-old Spanish-English bilinguals, Fabiano-Smith and Goldstein (2010) found that their phonological accuracy in word production was higher in sounds that were shared by both languages. Overall, the available evidence from production studies suggests that bilingual children begin to develop phonological categories of both of their languages from early on, yet their dual phonological systems may sometimes interact.

Furthermore, a couple of studies have analysed other phonological properties of young bilinguals' productions. Paradis (2001) studied French-English 2-year-olds patterns of truncation during a nonce-word repetition task. Truncations (e.g., *nana* for "banana") are typical of toddlers' productions and are known to be influenced by language-specific word-prosodic properties (Allen & Hawkins, 1980). While the truncation patterns of the bilingual toddlers in each of their languages were found to generally match those of the corresponding monolingual French and English toddlers, there was also some evidence of cross-linguistic transfer. Thus, their phonological systems appeared to be differentiated, yet not fully independent. In another study using a picture naming task, Nicoladis and Paradis (2011) explored French-English bilingual children's production of liaison, a complex phonological rule in French that causes word-final silent consonants to be pronounced before a vowel-initial word (e.g., *petit* [pəti] "small" is pronounced [pətit] in *petit ours* "small bear"). They found that, in general, young children's French vocabulary — but not their age -– correlated positively with their production of liaison. Interestingly, when matched by vocabulary, bilinguals applied liaison less often than their monolingual peers', but only in low-frequency collocation frames. However, the sample size of this matched comparison was very small (6 monolinguals and 6 bilinguals between the ages of 3 and 5), making it difficult to draw conclusions.

In summary, while current evidence from perception and production studies suggests that young bilinguals start acquiring phonological properties of both of their languages from early on, their development does not seem to be equivalent to that of two monolinguals in one. Complex patterns of

phonological acquisition emerge depending on the specific language pairs and phonological properties that BFL learners set out to discover, sometimes showing a short-lived delay in the development of a phonological feature of one of the two languages, and sometimes showing cross-linguistic influences. However, the number of studies conducted so far remains very limited, leaving many open questions regarding their phonological development.

1.4 Lexical development

The acquisition of a lexicon involves multiple cognitive abilities that develop in early childhood. In order to learn words, infants must learn to segment sound sequences from the continuous acoustic signal, store a mental representation of their phonological form, assign a meaning to them, and eventually be able to produce them orally. Monolingual infants achieve these feats during the first two years of life, although lexical development is a process that continues throughout the entire childhood. Evidence from behavioral experiments shows that monolingual infants begin to use transitional probabilities and prosodic cues to segment new words from spoken utterances between the ages of 6 and 9 months (Curtin, Mintz, & Christiansen, 2005; E. K. Johnson & Jusczyk, 2001; Saffran, Aslin, & Newport, 1996; Thiessen & Saffran, 2003, 2007), and that they are able to recognise the sound of familiar words without a visual referent at 11 months (Hallé & de Boysson-Bardies, 1994; Swingley, 2005; Vihman, Nakai, DePaolis, & Hallé, 2004; Vihman, Thierry, Lum, Keren-Portnoy, & Martin, 2007). Furthermore, the development of word-referent mappings for very frequent words may start to develop as early as 6 months (Bergelson & Swingley, 2012; Tincoff & Jusczyk, 1999, 2012). Towards the second year of life, once infants have acquired their first few lexical items, they can use a variety of learning strategies to discover the meaning of new words. For example, young monolinguals have a mutual exclusivity bias: upon hearing a new word, children are likely to assume that the new label cannot refer to an object or concept for which they already know the word, and therefore must denote a new referent, be it a new object, or a part or property of a known object (Halberda, 2003; Markman, Wasow, & Hansen, 2003; Mather & Plunkett, 2011; Merriman, Bowman, & MacWhinney, 1989). Speech production also emerges in the first year of life. Infants typically produce canonical babbling (i.e., the repetition of syllables composed of a consonant and a vowel, such as “dadada”) by the age of 6 to 7 months (Eilers et al., 1993), say their first words around their first birthday, and by the time they are 18-months-old they can produce, on average, some 50 words (Fenson et al., 1994).

1.4.1 Bilinguals' milestones and learning strategies

As in many other areas of research on childhood bilingualism, data on lexical acquisition in young bilinguals is limited. Nevertheless, current evidence suggests that BFL learners achieve several of the previously mentioned milestones at the same age as their monolingual peers. Examples of this are the onset of canonical babbling (Oller, Eilers, Urbano, & Cobo-Lewis, 1997), the recognition of familiar word forms at 11 months (Vihman et al., 2007), and the age of production of their first words (Petitto et al., 2001).

Bilingual children may, however, differ from young monolinguals in many aspects of their lexical development. One of them is their use of word learning strategies. For instance, using a visual disambiguation task, Byers-Heinlein and Werker (2009) studied monolingual, bilingual and trilingual 17-month-olds' use of mutual exclusivity. Infants were presented with two pictures on a screen, one corresponding to a familiar object and the other to a novel object, while hearing a phrase prompting them to look at one of them. Their results showed that monolinguals and multilinguals differed in their use of mutual exclusivity: while monolinguals looked significantly longer at the novel object upon hearing a novel word, bilinguals did so only marginally, and trilinguals did not give evidence of using mutual exclusivity at all. Similar results, showing a reduction or absence of a mutual exclusivity bias in bilinguals, have been found in other experiments with infants and pre-schoolers (Davidson & Tell, 2005; Houston-Price, Caloghiris, & Raviglione, 2010; Kandhadai, Hall, & Werker, 2017). The fact that bilinguals rely less on mutual exclusivity to guess new word meanings has been argued to stem from their language experience: unlike monolinguals, bilinguals are likely to hear more than one label per object - i.e., one label per language - and may thus not develop a strong constraint on the number of words that can map to the same concept (Byers-Heinlein & Werker, 2009; Houston-Price et al., 2010).

Other studies have explored the interactions between bilinguals' phonological and lexical development. For instance, using a cross-modal word learning paradigm, Fennell, Byers-Heinlein, and Werker (2007) observed that 17-month-old bilingual infants exposed to English and a second language failed at learning a new minimal pair, *bih* - *dih*, differing only in the /b/-/d/ contrast (which is discriminated by bilingual infants at 14-months-old, Fennell, 2005), while monolingual English infants succeeded at this age. Bilinguals eventually achieved this feat at the age of 20 months, indicating a delay in their ability to use phonetic detail for word learning. The authors attributed this delay to the

increased demands of language acquisition in bilingual settings. However, subsequent studies testing other consonant contrasts and language pairs show varied results. In some experiments with French-English bilinguals, no evidence was found of a bilingual delay in minimal pair learning using the /b/-/g/ contrast (Mattock, Polka, Rvachew, & Krehm, 2010) or the /k/-/g/ contrast (Fennell & Byers-Heinlein, 2014) when the speech samples were produced by a bilingual speaker. Havy et al. (2016) explored the effect of cross-linguistic similarities on bilinguals' ability to learn minimal pairs. They tested 16-month-old bilinguals acquiring French plus a Romance language with similar realisations of their stop consonants /p,t,k,b,d,g/ (Spanish, Italian and Portuguese), or French plus a Germanic language (English, German) which differ in the realisation of these stops. The task consisted in learning a minimal pair of words with a 1-feature contrast either in voicing (e.g., /p/-/b/) or in place of articulation (e.g., /p/-/t/). Their results showed that infants learning a close language pair (e.g., French-Spanish) succeeded in this word learning task, while infants learning French plus a Germanic language failed to learn the word-object pairings. This suggests that similarities between properties of the two languages might play a role in lexical acquisition, but given the contradictory results between this and previous studies, more evidence would be needed before a conclusion can be drawn.

A related line of research has begun to explore the role of bilingual phonological acquisition on word recognition. Ramon-Casas, Swingley, Sebastián-Gallés, and Bosch (2009) studied Spanish, Catalan and Spanish-Catalan 18- to 25-month-old infants' ability to detect a vowel mispronunciation in familiar nouns. Infants were tested on a preferential looking task, in which the name of one of two objects presented on screen was either pronounced correctly or with an /e/-/ɛ/ vowel change. Consistent with their native phonological systems, monolingual Catalan infants detected the mispronunciations in Catalan nouns (producing shorter looking times towards the target image), while Spanish monolinguals did not notice the vowel change in Spanish nouns. Bilingual toddlers were tested only on Catalan words. Despite their familiarity with this vowel contrast (which, as we mentioned in Section 1.3.1, they can detect by the age of 12 months), they did not detect the mispronunciations. When tested on other more salient vowel distinctions present also in Spanish (e.g., /e/-/i/ or /e/-/a/), Spanish monolinguals and bilinguals both succeeded in this task. An additional experiment testing 3-year-old bilinguals indicated that Catalan-dominant preschoolers did detect the /e/-/ɛ/ change. However, in these series of experiments, the Catalan words on which bilinguals were tested all had cognates in Spanish, e.g., the word "bee" is pronounced [ə'βɛʎə] in Catalan and [a'βexa] in Spanish. In a follow-up experiment, Ramon-Casas and Bosch (2010) tested 2-year-old bilinguals on a similar task but using only Catalan words that have no cognates in Spanish. In this case, bilingual toddlers succeeded in

detecting the /e/-/ɛ/ mispronunciation. These findings suggest that cognates may have underspecified phonological representations, yet this does not impede the acquisition of phonological categories.

1.4.2 Vocabulary size and composition in bilinguals

Another aspect of lexical development in bilingual children that has attracted much attention is the size and growth rate of their dual vocabulary. Although bilinguals may reach some of the first lexical milestones at the same time as their monolingual peers, their two vocabularies may not necessarily develop at the same speed. In order to assess young children's vocabulary, several methods have been used, depending on the age of the child. A widely used method in infants and toddlers is the use of vocabulary lists filled out by parents, most typically the MacArthur-Bates Communicative Development Inventories (*aka* CDI, Fenson et al., 1994), which are available in many languages. In older toddlers and preschoolers, some studies have used alternative methods that do not depend on parental report, such as the Computerized Comprehension Task designed by Friend and Keplinger (Friend & Keplinger, 2003; Poulin-Dubois, Bialystok, Blaye, Polonia, & Yott, 2013) or the Peabody Picture Vocabulary Test (*aka* PPVT, Dunn, Dunn, Lenhard, Lenhard, and Suggate, 2015).

In one of the first studies of its kind, Pearson, Fernández, and Oller (1993) conducted a semi-longitudinal study with 25 Spanish-English bilingual infants between the ages of 8 and 30 months, and compared them with a group of monolinguals. Using the Spanish and English CDIs, they computed four vocabulary measures: English vocabulary, Spanish vocabulary, Total vocabulary (TV, i.e., English plus Spanish vocabularies combined), and Total conceptual vocabulary (TCV, i.e., Total vocabulary minus translation equivalents, resulting in the number of concepts for which they know at least one label). Although statistical analyses were limited due to the small sample size in each age group, their results showed that, when considering their TV or TCV, bilinguals' lexicons were comparable to that of monolinguals. However, the vocabulary scores in each of their two languages, when analysed individually, were sometimes smaller than that of their monolingual peers. Furthermore, they compared the evolution in time of bilinguals and monolinguals' production scores, from the age of 16- to 26-months-old, showing similar growth rates for Total vocabulary across both groups of infants. Since then, a number of studies have investigated bilinguals' vocabulary scores in early childhood, finding, in general, that bilinguals' vocabulary in one or both of their languages may be smaller than that of their monolingual peers, while their Total vocabulary scores are often comparable or even larger (Bialystok, Luk, Peets, & Yang, 2010; De Houwer, Bornstein, & Putnick, 2014; Hoff

et al., 2012; Junker & Stockman, 2002; Marchman, Fernald, & Hurtado, 2010; Poulin-Dubois et al., 2013). Furthermore, infants' vocabulary in each of their two languages may develop at different speeds (David & Wei, 2008; Pearson & Fernandez, 1994). Research has aimed at understanding which factors influence the development of each language's vocabulary, suggesting strong effects of relative amount of exposure, among many other input and environmental factors (e.g., Hoff and Core, 2013; Pearson, Fernandez, Lewedeg, and Oller, 1997; Place and Hoff, 2011), which we will discuss in Chapter 3.

A related line of research has focused on the content of the bilinguals' dual lexicon. While individual differences exist between children, and particularly in the growth rate of each of their two languages, bilinguals have been reported to acquire semantic categories in both languages in the same order as they typically appear in monolinguals, e.g., social words emerge before nouns, and these before verbs (Conboy & Thal, 2006; David & Wei, 2008). Furthermore, several studies explored the existence of translation equivalents (TEs) in young bilinguals' lexicons (Bosch & Ramon-Casas, 2014; De Houwer, Bornstein, & De Coster, 2006; Legacy et al., 2017; Pearson, Fernández, & Oller, 1995; Poulin-Dubois et al., 2013). Translation equivalents are words that can be considered to represent the same meaning across two languages, such as the English word *tree* and the French word *arbre*. Research shows that soon after they produce their first words, bilinguals start acquiring translation equivalents (De Houwer et al., 2006; Pearson et al., 1995), thus indicating that they can accept and learn more than one label for the same object. The existence of TEs in the bilinguals' lexicon has been interpreted as evidence that young bilinguals are able to differentiate, at least to some extent, their two languages (Genesee & Nicoladis, 2006; Patterson & Pearson, 2004).

In summary, lexical acquisition in young bilinguals seems to follow, as a whole, a similar timeline as in monolinguals. However, due to the fact that bilinguals encounter each language in different proportions and contexts, each respective vocabulary may develop at its own rhythm. Furthermore, bilinguals' lexical development shows signs of early language differentiation, but subtleties exist in the way their dual phonological and lexical systems interact. When forming their first phonological representations of words, similarities and partial overlaps between linguistic properties of the two languages may pose a particular challenge for the bilingual learner. Once more, the scarcity of data available so far limits the possibility to draw conclusions, but current evidence outlines the importance of investigating the interaction between bilinguals' phonological and lexical development.

1.5 Thesis overview

As we have shown throughout this review, while great progress has been made over the past 30 years in understanding bilingual first language acquisition, the available evidence remains scarce and scattered. Given the diversity of language pairs and scenarios in which bilingual children acquire their two languages, much research remains to be done in order to know which aspects of language development are general to all bilinguals, and which depend on each child's specific "constellation" of dual exposure. Important questions regarding when and how infants begin to separate their two languages, and to what extent they keep them apart throughout development, remain largely open. In this thesis, we will investigate three different aspects of bilingual development related to the problem of language separation from birth to pre-school years: the early discovery of two languages in the input, the potential role of environmental language separation on early lexical development, and the perceptual separation of phonological rules. We propose to take a multidisciplinary approach, using both empirical and computational techniques, which can provide different insights on the task of early language separation. The remainder of this thesis proceeds as follows:

First, in Chapter 2, we investigate the process of discovering two languages in the input during the first months of life. A relatively large number of experimental studies conducted to date on language discrimination in young infants has inspired researchers to begin postulating quantitative theories of how they achieve this task. A notable example of this is Ramus et al.'s 1999 work, in which they proposed a first computational account of how rhythmic properties may translate into infant behavior as observed in experiments. The use of computational models is of great utility to the investigation of cognitive development in humans (Dupoux, 2018). Indeed, research on early language acquisition has traditionally relied on observations of infant behavior in controlled laboratory experiments and on various measures of language outcome. This methodological approach has proven successful in collecting evidence of early cognitive and linguistic abilities, as well as determining an approximate timeline of language development. However, current experimental methods are limited to indirect observations of language perception, as the specific mechanisms and representations that occur in the infant's mind remain a "black box". Using computational methods, researchers can propose explicit models of the internal processes that underlie speech perception, and evaluate their plausibility through comparison with experimental data. As the field of developmental research advances, the construction of explicit models becomes crucial to take the extra step from current *conceptual* theories

of language acquisition (such as the PRIMIR framework proposed by Werker and Curtin, 2005) to detailed descriptions that can explain infant behavior and make accurate predictions. Following this rationale, and inspired by previous attempts at modelling language discrimination, in this chapter we propose a model of speech perception, and explore its ability to replicate infants' behavior when faced with multilingual speech.

Other aspects of bilingual acquisition have not yet been sufficiently documented to propose explicit models. Thus, for the remaining two research questions in this thesis we focus on collecting data that will further our knowledge of early bilingualism. In Chapter 3, we investigate several environmental aspects of bilingual exposure in 11-month-olds. In order to capture the natural diversity in infants' experience, we use a Language Diary method (De Houwer & Bornstein, 2003; Place & Hoff, 2011, 2016), in which parents of bilingual infants report the speakers and languages used in the child's environment throughout the day. These diaries allow us to characterise the variability in young bilinguals' dual language experience with better detail than traditional questionnaire methods, particularly on how often a bilinguals' two languages co-occur throughout a typical day. Furthermore, using parental reports of infants' vocabulary, we explore the potential impact of different environmental factors on lexical acquisition.

In Chapter 4, we investigate bilingual preschoolers' perception of language-specific phonological rules. Unlike other properties of young bilinguals' phonological systems, their acquisition and separation of phonological rules has barely been explored, with the only existing evidence coming from production studies. Phonological rules alter the way in which certain sounds are pronounced, depending on their phonological context within a word or a sentence. As bilinguals' input is likely to contain more phonetic variability than in monolinguals, to what extent young bilinguals are sensitive to context-specific alterations, and whether they can keep language-specific rules apart, is not clear. In this chapter, we use a touchpad videogame to test French-English bilingual children's perception of *assimilation*, a phonological rule existent in both French and English, but which affects different phonological properties in each language: in French, it affects the *voicing* of certain consonants, while in English it alters their *place of articulation*. By testing both French and English assimilation rules in one same language, we can investigate both questions regarding their sensitivity to context-specific alterations and the language-specificity of their knowledge.

Finally, we conclude in Chapter 5 with a general discussion of our findings and future lines of research.

Chapter 2

Modelling early language discrimination

2.1 Introduction

The task of language discrimination can be described as the separation of utterances belonging to two or more languages along some perceptual dimension(s). As is well known, the world's languages differ in a large variety of properties, ranging from sub-lexical (e.g., phoneme inventories, phonotactics and prosody) to lexical and syntactic. While adults can perform the task by using more or less explicit knowledge of these properties, young infants exposed to multilingual speech are faced with the challenge of naturally discovering the underlying languages without even knowing how many of them are spoken in their environment. Their ability to do so may depend on a large number of factors. For instance, the perceptual dimensions that infants pay attention to could determine whether differences in linguistic properties between the target languages will be salient or not. Furthermore, variability in the input from such sources as speaker, context or mood may all have an impact on language separation. Along with experimental work on language discrimination and on the role that it may have on later language outcomes (which we have reviewed in Chapter 1), the use of modelling may help understand the impact of different factors on this task. In the following section, we will first describe previous attempts at modelling language discrimination from a psycholinguistic point of view, discussing what remains to be explained. Then, we will review different engineering approaches, which will serve as an inspiration to tackle the shortcomings of the psycholinguistic approaches.

2.1.0.1 Psycholinguistic approaches

Cognitively- and linguistically-motivated models of language discrimination have mainly focused on the role of rhythmic properties of language (Dominey & Ramus, 2000; Galves, Garcia, Duarte, & Galves, 2002; Ramus et al., 1999; Varnet, Ortiz-Barajas, Erra, Gervain, & Lorenzi, 2017). Inspired by evidence suggesting that newborn infants can discriminate languages across rhythmic class in low-pass filtered speech (Mehler et al., 1988; Nazzi et al., 1998), several studies have used temporal structure as the relevant dimension for language separation.

Ramus et al. (1999) first investigated this problem by studying how different measures related to speech rhythm could separate languages into broad rhythmic classes. In this study, it was assumed that infants perceive speech as an alternation of vowel (V) and consonant (C) segments. Speech samples were thus manually segmented into CV sequences, and for each utterance three different measures capturing the proportion of V segments ($\%V$) and the standard deviation of V and C segments (ΔV & ΔC) were computed. Their results showed that $\%V$ combined with ΔC resulted in an apparent clustering of the languages (shown in Figure 2.1) into rhythmic classes. Based on these results, they proposed to use $\%V$ to model a language discrimination task that mirrored the habituation experiments conducted with infants. For each step in the simulation, they defined arousal as the distance between the current utterance's $\%V$ and the average of this value over all previous utterances. After habituation with one language, they continued with new utterances either from the same language, or from a different one. Statistical tests performed on the test utterances comparing both conditions showed similar patterns to infant experiments: languages were generally easier to discriminate across rhythmic classes than within the same rhythmic class.

One limitation of this study is that the model depends on a manual segmentation of the signal into CV segments. A clear-cut distinction between vowels and consonants is sometimes difficult to make[1]; infants might thus produce different speech segmentations which could yield different results. Furthermore, while 4 different speakers per language pair were used in the simulation, the samples were in fact composed of read speech, and utterances were selected to be of similar duration and number

[1]While many vowels and consonants can be easily classified, this is not always the case: for instance, some languages contain syllabic consonants, i.e., consonants that take the position of the nucleus of a syllable, replacing vowels. Notorious examples are the Czech language, where words and even full phrases may be composed solely of consonants (as illustrated by the famous tongue-twister "strč prst skrz krk", meaning *stick a finger through the throat*), or a dialect of Berber studied by Dell & Elmedlaoui (1985). Furthermore, when hearing filtered speech, the difference between certain consonants (e.g. nasals) and vowels may be difficult to perceive. To work around this problem, a different modelling approach based on sonority measures has been proposed by Galves et al. (2002)

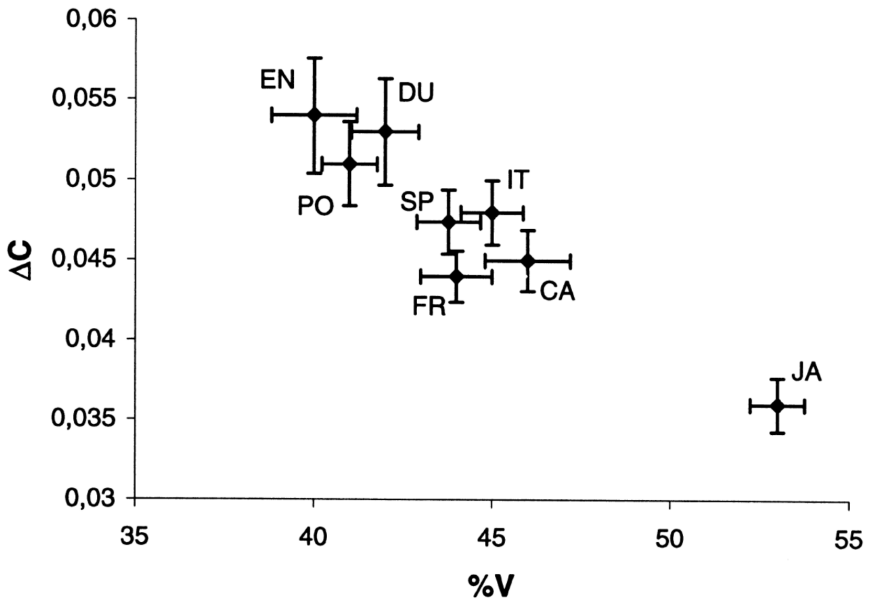


Figure 2.1: Classification of 8 languages along rhythmic dimensions $\%V$ and ΔC. Figure reproduced from Ramus, Nespor, and Mehler (1999).

of syllables. Thus, the natural speaker variability in speech rhythm could have been unintentionally toned down by this manipulation (Ramus, 2002a). While this does not pose a problem for the comparison with experimental results, where stimuli are often controlled in similar ways, it is problematic for the interpretation of the impact for real-world situations that infants might be exposed to. In a recent study, Arvaniti (2012) analysed the performance of these and other rhythmic metrics in grouping languages into rhythmic classes, using speech from several speakers and speech registers (e.g. spontaneous or reading) per language. Arvaniti found that rhythmic variability between speakers and speech registers was often comparable to or even larger than the variability between languages.

Dominey and Ramus (2000) proposed a different approach, using a temporal recurrent network (TRN) inspired on the neuroanatomy of the fronto-striatal system in primates[2]. As in Ramus et al. (1999), this model used a transcription of speech utterances into CV sequences (here sampled every 5 ms). The TRN was designed to encode the serial order and duration of events, and thus had the potential to learn temporal properties from the input. They found that this model could learn to discriminate languages from different rhythmic classes, while it failed to do so within rhythmic class. However, these results were only observed when the model was specifically trained to separate the languages. The classification performance of a naïve model without training remained at chance level. Nonetheless, the model showed some evidence of sensitivity to rhythmic properties of the languages. For instance,

[2]The fronto-striatal circuit is believed to be involved in reinforcement learning (Schönberg, Daw, Joel, & O'Doherty, 2007)

they found that a correlation between output vectors and target languages was only significant if the languages belonged to two different rhythmic classes, but not if they belonged to the same rhythmic class. This approach improves on Ramus et al.'s (1999) model as it does not take a pre-defined measure of rhythm, but rather learns the rhythmic properties in a more organic way from the input sequence. However, it still requires a manual segmentation of the speech into CV segments.

In a recent study, Varnet et al. (2017) studied how amplitude and frequency modulations[3] (AM, FM) of the speech signal could reflect language differences. Their results showed that languages did not differ significantly in their FM spectra, while amplitude modulation resulted in a grouping of languages consistent with the traditional rhythmic classes. While they did not propose an explicit model of infant's language discrimination, the computation of the FM and AM spectra were fully unsupervised and inspired in human auditory perception. Infants could potentially detect these properties in a habituation experiment.

These cognitive approaches have succesfully shown that rhytmic information can be used to separate languages into classes. This is coherent with linguistic theories of language classification and with perceptual experiments with infants. However, speech has many sources of variability besides its rhythmic properties, and infants must deal with all these sources simultaneously. In this respect these models are limited; they do not explain why infants fail to discriminate languages belonging to two different rhythmic classes, when presented with unfiltered speech from several speakers (Ramus, 2002b). A model consistent with all experimental evidence (both from filtered and unfiltered speech), should therefore incorporate other sources of variability which could compete with rhythm in a comparable ground. Furthermore, we are interested in understanding how experience with one or more languages may shape the perception of language distance. It is therefore necessary to find a way to incorporate language experience to the model, a feature that only Dominey and Ramus' (2000) model allowed so far. To build a model that can cope with these problems, we propose to take inspiration from the large amount of work done in the field of speech engineering. We will now review different engineering approaches to language separation and discuss how they can be repurposed to model language discrimination in infants.

[3]Amplitude and frequency modulations characterize the fluctuations of the temporal envelope and of the spectral components of speech throughout the utterance.

2.1.0.2 Engineering approaches

In speech technologies, language discrimination plays an important role in *language identification* (LID) systems. LID methods aim to automatically recognize the language of a given speech sample. In order to do this, labeled utterances are first used to learn a representation of the linguistic properties of each language in a set, and then new unlabeled samples can be compared to each language to find the one with the maximum likelihood. In this context, finding the dimensions that best separate the languages is an implicit part of solving the problem. LID systems thus vary on the type of properties used to characterize the languages, as well as on the algorithms used to model these properties, and finally on the classification methods. While most of the LID pipeline requires supervision (i.e., language labels are needed in order to train classifiers), the extraction of relevant linguistic properties from the speech signal can be fully unsupervised. LID approaches using all possible properties have been proposed (Li, Ma, & Lee, 2013), but here we will focus solely on sub-lexical features, namely acoustic-phonetic, phonotactic, and prosodic. For each of these, we will give a general description of the feature representations and the most common LID methods using these properties as input. It should be noted that, as in many AI fields, deep neural networks (DNNs) are becoming increasingly popular in LID applications (e.g., Lopez-Moreno et al., 2014; Richardson, Reynolds, and Dehak, 2015). However, to this day these models require strong supervision and large amounts of training data, and often remain as much of a black box as the infant's mind. For these reasons, we will not consider these approaches here, but we believe ongoing efforts in reducing the amount of supervised training and in understanding what DNNs learn will make this approach an interesting model of infants' perception.

Acoustic-phonetic features

Many LID approaches rely on differences in the distribution of acoustic features across languages (Campbell, Singer, Torres-Carrasquillo, & Reynolds, 2004; Castaldo, Colibro, Dalmasso, Laface, & Vair, 2007; Dehak, Torres-Carrasquillo, Reynolds, & Dehak, 2011; Singer, Torres-Carrasquillo, Gleason, Campbell, & Reynolds, 2003; Zissman, 1996). To represent these properties, the speech signal is generally converted into a sequence of short time frames defined by *feature vectors* capturing frequency components. The most commonly used representation is the *Mel-Frequency Cepstral Coefficients* (a.k.a. MFCC, Davis and Mermelstein, 1990), which roughly approximates some characteristics of human auditory perception. MFCCs are vectors that contain spectral information of the signal within

short time windows (usually < 40 ms), and are often accompanied by their first and second order derivatives (known as *delta* features) to capture local transitions. In order to use acoustic features for LID, language-tagged feature vectors are used to train classifiers or to build language-specific models. A simple and popular approach is to model the distribution of acoustic features using a Gaussian Mixture Model (GMM), that is, a combination of Gaussians (Zissman, 1996). In this model, the feature vectors that compose an utterance are considered as independent observations of the acoustic properties of speech, losing information about the order of the sequence. These systems are often combined with additional supervised methods, such as Linear Discriminant Analysis or Support Vector Machines (Campbell et al., 2004; Castaldo et al., 2007), to improve their performance.

Phonotactic features

Other popular LID approaches rely on differences between the frequency of occurrence of sound sequences (Glembek, Matějka, Burget, & Mikolov, 2008; Matejka, Schwarz, Cernocky, & Chytil, 2005; Tong, Ma, Li, & Chng, 2009; Zissman, 1995). To represent these phonotactic properties, the speech signal is usually converted into a sequence of discrete acoustic tokens (e.g., phones), which are then used to train *n-gram* models[4]. This technique, called *phone recognition followed by language modelling* (PRLM), generally yields excellent results, but it requires huge amounts of labelled data (in this case, phonological transcriptions of the recordings) to build a speech tokenizer. To overcome this problem, practical applications often use pre-trained speech recognizers from one (or several) language(s) to decode the speech sequence of a different language. As young infants do not have access to such phone decoding systems during the first few months of their lives, this approach is not appropriate to model their first steps in language discrimination. However, it may be an interesting model at a later developmental stage, when infants have acquired a stable phonemic repertoire.

Prosodic features

Although not as popular as the acoustic and phonotactic methods, some approaches to LID have proposed using prosodic features to capture differences between languages (Martinez, Burget, Ferrer, & Scheffer, 2012; Ng, Leung, Lee, Ma, & Li, 2010; Pellegrino, Chauchat, Rakotomalala, & Farinas, 2002; Rouas, Farinas, Pellegrino, & André-Obrecht, 2005). Most prosodic LID systems perform an

[4]N-grams models compute the frequency of co-occurrence of all possible combinations of n phones, thus capturing phonotactic patterns

automatic segmentation of the speech signal into syllable-like units, from which prosodic features such as pitch, energy and duration of voiced intervals are calculated. For instance, Pellegrino et al. (2002) developed a system that automatically segments speech into pseudo-syllables composed of consonant and vowel intervals. These are then transformed into feature vectors, capturing several prosodic properties. Finally, these vectors are used as input to traditional supervised algorithms – such as GMMs – obtaining in general good classification results. In a follow up study, Farinas, Pellegrino, Rouas, and André-Obrecht (2002) showed that combining this rhythmic approach with acoustic modelling yields better performance than either method used on its own. In general, prosodic systems provide an improvement in classification performance when fused with other LID approaches, such as acoustic + prosodic as demonstrated by Farinas et al., or phonotactic + prosodic (Ng et al., 2010).

Dynamic features

An interesting mid-point between all three kinds of features is attainable through the use of dynamic features. Particularly, the performance of the acoustic models improves drastically by using *Shifted Delta Coefficients* (SDC, Torres-Carrasquillo et al., 2002), which capture the evolution of the feature vectors over a time window of roughly the length of a syllable. By using SDCs in addition to static MFCCs, GMM-based models can capture regularities not only of the distribution of sounds, but also of the transitions and co-occurrences of different sounds, offering a simple way of incorporating phonotactics. Furthermore, SDCs can also be computed over prosodic features, such as pitch and energy, thus merging all three sub-lexical approaches in one. The introduction of SDC features has proven extremely successful as it increases the classification performance of GMM acoustic models while retaining their simplicity.

Global features: i-vectors

The features discussed so far all represent, to a greater or lesser extent, local features. Even in the case of phonotactic and prosodic properties, these features do not cover more than the length of a syllable. LID models using these types of features thus base their decisions on local evidence, while at the same time evaluating each feature dimension independently. Furthermore, the language recognition systems described so far require large amounts of labelled multilingual data, making them implausible

representations of young infants' perception. In recent years, a different approach emerged that quickly became state-of-the-art: the *i-vector* model (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2011; Dehak, Torres-Carrasquillo, et al., 2011). In contrast with previously described systems, the i-vector model produces *global* features that represent properties of the utterance as a whole. This approach (originally designed for speaker recognition) derives from a family of models developed over the past 20 years, in which a language- (or speaker-) independent GMM is used as a *Universal Background Model* (a.k.a. UBM, Reynolds, Quatieri, and Dunn, 2000), that is, a model representing general properties of speech. In traditional GMM-UBM models, specific language (or speaker) models can be obtained by adapting (i.e., *shifting*) the parameters of the UBM. In the i-vector model, the dimensionality of the shift is reduced by finding the feature dimensions (or combinations thereof) that best capture variability between utterances. Any new utterance can thus be approximated by an offset from the background model in this smaller subspace. In other words, the resulting shift vector – i.e., the i-vector – represents global properties of the utterance. This process does not require any language or speaker labels, as it ignores the identity of the utterances during training, making it completely unsupervised.

In this representation, speech coming from the same language (or speaker) will tend to shift in a similar direction, thus facilitating the classification of utterances. This model has many applications in speech processing systems, such as speaker, gender and mood identification, and speech diarization, among others (for a review see Verma and Das, 2015). In LID applications, the i-vectors are then combined with supervised methods to reduce undesired variability and train language-specific models or classifiers. However, they can be used on their own as an utterance-level representation requiring no supervision, and thus may be feasible models of language perception.

From a cognitive perspective, the i-vector pipeline makes two main assumptions: first, that infants have good acoustic perception, and second, that they are sensitive to statistical regularities in their input. Experimental evidence suggests that infants fulfil both of these assumptions (Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Kuhl, 2004; Maye, Weiss, & Aslin, 2008; Maye, Werker, & Gerken, 2002; Saffran et al., 1996). Furthermore, this model requires learning one single distribution of speech features, unlike other LID models where separate distributions are learnt for each language. The single distribution is a more feasible model of accumulated experience for young infants who have yet to start sorting their input. Finally, the learning algorithm involved in the computation of the i-vectors makes no assumptions about phonemes, syllables, or words, as it learns the regularities (and the dimensions of variability) directly from the input using simple statistical methods.

The use of i-vectors has many advantages over previous psycholinguistic models. Particularly, it allows to have a language background representing previous experience, and it allows to incorporate different sources of information into the same perceptual space. For instance, by using MFCCs combined with prosodic features and SDCs, it is possible to model simultaneously the acoustic and prosodic space, while the dimensionality reduction step automatically decides which of these dimensions (or combinations thereof) are relevant. Thus, i-vectors overcome some of the previously mentioned shortcomings of psycholinguistic models.

In this project, we propose to investigate the potential use of i-vectors to model language discrimination in infants. First, we will describe the i-vector pipeline, from feature extraction to building the i-vector system. Next, we will present a series of experiments that begin to explore the extent to which these models can mimic infant behavior, and whether they could provide new insights on early language separation. Finally, we will discuss the directions of future work.

2.2 I-vector pipeline

In this section we will briefly describe the pipeline of the i-vector model (shown in Figure 2.2). For a detailed description of each step, see the Appendix of the present chapter (Section 2.A).

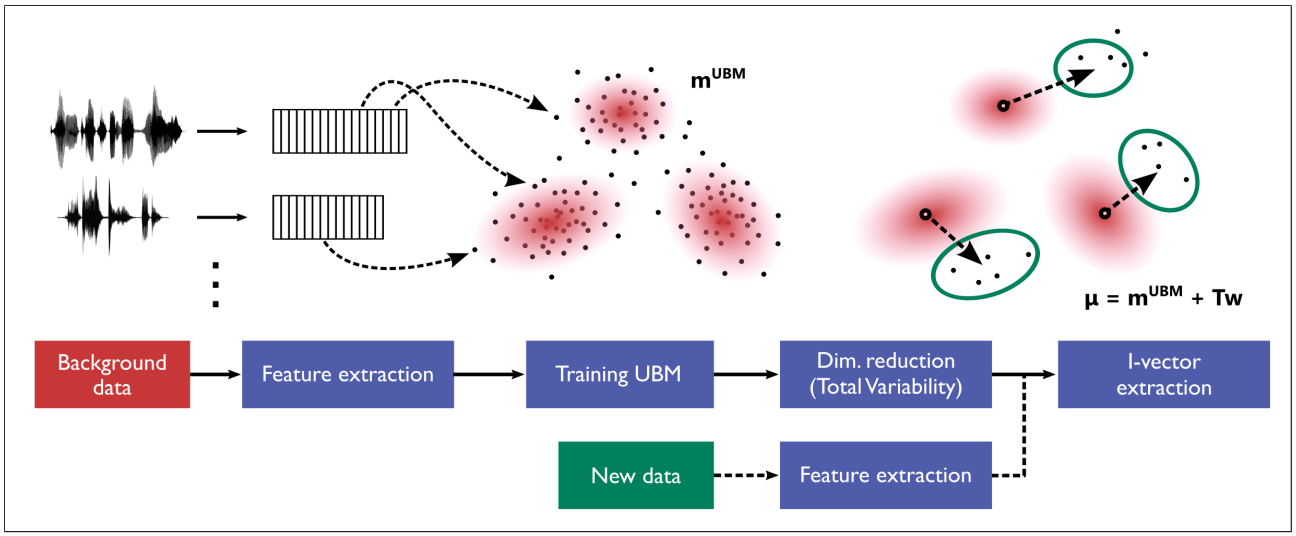


Figure 2.2: Pipeline for i-vector extraction.

The pipeline begins by extracting feature vectors from a set of background (training) utterances. For this purpose, we use the HTK Speech Recognition Toolkit (Young et al., 2006) to extract the traditional MFCCs (Davis & Mermelstein, 1990), which represent spectral properties of the signal based on a

frequency scale that approximates human auditory perception (the Mel-frequency scale). These feature vectors are computed in small moving windows of 25 ms width, shifted every 10 ms along the signal. The MFCCs are then expanded with energy (corresponding to the 0^{th} MFCC coefficient) and pitch (*F0*, computed with the Kaldi Toolkit, Povey et al., 2011). Finally, to capture dynamic information, Shifted Delta (SDC) features are computed from the MFCC-F0 vectors (Torres-Carrasquillo et al., 2002). Using the typical SDC configuration (7-1-3-7), the resulting MFCC-SDC features will form a vector with $D = 64$ dimensions. These features capture both static and dynamic properties (i.e., acoustic-phonetic and prosodic information).

The extracted features are then used to build a model of the background distribution (i.e., the UBM). The UBM is a Gaussian Mixture (GMM), defined by a *supervector* containing the means of each Gaussian component, $\mathbf{m}^{\text{UBM}}$ (Reynolds et al., 2000). The dimension of this supervector is $DK \times 1$ (D: number of features, K: number of Gaussians). Typical LID systems have approximately 1 to 2 thousand Gaussian components. The parameters (means and covariances) of the UBM are estimated using an Expectation Maximization (EM) algorithm (a description of this algorithm can be found in Section 2.B). It should be noted that, while EM is a batch algorithm, it is possible to train a GMM incrementally (Zhang, Chen, & Ran, 2010). In this project we will use the traditional batch EM, but future work should investigate how the model's behavior changes as it adapts to new input data with an incremental algorithm.

Next, an unsupervised dimensionality reduction algorithm – similar to a factor analysis – is used to find the dimensions of largest variability between utterances (Dehak, Kenny, et al., 2011). This low-dimensional subspace, characterized by the matrix $\mathbf{T}$, is referred to as the *Total Variability* space. The number of factors (F) in this subspace (i.e., the dimensionality of the i-vectors) is defined prior to training. Typical LID applications use approximately 400 Total Variability factors.

Finally, any new utterance[5] $\boldsymbol{\mu}$ can be approximated by a shift from the UBM, constrained to the subspace $\mathbf{T}$:

$$\boldsymbol{\mu} = \mathbf{m}^{\text{UBM}} + \mathbf{T}\mathbf{w} \tag{2.1}$$

where $\mathbf{w}$ is a low-dimensional, fixed-length shift vector: the *i-vector*. Both the background training and the i-vector extraction are computed using the MSR Identity Toolbox (Sadjadi, Slaney, & Heck, 2013).

[5]In the traditional i-vector notation, the utterance supervector is referred to as $\mathbf{M}$. We avoid this notation as it can be misinterpreted as a matrix.

In LID applications, these models are trained with large amounts of data and then combined with supervised classification techniques, yielding excellent performance in language identification tasks. In this project, however, we will scale down the amount of training data, and completely remove all supervised algorithms, thus using the i-vectors as simple vector representations of speech. As mentioned previously, i-vectors represent global features of the utterance, as seen from the perspective of the background model. Given their vectorial properties, it is possible to measure the distance between any two utterances, for instance using the cosine distance (i.e., the angle of separation between the vectors). In the following studies, we will use this distance to quantify the separation of the utterances from different languages and speakers.

In the following section we will describe a series of six computational experiments using i-vector models.

2.3 Computational experiments

In this section we will present a series of experiments investigating the behavior of the reduced i-vector model when faced with multilingual speech. As we are interested in modelling language discrimination abilities observed in very young infants, we will train our background models with small training datasets, and then test their ability to discriminate new utterances in several languages. Whenever possible, we will compare the model's behavior with existing experimental data. If the representations generated by the model are similar to the representations a young infant may form, then we expect the model to make similar mistakes when confronted with difficult language pairs.

To evaluate the discriminability between languages, we will use a non-parametric measure inspired in a human perceptual task: the computational ABX score (Schatz, 2016; Schatz et al., 2013). This measure, which we will describe in Experiment 1, allows us to quantify the degree of separation of two classes (here, two languages) based on the distance between their elements. Other techniques, such as Principal Component Analysis (PCA), hierarchical clustering, and models of the infant habituation task, will be used along with ABX to better understand the shape of the representations and how these may translate into overt behaviors.

2.3.1 Experiment 1: A proof of concept (Article 1)

To begin investigating the plausibility of i-vectors as models of speech perception, we conducted a first study comparing the discrimination of two language pairs: one with distinct acoustic properties (French & English), and one with highly overlapping properties (Spanish & Catalan).

This study was published as: Carbajal, M.J., Fér, R. & Dupoux, E. (2016) Modeling language discrimination in infants using i-vector representations. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Modeling language discrimination in infants using i-vector representations

M. Julia Carbajal [1] **(carbajal.mjulia@gmail.com)**
Radek Fér [2] **(ifer@fit.vutbr.cz)**
Emmanuel Dupoux [1] **(emmanuel.dupoux@gmail.com)**

[1]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS; 29, rue d'Ulm
75005 Paris, France
[2]Speech@FIT, Faculty of Information Technology, BUT; Božetěchova 2
612 66 Brno, Czech Republic

Abstract

Experimental research suggests that at birth infants can discriminate two languages if they belong to different rhythmic classes, and by 4 months of age they can discriminate two languages within the same class provided they have been previously exposed to at least one of them. In this paper, we present a novel application of speech technology tools to model language discrimination, which may help to understand how infants achieve high performance on this task. By combining a Gaussian Mixture Model of the acoustic space and low-dimensional representations of novel utterances with a model of a habituation paradigm, we show that brief exposure to French does not allow to discriminate between two previously unheard languages with similar phonological properties, but facilitates discrimination of two phonologically distant languages. The implications of these findings are discussed.

Keywords: language discrimination; speech; acoustics; computational models; habituation

Introduction

When infants acquire their first language, they meet the formidable challenge of dealing with massive variability and ambiguity at all levels of acoustic and linguistic structure. Infants growing up in a multilingual environment must face an additional level of variability due to the presence of two (or more) languages with independent yet partially overlapping acoustic and structural properties. Although the task may seem hard, a large number of studies show that the ability to discriminate spoken languages is present early on in life (Mehler et al., 1988; Nazzi, Bertoncini, & Mehler, 1998; Nazzi, Jusczyk, & Johnson, 2000; Bosch & Sebastian-Galles, 2001; Ramus, 2002; Byers-Heinlein, Burns, & Werker, 2010). For example, using a habituation paradigm, Mehler et al. (1988) showed that French newborns, in spite of their brief experience with language, are able to discriminate their native language from a foreign one (in this case, Russian) as evidenced by an increase in their arousal following a switch from Russian to French utterances. This discrimination was still observed when infants were presented with low-pass filtered speech, and a preference for their native language was suggested by an asymmetry in the arousal depending on the language presented during habituation.

Further research extended these findings, supporting the claim that newborns can distinguish any two unheard languages if they belong to different rhythmic classes, such as Japanese and English, but that they fail to do so if they belong to the same rhythmic class, e.g., English and Dutch (Nazzi et al., 1998; Byers-Heinlein et al., 2010). These results point at prosody as a strong cue for language discrimination at an early developmental stage. However, languages often differ in many other dimensions, such as their phonemic inventories and phonotactic rules. These cues may become relevant through further exposure to one or more languages and thus facilitate their discrimination: by 4 to 5 months of age, both monolingual and bilingual infants can discriminate two languages even within the same rhythmic class, such as Spanish and Catalan, if they were exposed to at least one of them before (Nazzi et al., 2000; Bosch & Sebastian-Galles, 2001).

While these studies suggest that language distance plays an important role, the specific acoustic features and mechanisms that may allow language discrimination throughout the first year of life, and the impact of prior exposure to one or more languages, are not yet fully understood. In the present study we explore how state-of-the-art speech technology tools can help us understand this feat. As a first step in the application of these novel techniques to the study of infant perception, we propose the use of *i-vectors*, a method to represent any given utterance as a pattern of deviations from a previously constructed background acoustic distribution, to implement an unsupervised model of language discrimination. The i-vector representation, in combination with discriminative classifiers, was originally developed for automatic Speaker Recognition (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2010), and in recent years has been adapted to Language Identification systems showing excellent performance (Martínez, Plchot, Burget, Glembek, & Matějka, 2011). These models are typically trained on large datasets containing many different speakers/languages to capture all possible sources of variability. Here, we simplify the model to represent the brief experience of an infant exposed to a single speaker of French, and then test the system's ability to discriminate new unheard utterances of two languages that differ in many phonological dimensions, such as rhythm, syllabic structure and phonemic repertoire (French and English), and two languages with largely overlapping phonologies (Spanish and Catalan). As most studies of language discrimination have made use of habituation paradigms, we also propose a computational model of the habituation task, which will allow us to compare the performance of our system with what has been observed in young infants.

The remainder of the paper unfolds as follows. We first introduce the concept of Universal Background Model and i-vector representation, discussing how these representations can be adapted to model infants' experience. Next, we describe the datasets that we selected for the modeling of the background space and the language discrimination tests. Then, we present a model of the habituation task that uses the extracted i-vectors as input, and two additional measures of discriminability. Finally, the results are described and discussed with respect to current experimental data, followed by a perspective on future work.

Methods

Universal Background Model and i-vectors

The first step of the modeling consists in constructing a representation of the acoustic space formed through the infant's exposure to a given linguistic environment, i.e., their "native" language. To model the distribution of speech features, speech technologies typically use Gaussian Mixture Models (GMM). With a sufficient number of mixture components, GMMs can model any arbitrarily complex distribution. The typical number of components for a Language Identification (LID) system is around one thousand.

The parameters (weights, means and covariances) of the model can be estimated by Maximum Likelihood using an Expectation-Maximization algorithm (Bishop, 2006). A GMM trained on a large database of several hundred hours of speech containing many different speakers, languages and other sources of variability, can be used to represent the overall feature distribution. In the context of speaker and language recognition, this is called the *Universal Background Model* (UBM). Evidently, young infants cannot count on such a large and variable amount of data to build their representations of the acoustic space, however, nothing prevents UBMs from being trained on a much smaller dataset. In the present study, we train a small UBM with speech from one single French speaker to represent the brief exposure that even a 4-day-old infant may have already encountered.

Once the UBM has been trained, data-specific models representing feature distributions of different utterances can be derived from the UBM by Maximum a Posteriori (MAP) adaptations. Usually, only the component means are shifted during the adaptation. Using factor analysis, the adaptation offset with respect to the UBM can be confined to a low-dimensional subspace, called the Total Variability space. If we denote by $\mathbf{m}$ the stacked vector of UBM component means, the generative subspace model has the form:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{v},$$

where $\mathbf{T}$ is a low-rank matrix (Total Variability matrix) defining the bases for the subspace, and $\mathbf{v}$ is a hidden variable with standard normal prior. As with the UBM, this subspace is typically trained on a large number of speech recordings using EM algorithms (Dehak et al., 2010), but for the purpose of our model it will be trained on the data of a single speaker.

Finally, given an utterance or any other segment of a speech recording, the posterior distribution of the hidden variable can be estimated. The MAP point estimate of this distribution is conventionally called an *i-vector*, and can be used as a low-dimensional fixed-length representation of the speech segment. In other words, any unheard utterance can be approximated as a deviation from the background "native" model. We propose to use this simple representation to model the infant's acoustic perception of previously unheard speech, computing an i-vector for every utterance in our test dataset. The advantage of this vectorial representation of speech is that a measure of distance can be defined between any two utterances.

In LID systems, the typical dimensionality of the subspace is around 400. However, for our experiments, the i-vector dimensionality is set to 200, and we use a UBM with 256 mixture components and diagonal covariance matrices. The reason for such a small model is that the database we propose to use in order to model a brief exposure to French is not large enough to robustly estimate all the parameters of a conventional LID model. Furthermore, since our database contains only a limited amount of variability (UBM trained on one single speaker and language), it is unnecessary to increase the number of dimensions.

We argue that i-vectors are reasonable as models of infants' representation of languages for the following reasons: (1) The entire pipeline (construction of UBM and i-vector extraction) only requires two skills, which have been documented in infants: a good acoustic perception (Eimas, Siqueland, Jusczyk, & Vigorito, 1971), and the ability of performing statistical learning (Saffran, Aslin, & Neport, 1996; Maye, Werker, & Gerken, 2002). (2) The learning algorithm is completely unsupervised, requiring no external information about phonemes or words, nor any information about speaker identity, or number and properties of different languages. The only linguistic hypotheses of this model are that utterances are relevant units for performing language discrimination, that they can be modelled through gaussian mixtures, and that they can be segmented out of continuous speech.

Feature extraction A common representation of the acoustic features of a speech signal used in many speaker and language identification systems are *Mel-Frequency Cepstral Coefficients* (MFCCs), which are based on a transform of the power spectrum on a frequency scale that approximates human auditory perception. For our modeling purposes, these features were calculated using the HTK Speech Recognition Toolkit (Young et al., 2006) in 25 ms windows with a 10 ms shift. We retained the first 7 coefficients (including *C0*, which represents the energy) and added a measure of *F0* (pitch) computed with the Kaldi Toolkit (Povey et al., 2011).

In addition, *Shifted Delta Coefficients* (SDC, a stacked version of delta coefficients calculated across several frames, Torres-Carrasquillo et al., 2002) were included to capture the temporal evolution of the MFCC-F0 features. The SDCs were calculated using the usual 7-1-3-7 configuration, resulting in

an approximation of the contour of the MFCC-F0 features over a span of 200 ms. The resulting 64-dimensional MFCC-F0-SDC vectors contain both spectral and prosodic information presumed available to the human auditory system.

Materials

Training data In order to train the UBM to represent the prior experience of an infant with a brief exposure to French, we used a dataset of casual speech recorded from an adult female French speaker selected from the *Corpus of Interactional Data* (Bertrand et al., 2008). The selected dataset is composed of 602 pre-segmented utterances with a mean length of 2.54 seconds ($min = 0.43$ s, $max = 9.01$ s), giving a total of approximately 25 minutes of clean speech. The original recordings were downsampled to 16kHz.

Test data Similarly to previous experimental studies, to test the discrimination of languages we recorded two proficient bilingual speakers: a male French-English bilingual speaker and a female Spanish-Catalan bilingual speaker. The use of bilingual speakers for the test data aims at reducing any sources of variability not due to the target languages. During each recording session, the speakers read the first two chapters of the book *The Little Prince* in one of their languages, and immediately afterwards they were asked to discuss what they had read. This procedure was then repeated for their second language. All recordings for each speaker were done on a single session.

The audio recordings were semi-automatically segmented into utterances with a 300 ms silence threshold using the speech analysis software *Praat* (Boersma & Weenink, 2014), and subsequently downsampled to 16kHz. The resulting dataset is composed of 319 utterances (French: 65, English: 75, Spanish: 99, Catalan: 80), with a mean length of 3.69 seconds ($min = 2.00$ s, $max = 10.63$ s).

Model of habituation task

Experimental studies of language discrimination in infants use an habituation paradigm (Mehler et al., 1988; Nazzi et al., 1998). In this paradigm, infants are presented with a set of stimuli from one language (L1), and their arousal is measured (in newborns, it is measured with a pacifier connected to a pressure detector). After an initial increase, infants' arousal decreases, indicating habituation. When a threshold has been reached, half of the infants continue with the same class of stimuli, and the other half are switched to a second class (L2). The difference of arousal after the switch in the two groups is used as a measure of discrimination.

Here, we will model this paradigm using an on-line clustering algorithm. In the habituation phase, the system gradually incorporates data from one language (L1) until it reaches a statistical threshold. In the test phase, as for infants, new utterances of L1 (*same* condition) and L2 (*switch* condition) are compared to the habituated model. The input of this model consists of the i-vectors of the test utterances as extracted by our previously trained system. To reduce spurious effects caused by specific subsets of utterances, the habituation task was run 100 times for each language pair using randomly selected subsets in each trial.

Habituation phase The model starts with an initial set of 10 i-vectors $\{v_1, ..., v_{10}\}$ of one language (L1) chosen randomly from our dataset. Firstly, the centroid μ_1 of this initial set (i.e., the mean i-vector) is computed, and the cosine distance of each of the 10 composing vectors to the centroid $d_c(v_i, \mu_1)$ is calculated. Secondly, a new random set of 10 i-vectors $\{v_{11}, ..., v_{20}\}$ of the same language L1 is selected, and their cosine distances to the initial centroid μ_1 are calculated. The distribution of distances of the initial and the second set of vectors are then compared with a t-test.

If $p \leq 0.05$, the two distributions are considered statistically different, that is, the model perceives a difference between the two sets of utterances, and therefore has not yet reached habituation. In this case, the last set of vectors is aggregated to the initial set and the centroid is recalculated, μ_2, as the mean i-vector of the whole set. Following the same procedure, a new group of 10 i-vectors from L1 is selected and their cosine distance to the new centroid $d_c(v_i, \mu_2)$, $\{i = 21, ..., 30\}$, are calculated and compared through a t-test to the distance of the previous 10 vectors to the new centroid $d_c(v_i, \mu_2)$, $\{i = 11, ..., 20\}$. This procedure is repeated as long as $p \leq 0.05$.

When $p > 0.05$ (defined as our saturation threshold), the two distributions are not statistically different and the habituation phase is therefore complete. As a final step, the last group of vectors is aggregated to the previous set and a final centroid is obtained, μ_F. The distance of the last 10 vectors to μ_F is then calculated and retained for the test.

Test phase In this stage, a new set of 10 i-vectors v_i is randomly selected from the same language used in habituation (L1, *same* condition) as well as 10 i-vectors u_j from the second language of the same bilingual speaker (L2, *switch* condition). For each set of vectors, the cosine distance to μ_F is calculated.

We finally perform two t-tests, one per condition, comparing the distribution of distances of the new vectors of L1 or L2 to the distribution of the last 10 habituation vectors. In the *same* condition, as the new utterances belong to the same language as those in habituation, the p-value of the t-test is expected to remain above the saturation threshold, $p > 0.05$. On the other hand, in the *switch* condition, the p-value will depend on the overlap between the distribution of the habituation (L1) and L2: a p-value below the 0.05 threshold would mean that the two distributions are significantly different, indicating discrimination of the two languages, while $p > 0.05$ would indicate a lack of discrimination.

Discriminability measures

To quantify the discriminability of the languages independently of our habituation-dishabituation model, we computed the pairwise ABX discrimination score, a nonparamet-

ric measure of category overlap. It consists in taking all possible ABX triplets of utterances from a language pair, where A corresponds to an utterance of L1, B corresponds to an utterance of L2, and X can be either L1 or L2. For each triplet, X is classified as belonging to L1 or L2 based on whether the cosine distance between X and A is smaller or greater than the distance between X and B. The percentage of correct classifications serves as an index of the discriminability between the two languages. Additionally, we performed a *Principal Component Analysis* (PCA) for each language pair as a way of visualizing the variance and distance of the i-vectors that compose each language.

Results

Habituation task

We ran the habituation model for both language pairs, and within each pair we tested the model with both possible languages in the initial habituation phase. The average amount of steps to reach habituation was similar for all languages (French: 2.1, English: 1.8, Spanish: 1.7, Catalan: 1.7).

As previously observed in infant experiments, the results of 100 trials for each test (presented in Figure 1) show a difference in the pattern of discrimination of the two language pairs. In the case of Spanish-Catalan (bottom panels), the p-values of both the *same* condition and the *switch* condition are significantly above the threshold value of $p = 0.05$, independently of the language presented in habituation (Habituation:Spanish -bottom right panel- *same*: $M = 0.48$, $SD = 0.26$, *switch*: $M = 0.40$, $SD = 0.26$; Habituation:Catalan -bottom left panel- *same*: $M = 0.52$, $SD = 0.28$, *switch*: $M = 0.54$, $SD = 0.27$), suggesting a lack of discrimination of these two languages. On the other hand, the second language pair (French-English, top panels) presented an asymmetry in the responses of the model to the *switch* condition, depending on the language of habituation. When the system is habituated to English as L1 and then switches to French (top left panel), the two languages are discriminated as indicated by a decrease of the p-value below the threshold in the *switch* condition (*same*: $M = 0.49$, $SD = 0.29$, *switch*: $M = 0.012$, $SD = 0.026$). However, if the system is initially habituated to French (top right panel), the switch to English is not detected, with both conditions showing similar p-values (*same*: $M = 0.54$, $SD = 0.29$, *switch*: $M = 0.48$, $SD = 0.25$). While a similar behavior was observed in infant habituation experiments (Mehler et al., 1988), additional analyses are required to understand this asymmetry.

ABX and Principal Component Analysis

To further explore the different response patterns of our model, we performed an ABX task for both language pairs and all possible X categories (ABA, ABB). The results of this test, shown in Table 1, present a similar pattern to the one observed in the habituation task. In the case of Spanish-Catalan, both ABA and ABB trials presented scores slightly above chance level (50%), meaning that nearly half of the Spanish

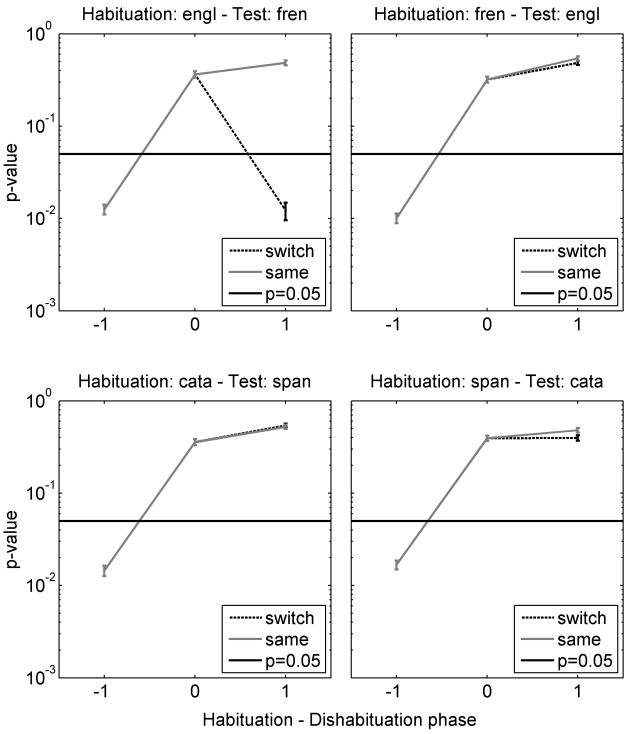


Figure 1: Average p-values over 100 trials of the habituation task for French-English discrimination (top) and Spanish-Catalan discrimination (bottom). The *x* axis represents the steps of the habituation and test phase, where *0* indicates the step where the habituation threshold ($p = 0.05$) was reached. Accordingly, *step -1* represents one step before habituation, and *step 1* represents the test (dishabituation) phase.

utterances were incorrectly categorized as Catalan utterances (and vice-versa). On the other hand, French-English trials presented an asymmetry: a majority of English utterances were correctly classified, while the classification of French utterances remained near chance level. This means that the distance between two given French utterances in the test set is often larger than the distance between a French and an English utterance, pointing at a possible imbalance in the variance of the distributions of their i-vector representations.

Table 1: Summary of ABX results: Percentage of accuracy for the distant language pair (A = English, B = French) and the close language pair (A = Catalan, B = Spanish).

Language Pair	X=A	X=B
English (A) - French (B)	76%	46%
Catalan (A) - Spanish (B)	51%	57%

Finally, we performed a Principal Component Analysis on both language pairs in order to visualize the distribution of the utterances. A representation of the first two dimensions

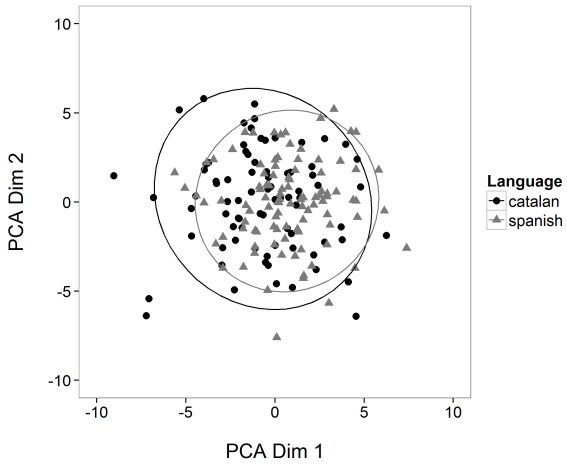


Figure 2: First two dimensions obtained through a Principal Component Analysis of the i-vectors of Spanish and Catalan utterances spoken by a bilingual speaker.

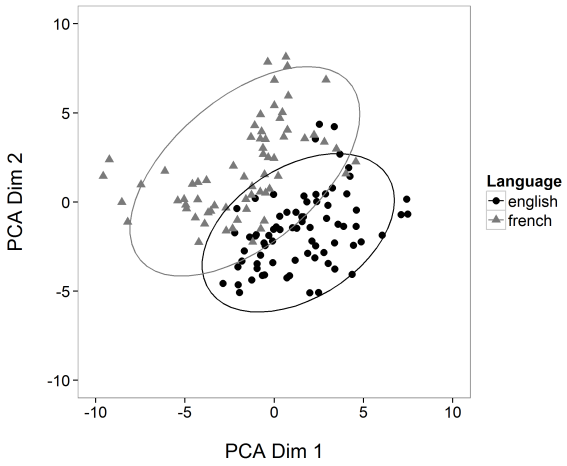


Figure 3: First two dimensions obtained through a Principal Component Analysis of the i-vectors of French and English utterances spoken by a bilingual speaker.

of the PCA for the Spanish-Catalan pair, shown in Figure 2, revealed a high degree of overlap in the distribution of the utterances of these two languages. On the contrary, the first two dimensions of the French-English PCA, presented in Figure 3, show a higher separation between the two languages. However, as suggested by the ABX score, the variance in these two dimensions appears to be larger within French utterances than within English utterances.

Together with the ABX results, this difference in the variance may explain the asymmetry observed in the habituation task: when the model is habituated to English, the variance of the i-vectors that are aggregated during this initial phase remains small, allowing the system to detect a switch to the second language. In other words, the within-language distance distribution is smaller than across-language. However in the inverse case, when the model is initialized with French, the variance of the habituation vectors is relatively large and therefore the switch to English remains unnoticed.

In summary, we found an overall difference in the degree of separation of the i-vectors of both language pairs, which reflected in the behavior of our habituation-dishabituation model. Spanish-Catalan utterances present largely overlapping distributions, causing a lack of discrimination in the habituation task, while French-English utterances have less overlapping yet more asymmetrical distributions, producing an equally asymmetric response of the system.

Discussion

In this paper we introduced a novel application of speech technology tools to model language discrimination in infants. Using a GMM-UBM trained on a small dataset of French utterances, we represented the acoustic space of a monolingual infant after a brief exposure to this language. To test the system's ability to discriminate languages, we modelled the acoustic representation of novel utterances as a pattern of shifts from the means of the UBM. Using this low-dimensional representation, called i-vector, we constructed a model of a habituation task similar to the experimental paradigm often used to test discrimination in infants.

The behavior of our model in the habituation task resembled that observed in previous experiments: our system, pre-exposed to French, was unable to discriminate between two previously unheard languages with highly similar phonologies (Spanish & Catalan), while it discriminated two phonologically distant languages (French & English). Interestingly, just as reported in previous infant studies such as Mehler et al. (1988), the ability to discriminate between French and English depended on the language presented during habituation. When the system was initially habituated to the previously unheard language (English), it detected a switch to the "native" language (French), but it failed at discriminating a switch to English when French was presented in habituation. Further analyses provided a potential explanation for our model's asymmetrical behavior: the variance of the i-vector representations of French utterances is larger than that of English utterances, causing the habituation model to create a broad category for French which hinders the discrimination of English. While in the context of infant studies this asymmetry was regarded as a preference for the native language, our modeling results suggest that the perceived acoustic variability might be responsible for this behavior, providing a new perspective on this issue.

There are three possible explanations for the larger variance of French as compared to English in our test data. First of all, this difference might be a characteristic of the specific bilingual speaker that was recorded for this experiment. To test this hypothesis, it would be necessary to repeat the test with a different French-English speaker. If the same pattern

was observed, it would indicate that the difference does not lie in the speaker but in the language. This could mean that, overall, French speech is acoustically more variable than English. However, and more interestingly, it is also possible that the difference was originated in the training of the Universal Background Model and the Total Variability subspace: as our system was pre-exposed only to French, the model may have developed a larger sensitivity to acoustic differences present in French speech than those in English speech, thus appearing more variable. To discern these two possibilities, the model could be re-trained using English as the background (i.e., "native") language. If the larger variance is due to the sensitivity of the model to its native language, then the asymmetry should be inverted. The results of these future modeling experiments may help us better understand the behavior observed in infants.

In addition, this methodology can be applied to model language discrimination in a variety of other cases. First, the UBM and the TV subspace can be trained with different languages and with varying amounts of data to investigate the impact of language exposure on discrimination (e.g., the model can be trained with a large dataset of Spanish speech and then tested on its ability to discriminate Spanish from Catalan). Second, the system could be trained with a bilingual background to study how multilingualism affects the construction of the acoustic space and consequently its ability to discriminate languages. This bilingual background can be composed of either monolingual speakers of two languages or bilingual speakers, giving further insight into the impact of different bilingual environments on the perceptual system. Third, the acoustic features provided to the model can be adapted (for example, by using filtered speech, or adding additional prosodic information to the feature vectors) to explore the role of different cues in language discrimination. The experimental data available to date provides a means of evaluation for the models, which in turn may generate new testable hypotheses that will help us better understand how young infants achieve this task.

Acknowledgements

We thank Alexander Martin and Laia Fibla for their participation in the recordings, and Hynek Heřmanský and Lukáš Burget for their helpful discussions. This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the École des Neurosciences de Paris Ile-de-France, the Region Ile de France (DIM Cerveau et Pensée), and an AWS in Education Research Grant award.

References

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID - Corpus of Interactional Data - Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, *49*(3), 1–30.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer [Computer program]*. Retrieved from `http://www.praat.org` (Version 5.3.86)

Bosch, L., & Sebastian-Galles, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy*, *2*(1), 29–49.

Byers-Heinlein, K., Burns, T. C., & Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychological Science*, *21*(3), 343–348.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Martínez, D. G., Plchot, O., Burget, L., Glembek, O., & Matějka, P. (2011). Language recognition in ivectors space. In *Proceedings of interspeech 2011* (Vol. 2011, pp. 861–864). International Speech Communication Association.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

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.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology*, *24*(3), 756–766.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by english-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*, 1–19.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011, December). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, *2*(1), 85–115.

Saffran, J. R., Aslin, R. N., & Neport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., & Deller, J. (2002). Approaches to language identification using Gaussian Mixture Models and Shifted Delta Cepstral features. In *ICSLP 2002* (pp. 89–92).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., ... others (2006). The HTK book (for HTK version 3.4).

2.3.2 Experiment 2: Generalizing to other language pairs

In our first study we explored the ability of an i-vector based model trained on a small corpus of French speech to discriminate two pairs of languages. While the observed results were in line with what has been observed in infants (that is, language discrimination was harder for the close pair Spanish-Catalan than for the distant pair French-English), the interpretation of the results may be limited by the fact that only two language pairs were tested, and one of them contained the language that the model had been trained on. In order to overcome this limitation and further explore the model's behavior when faced with different language pairs, we propose an extension of our experiment. Here, we will extract i-vectors for new bilingual speakers of 4 language pairs: French-English, Italian-English, Chinese-English and Dutch-English. Testing new utterances for French-English will allow us to evaluate the stability of the results observed in our first study. Furthermore, the inclusion of a close-distance pair (Dutch-English) and two additional language pairs with distant phonological and rhythmical properties (Italian-English & Chinese-English), none of which had been seen by the model during training, will allow us to discern the distance effect from a potential familiarity confound.

2.3.2.1 Materials

Our new test set was composed of bilingual speech from 4 male speakers from the *UCAM Bilingual Corpus* (EMIME project, http://www.emime.org). Bilinguals spoke French, Italian, Dutch or Chinese as native language, and English as a non-native language with a moderate foreign accent. For each speaker, utterances are composed of read speech in both of their languages. Table 2.1 shows the number of utterances per speaker and language. Recordings were down-sampled to 16 kHz. For the background model, we used the same UBM as trained in Experiment 1.

Table 2.1: Number of utterances in English and L2 for each speaker from the UCAM Bilingual Corpus.

Speaker	Native Language	N utt. (Eng.)	N utt. (Native Lang.)
s01	Dutch	130	130
s02	French	130	130
s03	Italian	130	130
s04	Chinese	89	89

2.3.2.2 Model parameters

Speech features were computed following the same parameters and procedure as described in Carbajal, Fér, and Dupoux (2016), that is, we calculated MFCC-F0-SDC features with a $7-1-3-7$ configuration using 25 ms windows and 10 ms shifts. For each utterance, we extracted 200-dimensional i-vectors using the same UBM and TV space trained with French speech in our previous study.

2.3.2.3 Results and discussion

First, in order to visualize the data, we computed a Principal Component Analysis (PCA) on the distribution of i-vectors for each speaker. Figure 2.3 shows the first two dimensions of the PCA for each of the language pairs. From this visualization, the distribution of utterances for each language pair seems to be in agreement with the expected outcomes. That is, the language pair showing the highest amount of overlap seems to be Dutch-English, two languages that share rhythmic properties and a good part of their phonemic inventories, while the other three language pairs show fairly good separation.

To quantify the separation, we computed the ABX score for each language pair based on the cosine distance between i-vectors. The procedure to compute ABX is the same as explained in our previous study. Results are shown in Table 2.2. The ABX scores confirm a higher overlap between Dutch and English compared to the other language pairs. However, the score for this particular pair (60%) was not as low as that of the Spanish-Catalan samples analysed in Carbajal et al. (2016), with an average of just 54%. This may mean that Spanish and Catalan share greater similarities than Dutch and English (and are thus even harder to discriminate), but given that we only have one bilingual speaker of each pair it is not possible to draw conclusions.

Table 2.2: Mean ABX scores (% correct classifications averaged over ABA and ABB trials) for the 4 language pairs.

Language Pairs	ABX
French - English	71%
Italian - English	67%
Chinese - English	70%
Dutch - English	60%

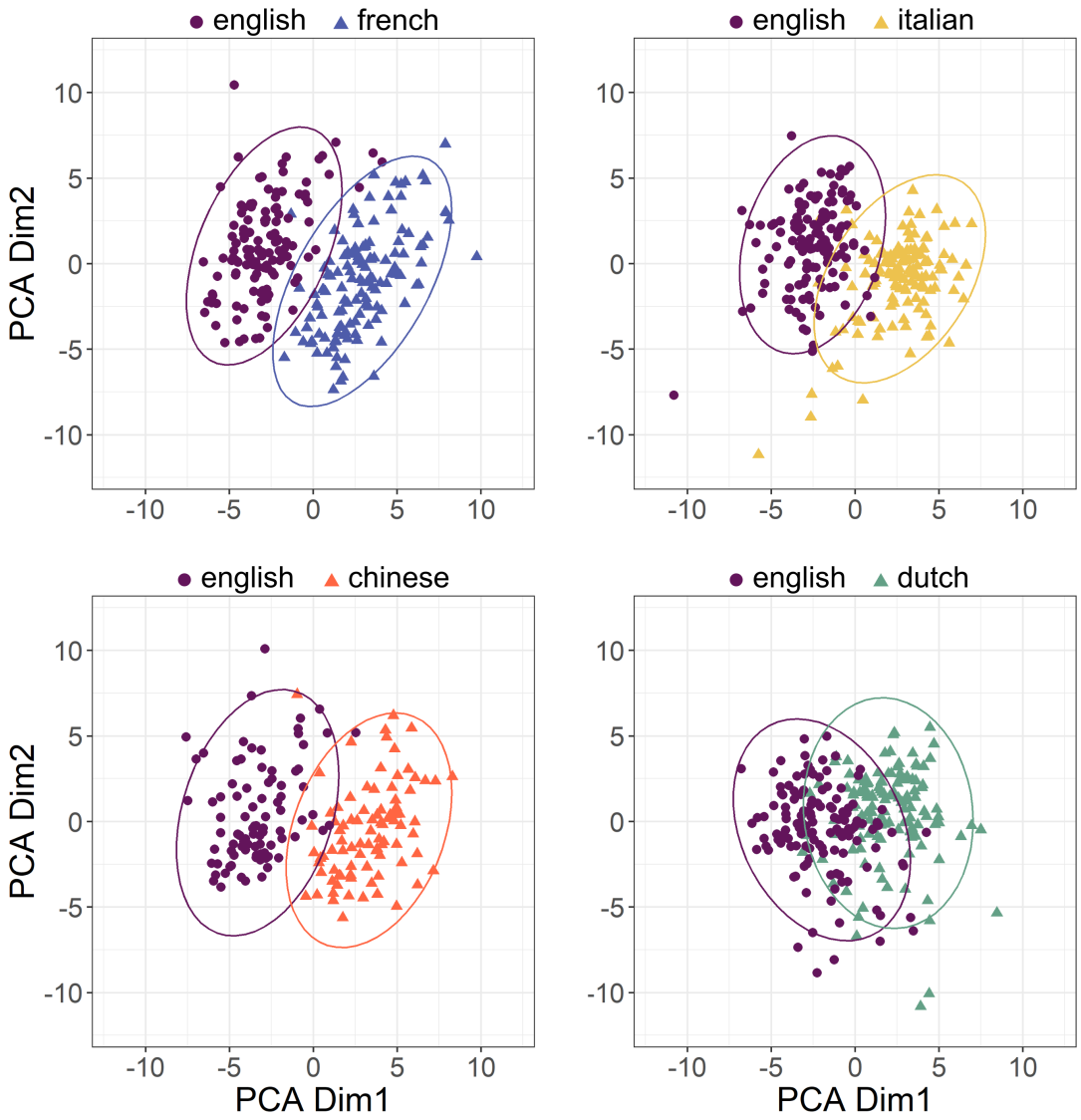


Figure 2.3: PCA visualization of utterances produced by 4 bilingual speakers from the UCAM corpus. Each dot represents one i-vector (i.e., one utterance).

Overall, these results – together with those from our first study – seem to confirm an effect of language distance on the degree of separation of the language pairs. Furthermore, we showed that the ability to separate languages that are both rhythmically and phonologically distant does not depend on the language that the model was trained on, as we have found discrimination for English-Italian and English-Chinese samples with a model trained on French. Interestingly, unlike in our first study, we did not find evidence of an ABX asymmetry for French-English samples (ABA: 69%, ABB: 72%). This may mean that the speech samples used in our first study (which were composed of a mix of read and casual speech) were more variable than the current samples (composed of read speech only). If this is the case, then it suggests that differences in speech register are an important source of variability, and their impact on language separation should be further investigated.

So far we performed separate tests per speaker, evaluating language discrimination within each bilingual's language pair. This was done in order to rule out an effect of speaker, but leaves open a question: is language distance within speaker larger than speaker distance within language? This question is of great importance to hypothesize whether infants would be able to easily group their bilingual input by language and not by speaker. In Figure 2.4 we visualize the first two PCA dimensions of the i-vectors of all 4 language pairs. From this visualization, it seems that the distance between speakers is generally larger than the distance between each speaker's two languages. Nonetheless, the language all speakers shared in common (English) appears to "pull" towards each other. This suggests that, for each speaker, their English samples are closer to other speakers' English samples than to other speakers' L2 samples.

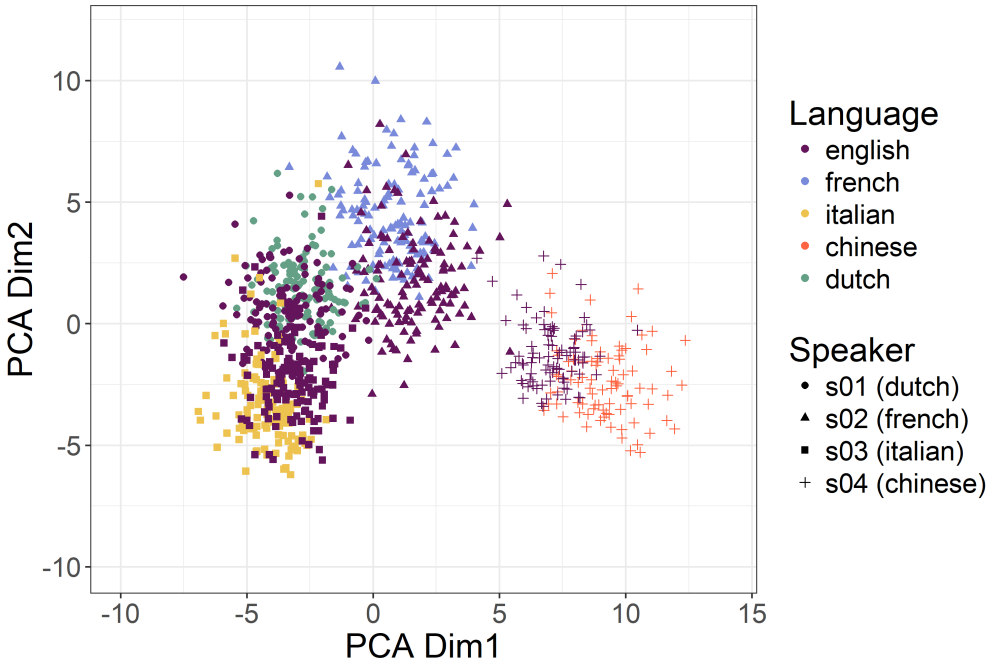


Figure 2.4: PCA visualization of all utterances produced by 4 bilingual speakers. Color represents the language, shape represents the speaker.

In order to confirm these two observations, we computed the average cosine distance between i-vectors v of each *speaker*×*language* subset: $D = \mathsf{avg}\, d(v_m, v_n)$. More specifically, for all speakers s_i and language categories $\{\mathrm{EN}, \mathrm{L2}\}$ (where EN refers to English and L2 refers to the second language spoken by each speaker), we computed:

(a) $D_a(i)$, within-speaker distance across their two languages ($v_m \in \{s_i, \mathrm{EN}\}$ and $v_n \in \{s_i, \mathrm{L2}\}$).

(b) $D_b(i, j)$, within-language (English) distance across speakers ($v_m \in \{s_i, \mathrm{EN}\}$ and $v_n \in \{s_j, \mathrm{EN}\}$, $i \neq j$).

(c) $D_c(i, j)$, across-speaker and across-language distance ($v_m \in \{s_i, \text{EN}\}$ and $v_n \in \{s_j, \text{L2}\}$, $i \neq j$).

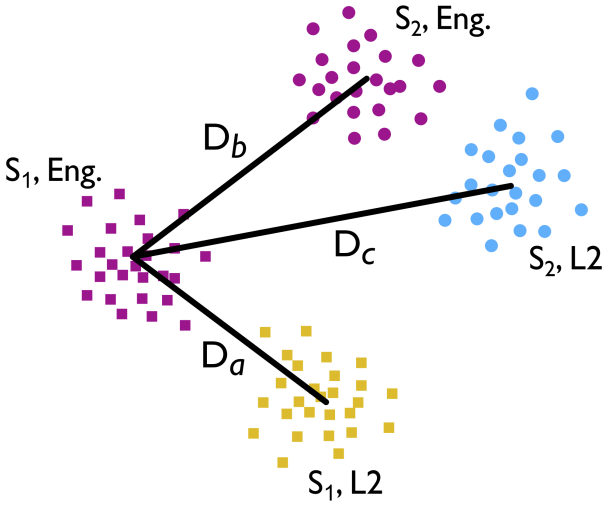


Figure 2.5: Example of distance measures across speakers and languages.

An example of these measures is shown in Figure 2.5. First, we tested whether language distance within speaker (D_a) is indeed smaller than within-language distance across speakers (D_b). The distribution of distances had means $M_a = 0.45$ ($SD = 0.04$) and $M_b = 0.53$ ($SD = 0.05$), respectively. A paired-samples t-test confirmed a difference between these two distances, $t(11) = 4.25$, $p = 0.001$, indicating that the English i-vectors of a given speaker are more similar to the same speaker's L2 than to other English vectors from other speakers. Second, we analysed whether English samples from different speakers (D_b) were closer to each other than to other speakers' L2 (D_c). The distribution of distances in the latter case had mean $M_c = 0.55$ ($SD = 0.06$). A paired-samples t-test comparing the distributions of D_b and D_c showed that, across speakers, English samples were closer to each other than to other languages, $t(11) = 3.95$, $p = 0.002$.

On the one hand, these results suggest that distance between speakers could potentially obscure the presence of multiple languages. If infants were to search for the most likely separation of their input, they could end up collapsing all speech by a single speaker, as their distance is smaller than the distance between speakers. On the other hand, although speakers voices are the most prominent difference between speech samples, speech belonging to a common language is still to an extent grouped together in the i-vector space. Whether infants could use alternative strategies to pick up on this similarity should be investigated in future work.

2.3.3 Experiment 3: Language discrimination with filtered speech

As we have seen in Chapter 1, many language discrimination experiments used filtered speech, showing that infants could use prosodic information in the signal to separate the languages. Here, we are interested in evaluating how filtering the speech samples may affect the separation of languages in the i-vector space. Based on infant experiments, if our model is sensitive to prosodic information, we would expect it to retain its ability to discriminate very distant language pairs, and to fail with very close language pairs. If, on the other hand, the model was relying mostly on fine phonological information, then it may be unable to discriminate any of the pairs. In order to investigate this, we will low-pass filter the bilingual speech samples used in Experiment 2, and run the same pipeline as presented before, based on the UBM and TV space trained in Carbajal et al. (2016). This will allow us to analyse whether the dimensions that our model has captured during training contain sufficient prosodic information to separate the languages.

2.3.3.1 Materials

Our test set was composed of the same utterances presented in Experiment 2 (4 male speakers from the *UCAM Bilingual Corpus*). As in infant experiments, we low-pass filtered the speech samples to 400 Hz and normalized the intensity to 70 dB using the software *Praat* (Boersma & Weenink, 2014). We kept a sampling rate of 16 kHz. For the background model, we used the same UBM as in Experiments 1 & 2, trained on French casual speech.

2.3.3.2 Model parameters

Feature extraction followed the same pipeline as described in Experiments 1 and 2. For each utterance, we extracted 200-dimensional i-vectors using the same UBM and TV space trained with French speech in Carbajal et al. (2016).

2.3.3.3 Results and discussion

As in previous experiments, we first computed a PCA on the i-vectors for each language pair in order to visualize the data. Figure 2.6 shows the first two dimensions of the PCA for each pair. As can be seen in the visualization, all language pairs appear to show a higher degree of overlap compared to the

unfiltered speech, suggesting that the model was relying to some extent on phonological differences. However, with the exception of English-Dutch, most language pairs seem to be still discriminable to a certain extent. In order to quantify their degree of separation we computed the ABX scores for each language pair. The results are shown in Table 2.3.

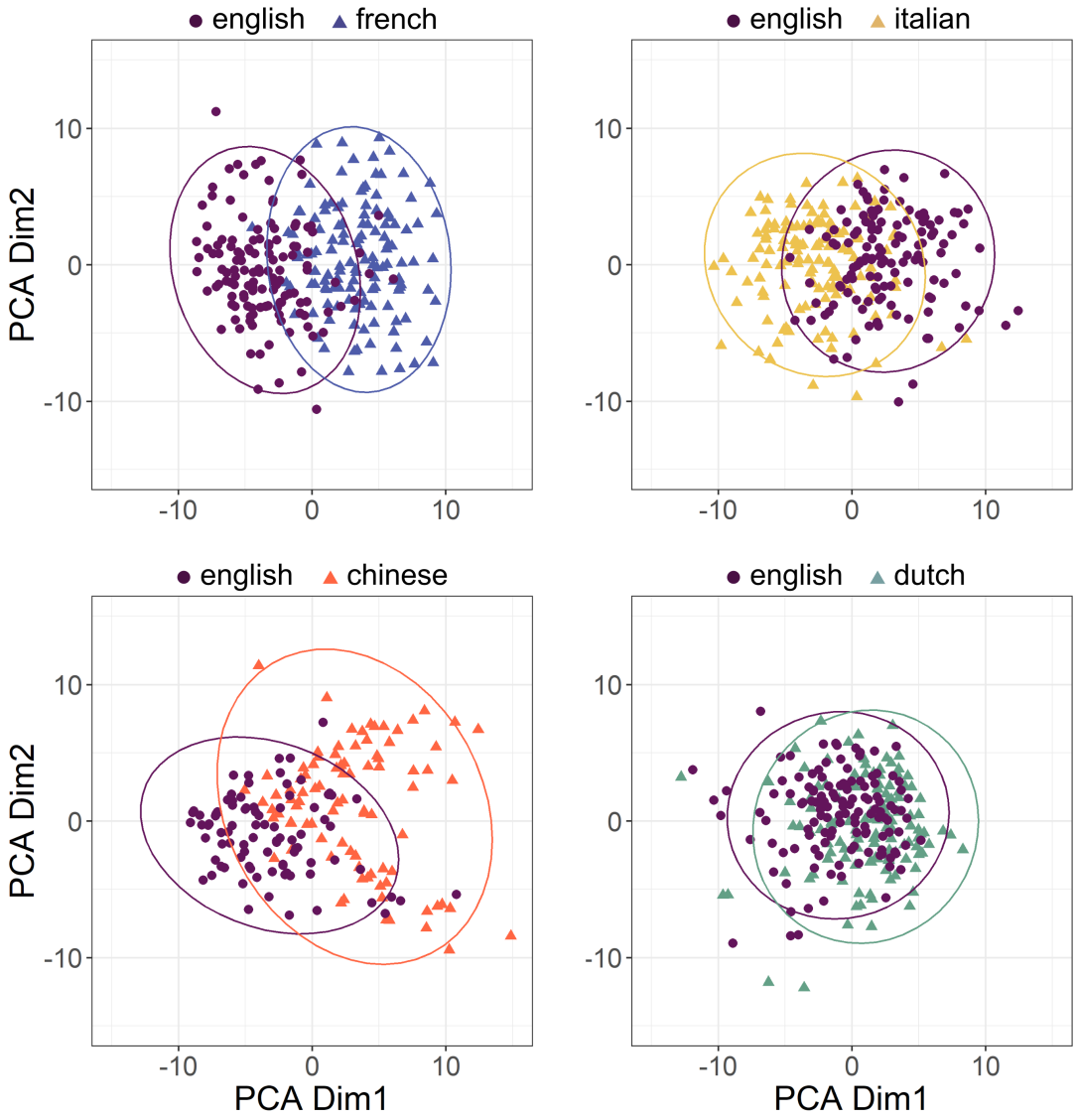


Figure 2.6: PCA visualization of low-pass filtered utterances produced by 4 bilingual speakers from the UCAM corpus. Each dot represents one i-vector (i.e., one utterance).

The ABX results confirm our observations. That is, with the exception of French-English, all languages show a small reduction of their discriminability (i.e., they have lower ABX scores compared to the ABX scores of unfiltered speech). In particular, the ABX score of the filtered speech samples in Dutch-English is as low as that of the unfiltered samples in Spanish-Catalan. In spite of the reduction in disciminability, these results seem to indicate that our i-vector model has been able to capture sufficient prosodic information to detect differences in very distant language pairs, even when the

speech is low-pass filtered.

Table 2.3: Mean ABX scores (% correct classifications averaged over ABA and ABB trials) for the 4 language pairs (filtered speech).

Language Pairs	ABX
French - English	72%
Italian - English	64%
Chinese - English	67%
Dutch - English	55%

In summary, in these two experiments we have extended the results from our first study, showing that language discriminability (based on ABX measures) is higher for distant language pairs than for pairs that have highly overlapping phonological and prosodic features. Furthermore, we showed that language discrimination of distant pairs does not require previous exposure to either language, as our model had been trained only on French and was able to discriminate, for instance, English from Chinese. Finally, we observed that discriminability was reduced by filtering speech, but remained above 60% for all distant language pairs, and almost at chance level for a very closely related language pair, English-Dutch (55%).

2.3.4 Experiment 4: Language discrimination across speakers

So far, we conducted experiments using speech from one bilingual speaker of each language pair. Would language discrimination still be possible in a context with multiple speakers of each language? In order to answer this question, we propose to replicate the computational experiments presented by Ramus et al. (1999), in which speech from multiple speakers of 8 different languages was used in simulations of the infant habituation task. To compare the models on similar grounds, we will replicate their habituation algorithm as closely as possible, while using i-vectors as input measure instead of $\%V$. An advantage of using i-vectors over previous cognitive models of language discrimination is that we can test both normal and filtered speech. We thus propose to run the experiments using both kinds of input.

2.3.4.1 Model of the habituation task

In the simulation proposed by Ramus et al. (1999), 10 utterances from 2 speakers of one language (5 per speaker) were used for habituation. In the test phase, 10 new utterances from either 2 new speakers of the same language (control) or 2 speakers of a different language (experimental/switch) were presented. At each step (i.e., utterance) in the simulation, they defined arousal A_n as the absolute difference between the current $\%V$ and the average of this value in all previous steps (P_{n-1}), $A_n = |V_n - P_{n-1}|$. As we have multidimensional i-vectors, we will compute the cosine distance between each vector at step n (X_n) and the centroid (i.e., average vector) of the i-vectors[6] from step 1 to $n-1$.

As proposed in Ramus et al. (1999), 40 different subjects (20 control, 20 experimental) will be simulated by randomly selecting the speakers and utterances for habituation and test. Finally, the p-values of a statistical test (Mann-Whitney signed-rank test, as used in Ramus et al., 1999) will be used to compare the arousal in test utterances under the two conditions (control vs experimental).

2.3.4.2 Materials

To keep the studies as similar as possible, we used the same speech corpus as used in Ramus et al. (1999). This corpus is composed of read speech from 4 female speakers of each of the following languages: Catalan, Dutch, English, French, Italian, Polish, Spanish, and Japanese. Table 2.4 shows the total number of utterances available per language (counting all 4 speakers in each). However, unlike Ramus et al.'s study, we did not select a subset of utterances matched by length and number of syllables. Instead, we used the whole set of possible utterances, within which the ones used in that study are contained. As in our previous studies, the recordings had a sampling rate of 16 kHz. We further constructed a filtered speech corpus by applying a low-pass filter with threshold at 400 Hz (and intensity normalization at 70 dB) on each utterance of the original test corpus using the software *Praat* (Boersma & Weenink, 2014).

As background model, we used the French UBM used in Experiments 1 - 3.

[6]This simulation is similar to the one we presented in Carbajal et al. (2016). The main difference relies in that, in our previous study, we had computed these values only every 10 utterances, while here it is computed at each step.

Table 2.4: Number of utterances available per language.

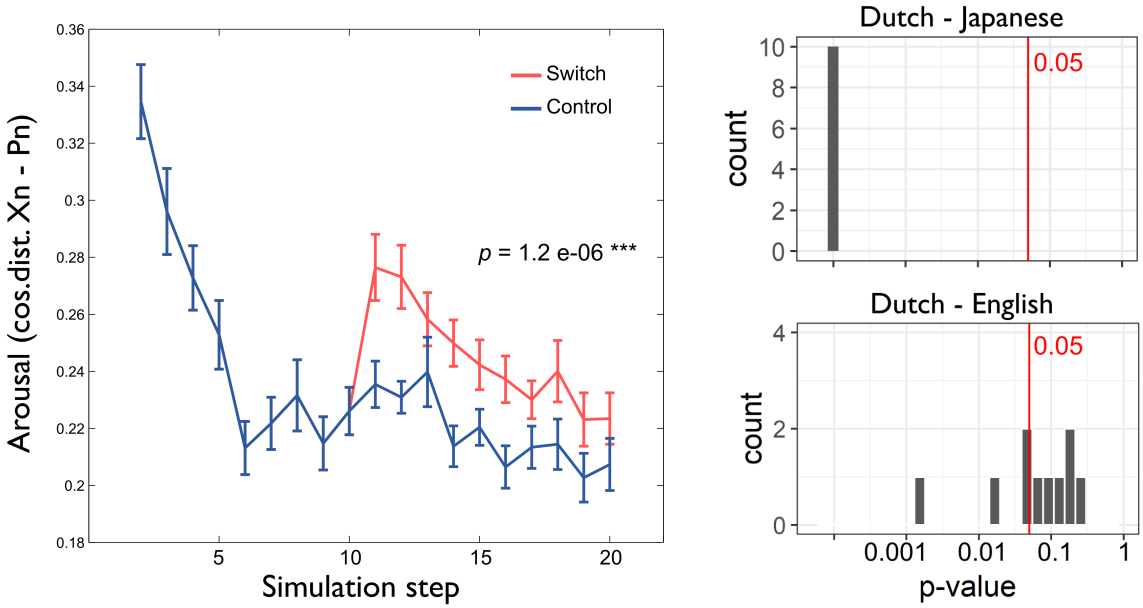
Language	N	Language	N
Catalan	216	Italian	215
Dutch	228	Polish	216
English	161	Spanish	212
French	216	Japanese	212

2.3.4.3 Model parameters

Feature extraction followed the same pipeline as described in our previous studies. As before, for each utterance, we extracted 200-dimensional i-vectors using the same UBM and TV space trained with French speech in Carbajal et al. (2016).

2.3.4.4 Results and discussion

As in Ramus et al. (1999), for each language pair available in the corpus, we simulated an experiment with 40 subjects (20 control and 20 experimental). An example of a simulated experiment is given in Figure 2.7a. Ramus et al. (1999) interpreted the p-value of control vs. experimental conditions as an indicator of the discriminability of each language pair, with $p < 0.05$ indicating a significant difference between both groups (thus implying discrimination of the language pair), and $p \geq 0.05$ meaning no significant group difference was found (i.e., no language discrimination). When running our replication, we first observed that repeating the simulation of 40 subjects several times sometimes changed the interpretation of the discriminability for certain pairs, with p-values going from significant to non-significant (see an example of a stable language pair, Dutch-Japanese, and a language pair with variable p-values, Dutch-English, in Figure 2.7b). These unstable results may be due to the large number of possible utterance sets, and to the fact that we did not control them for duration and syllable length. However, since Ramus et al. only reported one p-value computed over a single set of 40 simulated subjects, we cannot know how stable their results were. In order to overcome this problem while keeping our results comparable to Ramus' study, instead of incrementing the number of simulated subjects, we decided to repeat the simulation of 40 subjects 10 times. Thus, one p-value is obtained per simulated experiment, and the median of the 10 p-values is finally used as indicator of the discriminability of the language pair.



(a) Example of a simulated habituation experiment, using Dutch and Japanese samples.

(b) Distribution of p-values for 10 Dutch-Japanese and 10 Dutch-English simulations.

Figure 2.7: Examples of results obtained from simulations of the habituation experiment.

We will first discuss the results for filtered speech, as many behavioral experiments have used this manipulation (or alternatively a resynthesized signal with reduced phonetic variability) to test infants' language discrimination. Figure 2.8 shows a color-coded matrix with the median p-values obtained from the simulations of each language pair (right panel). A table with the numeric results can be found in the Appendix 2.C. For comparison, in the same figure (left panel) we show the results reported by Ramus et al. (1999). It should be noted that the matrices are symmetrical, that is, the upper half represents the same results as the lower half. Additionally, we have indicated with letters which language pairs have been tested on newborns using similar stimuli (that is, filtered or resynthesized). Experiments where discrimination was found are indicated with an asterisk.

As can be seen, within each traditional rhythmic class (syllable-timed languages marked with a red square on the top left corner; stressed-timed languages marked with a blue central square), most language pairs were not discriminated by our i-vector model. With the exception of Italian, which was discriminated from Spanish and Catalan, within-class experiments resulted in median p-values above 0.05. These results show overall agreement with Ramus et al. (1999), with only one language pair (Italian-Spanish) differing between the two models. On the other hand, the discrimination of language pairs across rhythmic class showed several differences. While both models found that Japanese and Italian are easily discriminated from other languages, the i-vector model failed to discriminate many

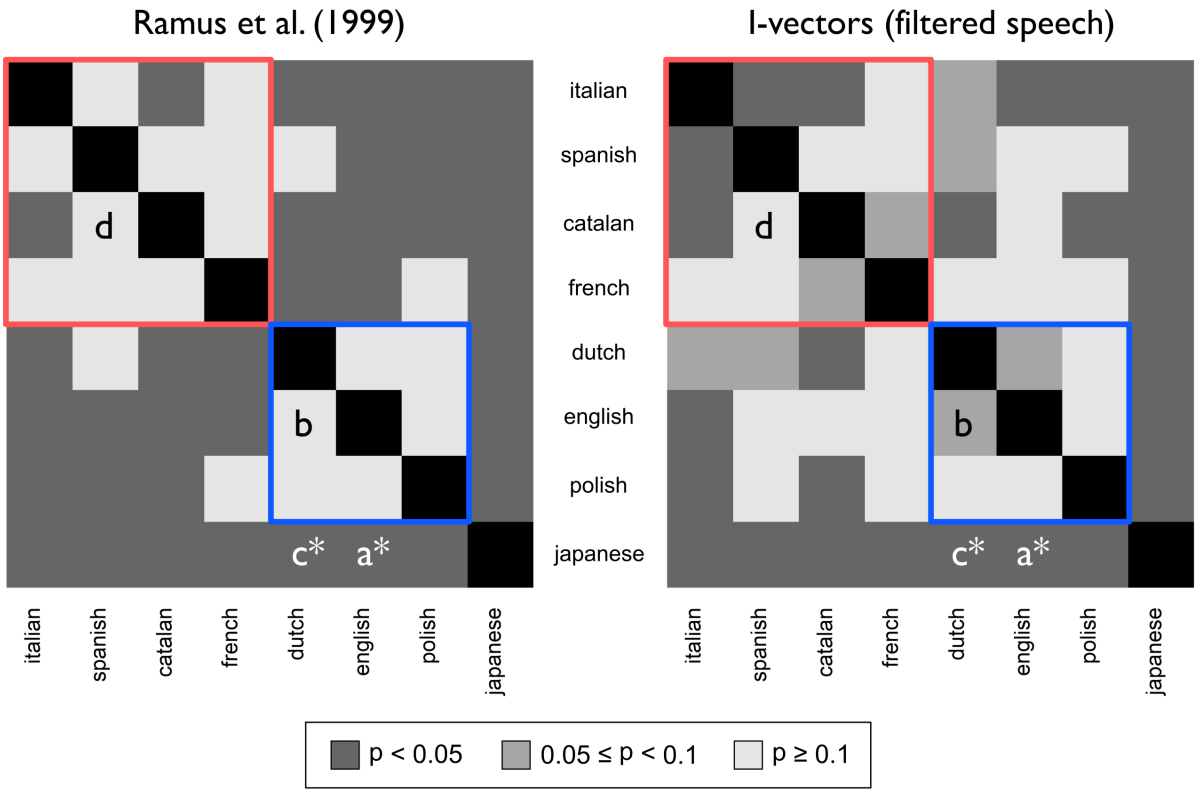


Figure 2.8: Color-coded matrix of p-values for each language pair, showing simulation results from Ramus et al. (1999) on left panel, and i-vectors using low-pass filtered speech on right panel. Red and blue squares indicate rhythmic classes (red: syllable-timed, blue: stress-timed). Pairs that have been tested in newborns on filtered or resynthesized speech (where only prosody and rough phonotactics were preserved) are shown with letters: (a) English-Japanese, filtered (Nazzi et al., 1998), (b) English-Dutch, filtered (Nazzi et al., 1998), (c) Dutch-Japanese, resynthesized (Ramus, 2002), (d) Spanish-Catalan, resynthesized (Ramus, unpublished). Asterisks mark experimental results where discrimination was found.

language pairs where Ramus et al.'s model had succeeded. Particularly, Spanish, French and English were rarely discriminated from other between-class languages. This difference could have been due to the larger variability of our test set compared to the subset used in Ramus et al.'s study. However, when repeating the simulations using only the utterances used by Ramus et al. (1999), our model still failed to discriminate these between-class pairs. It is thus possible that this failure was caused by the larger number of dimensions on which we computed the distance (that is, the 200 dimensions of the i-vectors, against a single dimension, $\%V$, used by Ramus et al.), which results in a higher overlap of the speech samples due to variability in other speech properties. Although these results seem to be in contradiction with the rhythmic class hypothesis, the particular language pairs that our model failed at have never been (to the best of our knowledge) tested in newborn infants. We therefore cannot rule them out as incorrect. Nevertheless, our model predicted the same results as Ramus et al. (1999) on all 4 language pairs that have been tested in newborns using filtered or resynthesized speech.

Next, we repeated the simulations using unfiltered speech. Figure 2.9 shows the matrix of results. A table with the numeric results can be found in the Appendix 2.C. Overall, cross-class language pairs showed an improvement in their discrimination with the addition of the full spectral information. Interestingly, Japanese – which was previously easily discriminated from any language – was confused with French, English and Dutch. While this may seem counter-intuitive, similar results were observed in infants tested on unfiltered speech spoken by multiple speakers in Japanese and Dutch (Ramus, 2002b). Moreover, the discriminability of languages within class gave similar results as with filtered speech: most language pairs showed no discrimination. The only difference from the previous results was in the stress-timed language group, where English remained confused with Dutch, but was discriminated from Polish.

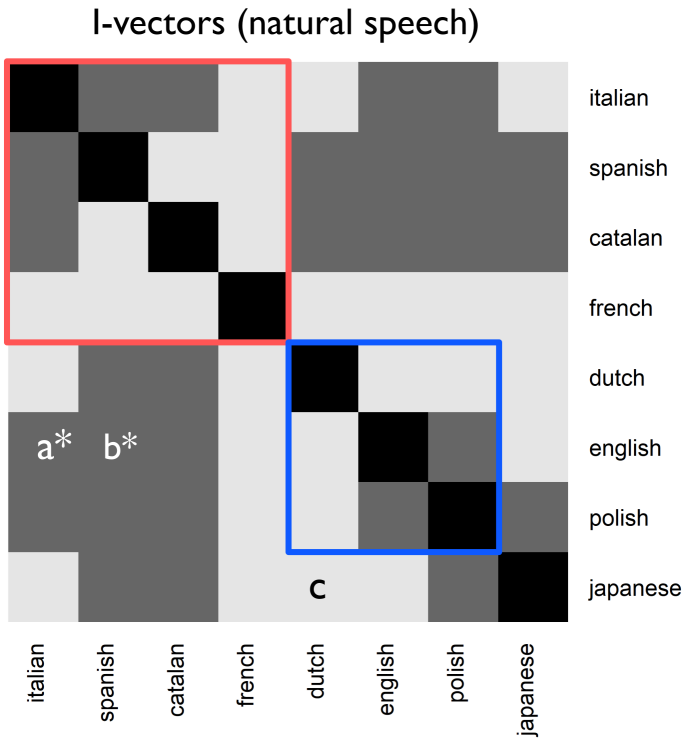


Figure 2.9: Color-coded matrix of median p-values for each language pair, showing simulation results for i-vector model using natural (unfiltered) speech. Red and blue squares indicate rhythmic classes (red: syllable-timed, blue: stress-timed). Pairs that have been tested in newborns on natural speech are shown with letters: (a) English-Italian (Mehler et al., 1988), (b) English-Spanish (Moon et al., 1993), (c) Dutch-Japanese (Ramus, 2002). Asterisks mark experimental results where discrimination was found.

A surprising result of this experiment is the confusion of French samples with all other languages. This confusion may have been caused by the familiarity of the model with this language. However, experimental results suggest that the opposite should occur: familiarity with a language, at least at a later age, usually results in better discrimination for that language against any other language, even

within class (Bosch & Sebastián-Gallés, 1997; Nazzi et al., 2000). Further computational experiments should thus be conducted with a larger background model to examine whether this confusion is reverted at a later stage.

In conclusion, these simulations showed similarities between the predictions made by Ramus et al.'s (1999) model and our i-vector model regarding the language pairs that have been tested in newborn infants with filtered or resynthesized speech, and more generally on the discriminability of within-class language pairs. Disagreements in other language pairs, especially in pairs across rhythmic class, remain to be resolved with further behavioral experiments. Furthermore, the i-vector model allowed us for the first time to make separate predictions on the discriminability of languages with natural vs filtered speech. Unlike previous psycholinguistic models that focused only on rhythmic properties, the i-vectors were able to capture additional acoustic properties of the unfiltered signal, showing agreement with experimental results suggesting that non-rhythmic speaker variability can hinder language discrimination, even across rhythmic class.

2.3.5 Experiment 5: The role of the background model

Experiments 1-4 were all based on the same background model, trained on approximately 25 minutes of casual speech from one single French speaker. While the results seem overall compatible with infant behavior, these may have been a peculiarity of this specific background model. In order to investigate the stability of the results, it is necessary to compare the i-vector representations learnt from different backgrounds. In this experiment, we propose to evaluate the role of the background model by training two new French and two English UBMs (and their respective Total Variability space). To test their effect on language separation, we will replicate the ABX tests of Experiment 2 with the i-vectors extracted from these new backgrounds.

2.3.5.1 Materials

To train the French background models, we selected two additional female French speakers from the same casual speech corpus as used in the background model of Carbajal et al. (2016), namely the *Corpus of Interactional Data* (Bertrand et al., 2008). For the English background models, we selected two female English speakers from the *Buckeye Corpus of Conversational Speech* (Pitt et al., 2007). Speech from each corpus was resampled at 16 kHz and segmented into utterances using Praat (Boersma

& Weenink, 2014), setting a silence threshold of 300 ms. Table 2.5 shows a summary of the number and duration of the utterances in each of these corpora. For the test, we used the same utterances from the 4 bilingual speakers presented in Experiment 2.

Table 2.5: Number and length of utterances in background model of Experiment 1 (Carbajal et al., 2016) and four new background models.

Model	Language	N utt.	Mean length (range)	Total
Carbajal et al. (2016)	French	602	2.54 s (0.43 s - 9.01 s)	25 min.
Background Model # 2	French	852	2.19 s (0.43 s - 9.81 s)	31 min.
Background Model # 3	French	546	2.51 s (0.43 s - 9.51 s)	23 min.
Background Model # 4	English	425	3.42 s (0.60 s - 20.57 s)	24 min.
Background Model # 5	English	336	5.07 s (0.60 s - 22.12 s)	28 min.

2.3.5.2 Model parameters

Feature extraction followed the same pipeline as described in our previous studies. For each background corpus, we trained a UBM with 256 Gaussian mixtures and a Total Variability space with 200 factors. Next, we extracted 200-dimensional i-vectors for each utterance in the test corpus using the UBM and Total Variability space of each background model.

2.3.5.3 Results and discussion

Figure 2.10 shows a comparison of the PCA visualization of the language pairs based on all 4 new background models. As can be seen, the new French models (top row) do not appear to differ significantly from the PCA obtained in Experiment 2 (Figure 2.4). The distribution of utterances based on the English models (bottom row) shows overall some resemblance with the one obtained with the French models. However, they seem to show less overlap between English utterances and the other languages, particularly so for the French-English pair.

We computed the ABX score for each of the 4 language pairs using the i-vectors extracted from the four new models. Figure 2.11 shows a comparison of all the models, including the one from Experiment 2. As can be seen, the results obtained for the three French background models are very similar, indicating that the models have learnt similar features from their input despite having been exposed to different speakers. On the other hand, the results from the English models show better performance in the Chinese-English and French-English language pairs, while they do not differ from

the French model in the Italian-English and Dutch-English pairs. This suggests that brief exposure to one language (here, English) is not sufficient to discriminate it from a language with many overlapping acoustic properties.

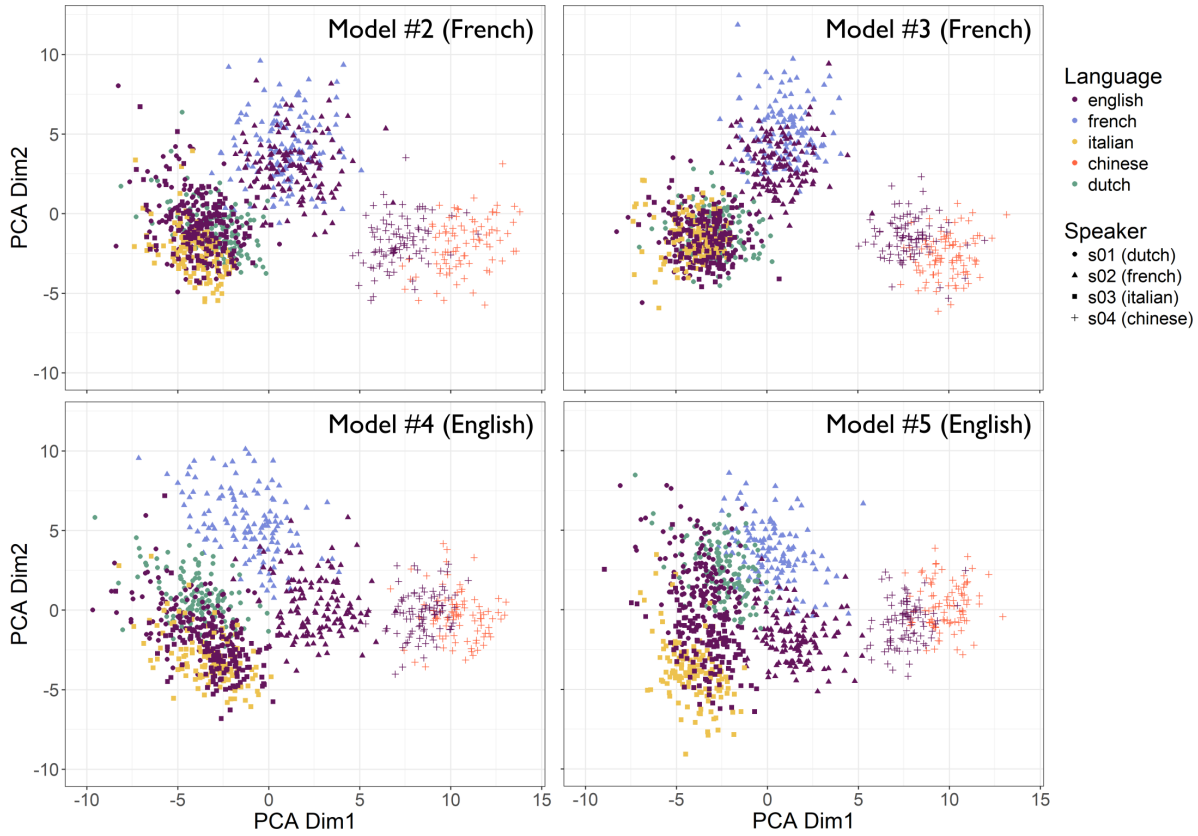


Figure 2.10: PCA visualization of the utterances of all 4 test language pairs, obtained with new background models (2 French, 2 English).

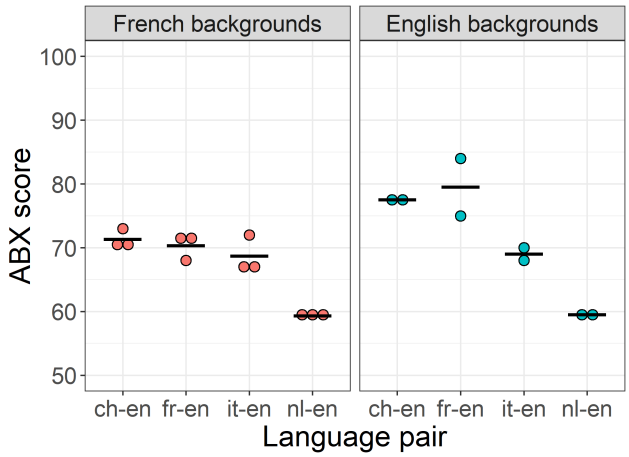


Figure 2.11: ABX scores for 4 language pairs, obtained with three different French background models (including the results from Exp. 2) and 2 English background models. Each dot represents the score from one background model. Horizontal lines represent the average score for a given language pair.

2.3.6 Experiment 6: The impact of a bilingual background (Article 2)

In the experiments presented so far, we have explored language discrimination based on monolingual backgrounds, showing that a brief exposure to one language is sufficient to discriminate many language pairs, as long as their acoustic properties are sufficiently distinct. Many infants, however, are born to bilingual families, hearing both languages regularly since birth. Would the exposure to two languages during background training affect language discrimination? To investigate this question we evaluate the separation of two languages with distinct acoustic properties (English and Xitsonga), comparing two monolingual background models that have been each exposed to one of the two languages, and a mixed model that has been exposed to both. Unlike in previous experiments, here we do not provide the model with prosodic information. Thus, these models can be taken to represent an extreme case where both languages are undistinguishable from their prosody.

This study was published as: Carbajal, M.J., Dawud, A., Thiollière, R. & Dupoux, E. (2016) The "language filter hypothesis": A feasibility study of language separation in infancy using unsupervised clustering of i-vectors. In *Proceedings of the 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-Epirob)*. Cergy-Pontoise, Paris, France.

2016 Joint IEEE International Conference on Development and Learning
and Epigenetic Robotics (ICDL-EpiRob)
Cergy-Pontoise, Paris, France, Sept 19-22, 2016

The "Language Filter" Hypothesis: A Feasibility Study of Language Separation in Infancy using Unsupervised Clustering of I-vectors

M. Julia Carbajal[1], Ahmad Dawud[1], Roland Thiollière[1] and Emmanuel Dupoux[1]

Abstract— **In order to avoid mixing up languages, infants immersed in a multilingual environment have to sort speech into language-homogeneous sets. To study the feasibility of this task, we use speech technology tools (Universal Background Models and i-vectors) in combination with unsupervised clustering to test language separation using speech from several speakers of two languages. We investigate the outcome of the clustering as a function of the variability of language experience (monolingual versus mixed background), and the availability of side information (speaker identity). Our findings show that in the absence of side information, language separation is relatively easier if the system has been pre-exposed to a single language (monolingual background), than it is for a system pre-exposed to both languages (mixed background). However, this initial disadvantage can be overcome by restricting the representation to the dimensions that most distinguish speakers using Linear Discriminant Analysis, suggesting that speaker identity side information may enhance language separation in a multilingual environment. The implications for language acquisition and computational modeling are discussed.**

I. Introduction

During their first years of life, human infants rapidly learn many linguistic properties of their native language (phonetics, phonology, prosody, lexicon, syntax, semantics) by merely being immersed in a language speaking community [1]. In spite of great advances in language development research, to this day relatively little is known about the computational mechanisms that they use to achieve this feat. One of the biggest unsolved puzzles is how language development can proceed smoothly even in situations where the environment contains more than one language, a scenario not uncommon across the globe [2]. To a large extent, bilingual infants achieve language developmental milestones following the same timeline as monolinguals [3], [4], [5], [6]. This is all the more remarkable considering that most learning mechanisms that have been proposed as the source of language acquisition are passive statistical mechanisms, such as tracking the distribution of sounds to learn the phonemes [7], tracking the transition probabilities for finding words [8], and tracking the patterns of repetition in order to acquire abstract rules [9], among others. Such statistical learning mechanisms indiscriminately accumulate sensory information. Therefore, if presented with a mixture of languages, they should either end up with a language chimera made up of the superposition of the language properties, or altogether fail to learn. Yet, even though in their early productions infants may appear to mix languages, this has been reanalyzed as a communicative optimization strategy in order to minimize production effort, rather than a true confusion between languages [5]. When tested in their comprehension, infants show a remarkable ability to discriminate languages from an early age [10], [11], [12].

Such an early ability has recently been vindicated within the PRIMIR framework (Processing Rich Information from Multidimensional Interaction Representations) [13], [14]. This framework proposes that infants apply their learning mechanisms not on raw data, but on the output of *dynamic filters* that allow them to focus their attention on the relevant dimensions and sections of the data. Additionally, PRIMIR incorporates the idea of a compare-and-contrast mechanism that helps infants to identify which sources of information go together and which should be kept apart. In this paper, we examine the proposition that infants are equipped with a mechanism which we call the 'language filter'. This mechanism (1) enables them to determine how many "language types" are spoken around them, (2) sort the utterances that they hear in terms of their language types, and (3) learn specific statistics for each language type. In other words, the child would not blindly accumulate statistics over all speech input, but rather actively separate the languages into types and acquire the linguistic properties of each type separately. Here, we approach the language filter hypothesis by quantifying the *feasibility* of language separation (subproblems (1) and (2)).

In order to do this, we propose the use of *i-vectors*, which represents utterances as a pattern of deviations from a background acoustic distribution, in combination with unsupervised clustering algorithms. Using databases of continuous speech from two languages —English and Xitsonga— we model three different background distributions: two monolingual backgrounds trained on speech from either English or Xitsonga, and one mixed or "bilingual" background trained on speech from both languages. We then extract the i-vectors corresponding to new utterances in both languages, and assess how well they cluster into language-homogeneous groups. Additionally, we explore the potential benefit of side information: we provide the system with speaker identity information and restrict the i-vector representations to the dimensions that most distinguish speakers using Linear Discriminant Analysis. Given that infants can plausibly discriminate the different speakers they interact with in their daily lives [15], [16], this type of information may serve as a cue to enhance language separation.

As with any clustering problem, one sensitive parameter is the number of clusters to be used. While too few clusters

[1]Département d'Etudes Cognitives, Ecole Normale Supérieure - PSL Research University, France. carbajal.mjulia@gmail.com, emmanuel.dupoux@gmail.com

978-1-5090-5069-7/16/$31.00 ©2016 IEEE

may result in impurities (i.e., clusters where several language types are mixed up), too many pure clusters may hamper language learning by fragmenting the learning materials and preventing within-language generalizations to occur. We explore this issue explicitly by evaluating cluster purity as a function of number of clusters.

The remainder of the paper is organized as follows. In Section 2, we describe the i-vector framework, the stimuli we use and the proposed analyses. In Section 3, we present and discuss the results, and in Section 4, we comment on the consequences of these results for theories of early language acquisition on the one hand, and for computational models of autonomous language learning on the other.

II. Methods

A. Universal Background Model and i-vectors

The i-vector paradigm was originally developed for automatic Speaker Recognition (SR) [17], and later adapted to Language Identification (LID) systems yielding excellent results [18], [19]. It consists in constructing a model of the acoustic space of speech, called the *Universal Background Model* (UBM), by fitting a Gaussian Mixture Model (GMM) over a distribution of speech features using an Expectation-Maximization algorithm [20]. These models are typically trained with several hundred hours of speech from many speakers and languages to capture all sources of variability.

After the UBM has been trained, any given utterance can be modeled as a deviation from the UBM by shifting the means of the GMM using Maximum a Posteriori (MAP) adaptations. These adaptations are usually restricted to a low-dimensional subspace —called the Total Variability space— that is assumed to cover all the important variability, and can be defined through factor analysis. Thus, based on the supervector (i.e. stacked vector) of UBM component means $\mathbf{m}$ and the Total Variability matrix $\mathbf{T}$ (a low-rank matrix that defines the bases for the subspace), the GMM supervector for any utterance or speech segment $\boldsymbol{\mu}$ can be modeled by:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{v}, \tag{1}$$

where $\mathbf{v}$ is a latent variable with standard normal prior. The MAP point estimate of $\mathbf{v}$ is called an *i-vector*, and can be used as a fixed-length vector representation of utterances irrespective of their duration.

In SR and LID systems, the extracted i-vectors of labelled data are typically used to train discriminative classifiers. Here, we will use the raw vectors as fixed length representations of each sentence, which will serve as input for unsupervised clustering algorithms. In contrast with traditional LID systems, we use a rather limited dataset for background training (around 90 minutes). We therefore restricted the number of Gaussian mixtures to 128 (typically more than 500 mixtures are used for LID), and 150 dimensions for the Total Variability subspace. To train our UBMs and extract i-vectors we used the MSR Identity Toolbox v1.0 [21].

B. Feature extraction

To represent the acoustic features of speech we used *Mel-Frequency Cepstral Coefficients* (MFCCs), which are traditionally used in speech processing and language identification systems. These features were calculated using the HTK Speech Recognition Toolkit [22] in 25 ms windows with a 10 ms shift. We retained the first 13 coefficients (including *C0*, which represents the energy).

C. Materials

We selected two speech data sets containing a large amount of speakers and utterances: the *TIMIT Acoustic-Phonetic Continuous Speech Corpus* of American English [23] and the *NCHLT Speech Corpus* of Xitsonga, a southern African language [24]. Both corpora are composed of read speech recorded from native speakers. From these corpora we selected different subsets of the speakers for training and test, as explained below.

1) UBM Training set: To evaluate the impact of language experience, we trained the UBM with three different subsets of the selected corpora to represent different background conditions: monolingual (trained on data from either one of the two languages), and mixed (trained on a mix of both languages). The number of utterances per speaker in the training data sets was specially chosen to form three similarly sized sets with equal number of male and female speakers. A summary of the resulting sets is shown in Table I. Note that since there are no bilingual speakers in these corpora, the mixed background set may only represent a bilingual environment where speakers never mix both languages.

TABLE I
Summary of datasets for training of UBM.

Background	N speakers (*n males*)	Total duration in minutes
English	168 (84)	86.0
Xitsonga	168 (84)	87.0
Mixed:	168 (84)	87.9
English	*84 (42)*	*43.7*
Xitsonga	*84 (42)*	*44.2*

2) Test set: To test each model's ability to separate the languages, we built a test set composed of 20 speakers (5 male, 5 female per language) that were not used during training, with equal number of utterances per speaker, giving a total of 100 new English utterances (mean length of 3.04 s) and 100 new Xitsonga utterances (mean length of 3.61 s).

D. Linear Discriminant Analysis

As a way of providing side information to the system, we performed a Linear Discriminant Analysis (LDA) on the test i-vectors based on speaker labels. That is, the LDA algorithm aims at finding the linear combination of i-vector features that maximizes the distance between speakers, independently of their language. Since the test set is composed of 20 speakers, the i-vectors can be projected into 19 LDA dimensions. To compute LDA we used the *lda* function provided with the *MASS* package [25] in the programming language R [26].

E. Evaluation methods

The language filter hypothesis states that infants have a special mechanism that helps them separate their languages by grouping together similar sources of information, while keeping others apart. While the exact algorithms that infants use to solve this problem are not yet known, we propose to assess the feasibility of the task using two methods that aim to evaluate how well utterances from the two target languages can be discriminated, and to what extent speech from different speakers of a same language can be clustered together.

1) ABX discrimination score: In our framework, each utterance is represented as a fixed-length i-vector, or equivalently as points in a multidimensional space. The separation of languages A and B can therefore be assessed as the overlap between the distributions of points in A with those of B. The overlap between the distribution of points can be typically measured using KL divergence or Mahalanobis distance, but this necessitates to estimate their densities using some known distribution. Alternatively, it is possible to assess category separability in a nonparametric way simply using the ranks of distances between points. The ABX method consists in computing the extent to which within category distances are smaller than between category distances. This method yields statistically stable results and does not depend on particular assumptions about the shape of the category densities [27]. The computational implementation of this method (which is directly inspired by the psychophysical task of the same name [28]) consists in taking all possible combinations of a, b and x, where a is a sample from one category A (e.g., an utterance from English), b is a sample from the second category B (e.g., an utterance from Xitsonga), and x is a third sample (i.e., utterance) either belonging to the category A or B. For each $\{a, b, x\}$ triplet, x is classified as belonging to A if its euclidean distance to a, $d(x, a)$, is smaller than $d(x, b)$, and as belonging to B if $d(x, b) < d(x, a)$. Finally, the percentage of correct classifications can be computed based on known language labels, providing a discrimination score.

This score reflects the extent of the separation between the two categories: if the two distributions are highly overlapping, the distance between a pair of elements across categories will often be smaller than the distance between two elements within one category, thus leading to classification errors. In this case, the ABX score will be close to chance level (50%). The larger the separation between the two categories is, the smaller the number of errors produced and therefore the higher the ABX score obtained. Perfect separation would correspond to a score of 100%, and insures that k-means and unsupervised clustering algorithms would find the two categories with no error [29].

2) Hierarchical clustering: Another way to more directly assess the separability of the languages is to run a clustering algorithm. Given that the result may be highly dependent on the particulars of the algorithm, we used a set of 8 unsupervised agglomerative hierarchical clustering algorithms using the *hclust* function provided with the default *stats* package in R [26]. These algorithms consist in a series of steps in which N data points or elements are clustered together based on a distance metric and a linkage rule [30], [31], without providing any class labels. The algorithm starts by considering all data points as individual 1-element clusters, and in each step the two nearest clusters are merged together into a new one, until all data have been merged into a single cluster. The result of this process is a hierarchical tree with $N - 1$ nodes or levels. Traditionally, the results of a clustering algorithm may be evaluated using different measures of accuracy based on known classes. One such measure is the *purity*, which represents how homogeneous the discovered clusters are [30]. For any given set of K clusters (i.e., at a specific level of the hierarchical tree), the purity of the set can be defined by

$$purity\,(\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j| \qquad (2)$$

where $\Omega = \{\omega_1, ..., \omega_K\}$ is the set of clusters at the specified level, $C = \{c_1, ..., c_J\}$ is a set of classes (for example, different languages) and N is the total number of data points. A purity of 1 would indicate that every cluster at the specified level is pure (i.e., each contains elements of one and only one class or category), while lower values of purity represent higher proportions of mixed data within each cluster. In our test set, since we have equal amounts of utterances of each language, the minimum value of purity is fixed at 0.5 for $K = 1$ independently of any other parameter.

This measure provides a way of evaluating the feasibility of the task of language separation in a bilingual environment. If the two test languages are well separated in the acoustic space (as represented by their i-vectors), then high purity would be expected at the level with $K = 2$ clusters. However, if the distance between the two languages is not large enough, a larger amount of clusters may be needed to guarantee homogeneity, consequently segregating the speakers. It is important to note that the results of the clustering algorithms are highly dependent on the chosen linkage rule, especially when the variance is not equal across speakers or across languages. For this reason, quantitative results from a specific algorithm should not be directly interpreted. Instead, qualitative comparisons of the different background conditions held across all clustering methods may give better insight on the problem. Here, we evaluate the purity of hierarchical clustering as a function of number of clusters using eight different linkage rules provided with the *hclust* package: *single, complete, average, median, centroid, mcquitty, ward.D* & *ward.D2* [31].

III. Results

In order to visualize the spatial distribution of the test utterances in both languages, we first performed a Multidimensional Scaling (MDS) of the English and Xitsonga test i-vectors extracted from our three background models (monolingual English, monolingual Xitsonga, and mixed). The first two coordinates of the MDS are shown in Fig. 1.

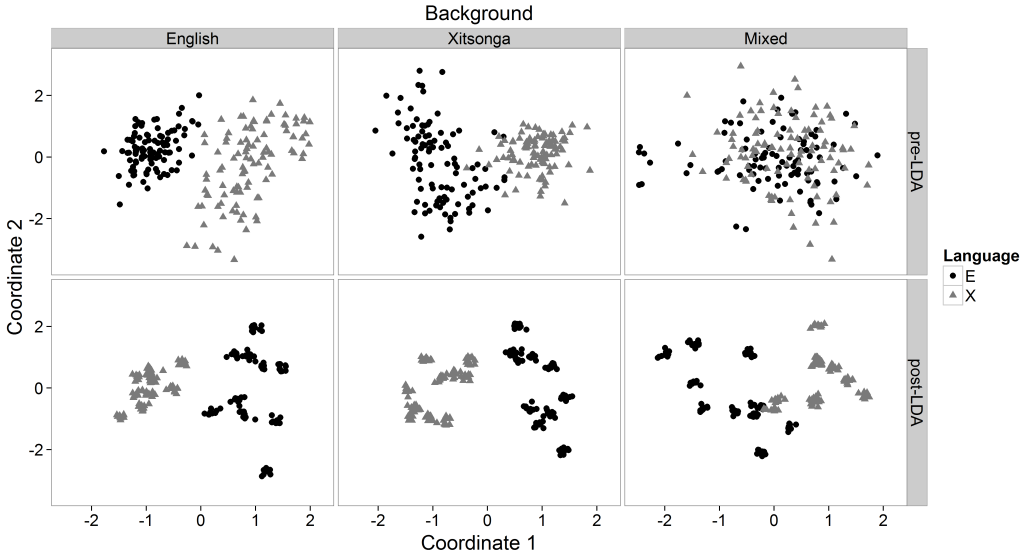


Fig. 1. Multidimensional Scaling plot of English (black dots) and Xitsonga (grey triangles) test utterances for three different language backgrounds (English, Xitsonga or mixed), pre- and post-LDA (top and bottom rows, respectively).

The top row represents the test i-vectors without any additional side information (pre-LDA), while the bottom row represents the projection of these i-vectors over the LDA dimensions that most discriminate speakers (post-LDA). For visualization purposes, we rescaled these two coordinates by their standard deviations without centering (all following computations were performed without rescaling).

The MDS projections of the test utterances prior to LDA show a clear difference between the monolingual and the mixed background conditions. While quite different in their distributions, test utterances given any of the monolingual backgrounds show relatively small amounts of overlap between English and Xitsonga. However, in the visualization of the utterances given a mixed background, it appears that both languages are largely overlapping. On the other hand, after performing LDA to discriminate speakers, the overlap observed in the mixed background condition is drastically reduced, while the two languages are pushed further apart in the monolingual backgrounds. Thus the addition of speaker information seems to enhance language separation in all three background conditions. In particular, the first dimension of the MDS (i.e., Coordinate 1) in the post-LDA distributions appears to separate the languages almost perfectly. This suggests that language identity is the most prominent source of between-speaker variance.

To have a more quantitative measure of the degree of overlap of the two languages we computed the ABX score. The results, shown in Table II, clearly demonstrate that raw i-vectors are moderately successful in enabling language discrimination given monolingual backgrounds (score around 70%), but do not enable separation at all in a mixed background (the ABX score is close to the chance level, 50%). However, post-LDA, the scores increase dramatically, and are over 80% correct for all backgrounds. In other words, the speaker-based LDA rescues the mixed background models from being utterly confused.

TABLE II
SUMMARY OF ABX RESULTS (IN % CORRECT).

Background	ABX score	
	pre-LDA	post-LDA
English	72.8	90.7
Xitsonga	69.8	92.1
Mixed	54.3	82.1

Finally, we applied several clustering methods to the test i-vectors. For each of the three background conditions, we computed clustering purity as a function of the number of clusters in each level of the hierarchical tree using eight different linkage methods. Fig. 2 shows the results before and after LDA for a *complete* linkage clustering[1] from $K = 1$ to $K = 20$.

The results shown in this example replicate what was observed in the ABX discriminability score: while the mixed background condition seems to have some disadvantage compared to both monolingual backgrounds, this is greatly overcome after enhancing speaker discrimination (post-LDA). However, a comparison across all linkage methods is required in order to evaluate the generalization of these findings. To summarise the performance of the different clustering methods we computed two derived measures: 1) in Fig. 3, the minimum number of clusters required to guarantee homogeneity in the whole set, i.e. to achieve $purity = 1$, and 2) in Fig. 4, the average purity over the 20 highest levels, i.e. from $K = 1$ to $K = 20$.

[1]In a complete linkage clustering, the two clusters with the smallest distance between their farthest pair of points are joined together [30].

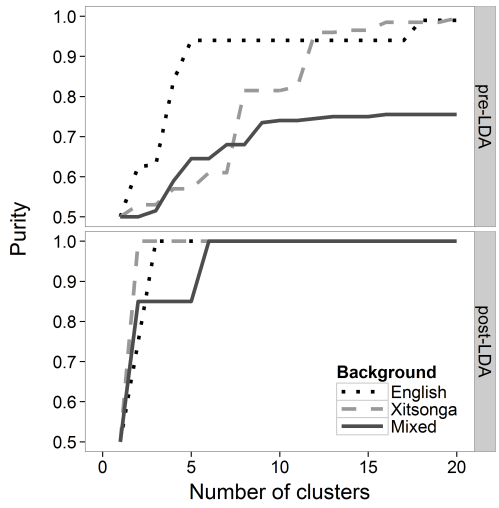


Fig. 2. Purity as a function of the number of clusters, pre- and post-LDA (top and bottom, respectively) for three different backgrounds. Clustering method: complete linkage.

Both derived measures support our main findings. Firstly, as shown in Fig. 3, in spite of the large variability across clustering methods, the minimum number of clusters required to guarantee intra-cluster homogeneity is consistently smaller for monolingual backgrounds than for the mixed background. Secondly, all background conditions greatly benefit from including talker side information (post-LDA), reducing the number of clusters to less than 20 —which would correspond to one cluster per speaker— irrespective of the clustering method.

Furthermore, in all three post-LDA conditions at least half of the clustering methods reached perfect purity within less than 5 clusters (7/8 methods for English, 6/8 for Xitsonga and 4/8 for Mixed background), and in all cases perfect purity with only 2 clusters (i.e., the *true* number of languages) was achieved by at least two methods. Analogously, the average purity of these clustering methods over pre-LDA utterances (Fig. 4) presents the same advantage of monolingual over mixed background, with higher purity values achieved in the monolingual cases. Again, the addition of side information (post-LDA) resulted in an enhancement of language separation in all background conditions.

IV. Discussion

We discuss separately the implications of this study for the field of early language acquisition and for the field of computational modeling of autonomous learning.

A. Relevance for language acquisition

Before discussing the results of our models, we first address the pertinence of the i-vector approach as a model of infants' representation of languages. The entire pipeline consists of three steps: (1) constructing the UBM and Total Variability space, (2) extracting i-vectors, and (3) performing speaker-based LDA. These steps assume a good perception of acoustic information [32], and the ability of performing statistical learning [8], [7], both of which have been documented in infants. Moreover, steps (1) and (2) are completely unsupervised, requiring no external information about phonemes or words, nor any information about speaker identity, or number and properties of different languages. The only linguistic hypotheses of these processing steps are that utterances are relevant units for performing language discrimination, that they can be modelled through Gaussian mixtures, and that they can be segmented out of continuous speech. Step (3) rests on the assumption that infants can recognize speakers on the basis of other non-linguistic information (e.g., visual, olfactory, etc.). Although there is data relevant to the recognition of the infant's mother [16], [33], [34], and that infants can match the gender of voices and faces [15], [35], more data would be needed to assess the true classification capacity of young infants.

In brief, the components of the pipeline we used could plausibly be used by infants. In our view, the Universal

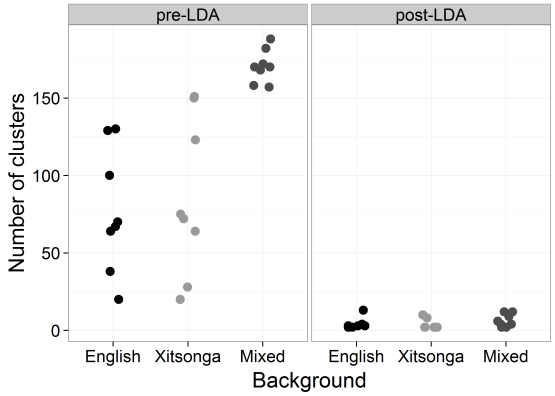


Fig. 3. Minimum number of clusters required to reach purity=1 (i.e., all clusters in the set are homogeneous), pre- and post-LDA (left and right, respectively) for three different background conditions. Each datapoint represents the results obtained with a different linkage rule.

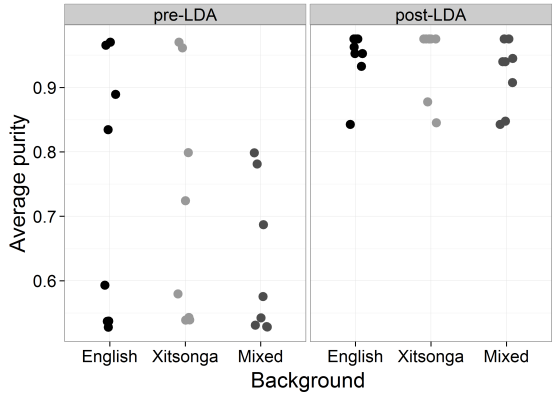


Fig. 4. Average purity over the last 20 clustering levels (i.e., K=1 to K=20), pre- and post-LDA (left and right, respectively) for three different background conditions. Each datapoint represents the results obtained with a different linkage rule.

Background Model could represent very early language experience gathered by the infant, prior to actually starting to do language separation for the purpose of language learning. This early background experience is needed for i-vectors to be definable, but needs not be extremely large (here, we used one hour and a half of speech). In fact, this background experience may start taking place even prenatally [10], [11]. However, we would like to raise three caveats.

First, even though each of the components of the processing pipeline we outlined could possibly be performed by infants, it seems implausible that they arise in a strict sequence of non-overlapping steps. Taken literally, our pipeline would imply that infants are unable to distinguish languages before they have accumulated enough evidence to construct an UBM, and that once the UMB is established, no further contact with new languages matters. Similarly, it would imply that before they have heard enough speakers to construct an LDA, infants cannot discriminate languages, and that once the LDA is set, no new speaker or language can matter. The empirical data, rather, shows that infants are able to discriminate languages from the first day of life, but that this capacity changes with experience. This would seem to suggest that the three steps are done concurrently and incrementally, with infants continuously updating their models of phonetic and speaker variability.

Second, it should be mentioned that the datasets used in this study are not an accurate representation of a child's experience, since they contain too many speakers, each of whom speak only a dozen sentences. In contrast, children are usually faced with a lot more speech from fewer speakers. This choice was driven by practical reasons, the recordings of infants' linguistic environments typically being very heterogeneous in recording quality. Moreover, since the selected datasets contain only monolingual speakers, our models may only represent a dual-language environment where speakers never mix both languages. While this scenario is not uncommon in families adopting a one-parent-one-language method [36], it certainly fails to represent many other cases of bilingual exposure that could pose an additional difficulty as suggested by recent research [37]. Whether speaker side information would help or hinder language separation in the presence of bilingual speakers remains unclear and will require further experiments with an appropriate dataset.

Third, it should be noted that the acoustic features used in our model (MFCCs) capture static spectral properties of speech, but lack potentially useful dynamic information that could further enhance language separation.

With all these caveats in mind, the current study unveiled that mixed language backgrounds may initially degrade the ability of infants to discriminate languages. Yet, as proposed by the "language filter" hypothesis, by focusing attention on the relevant dimensions (such as speaker differences) the separation of languages may be enhanced and thus greatly recover from a potential confusion. Altogether, this yields an interesting prediction for early language discrimination in infants, provided we can estimate the amount of language mixing that took place during the putative language background construction phase. Predictions about the longer term development (i.e. phonological, lexical, syntactic) of language in mixed linguistic backgrounds are more difficult to make, as this would require specifying a model of the language learner at these levels.

B. Relevance for computational models

Imagine that one were to construct a robot that can learn languages in any linguistic environment, including a multilingual one. How would one do it? The quantitative results of the present study reveal, first and foremost, that cues for language identity are coextensional with cues for speaker identity (i.e., they exist in the same subspace). This is both good news (we can use speaker identification to improve the representation) and bad news (this may not work as well if the same speaker speaks different languages). The second important aspect of the results is that although different language types look reasonably well separated when the language labels are known, it may be impossible to guess correctly the number of languages and get a 100% pure separation. What our results suggest, is that in order to guarantee a 100% pure input to a subsequent language learning system (which is important to avoid language chimeras), the language 'filter' should rather err on the side of over-segmentation, i.e., sort the input in slightly too many language types. A bit of redundancy may be a reasonable cost to pay in order to construct a learning system that is robust to multilingual environments.

V. Conclusions and Perspectives

In this project, we ran a feasibility study of language separation using i-vectors and corpora of two languages, and found that without speaker side information, language separation is relatively more difficult in mixed backgrounds than in monolingual backgrounds. However, with speaker side information this difficulty is almost completely eliminated. While language separation is overall surprisingly good given the limited amount of data that we used, it is yet not perfect, and would require more classes than the true number of languages to achieve perfect cluster purity.

The present study has used only a limited dataset. In order to fully characterise the feasibility of language separation, this methodology should be extended to study how the results change with larger datasets, and with other languages. Testing a pair of languages that are phonologically distant as English and Xitsonga enabled us to establish the evaluation methodology and to obtain a first baseline result. In future research it would be interesting to study the effect of the phonological distance between the languages, as well as the frequency distribution of each speaker and language, which is typically not uniform, but may be distributed according to a power law. In addition, the case of bilingual speakers (with perfect or imperfect accents) needs to be explored, as this situation may arise in multilingual communities. Finally, other kinds of side information (such as social and activity context) may help with language separation in ecological situations and should therefore be investigated.

VI. Acknowledgments

This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the École des Neurosciences de Paris Ile-de-France, and the Region Ile de France (DIM Cerveau et Pensée).

References

[1] S. Pinker, *The language instinct*. Harper, 1994.

[2] F. Grosjean, *Studying bilinguals*. Oxford University Press, USA, 2008.

[3] J. F. Werker and K. Byers-Heinlein, "Bilingualism in infancy: First steps in perception and comprehension," *Trends in cognitive sciences*, vol. 12, no. 4, pp. 144–151, 2008.

[4] L. A. Petitto, M. Katerelos, B. G. Levy, K. Gauna, K. Tétreault, and V. Ferraro, "Bilingual signed and spoken language acquisition from birth: Implications for the mechanisms underlying early bilingual language acquisition," *Journal of child language*, vol. 28, no. 02, pp. 453–496, 2001.

[5] F. Genesee, "Bilingual first language acquisition in perspective," *Childhood bilingualism: Research on infancy through school age*, pp. 45–67, 2006.

[6] V. Yip and S. Matthews, *The bilingual child: early development and language contact.* Cambridge: Cambridge University Press,, 2007.

[7] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, no. 3, pp. B101–B111, 2002.

[8] J. R. Saffran, R. N. Aslin, and E. L. Neport, "Statistical learning by 8-month-old infants," *Science*, vol. 274, no. 5294, pp. 1926–1928, 1996.

[9] G. F. Marcus, S. Vijayan, S. B. Rao, and P. M. Vishton, "Rule learning by seven-month-old infants," *Science*, vol. 283, no. 5398, pp. 77–80, 1999.

[10] T. Nazzi, J. Bertoncini, and J. Mehler, "Language discrimination by newborns: Toward an understanding of the role of rhythm," *Journal of Experimental Psychology*, vol. 24, no. 3, pp. 756–766, 1998.

[11] K. Byers-Heinlein, T. C. Burns, and J. F. Werker, "The roots of bilingualism in newborns," *Psychological Science*, vol. 21, no. 3, pp. 343–348, 2010.

[12] L. Bosch and N. Sebastian-Galles, "Evidence of early language discrimination abilities in infants from bilingual environments," *Infancy*, vol. 2, no. 1, pp. 29–49, 2001.

[13] J. F. Werker and S. Curtin, "PRIMIR: A developmental framework of infant speech processing," *Language learning and development*, vol. 1, no. 2, pp. 197–234, 2005.

[14] S. Curtin, K. Byers-Heinlein, and J. F. Werker, "Bilingual beginnings as a lens for theory development: PRIMIR in focus," *Journal of Phonetics*, vol. 39, no. 4, pp. 492–504, 2011.

[15] D. Bristow, G. Dehaene-Lambertz, J. Mattout, C. Soares, T. Gliga, S. Baillet, and J.-F. Mangin, "Hearing faces: How the infant brain matches the face it sees with the speech it hears," *Journal of Cognitive Neuroscience*, vol. 21, no. 5, pp. 905–921, 2009.

[16] T. M. Field, D. Cohen, R. Garcia, and R. Greenberg, "Mother-stranger face discrimination by the newborn," *Infant Behavior and development*, vol. 7, no. 1, pp. 19–25, 1984.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[18] D. G. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Proceedings of Interspeech 2011*, vol. 2011, no. 8. International Speech Communication Association, 2011, pp. 861–864.

[19] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via ivectors and dimensionality reduction," in *Proceedings of Interspeech 2011*, vol. 2011. International Speech Communication Association, 2011, pp. 857–860.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[21] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1. 0: A MATLAB toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, 2013.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, *et al.*, *The HTK book (for HTK version 3.4)*, 2006.

[23] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium, Philadelphia*, 1993.

[24] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," *Proc. SLTU*, pp. 194–200, 2014.

[25] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[26] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[27] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 1–5.

[28] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*. Psychology press, 2004.

[29] T. Schatz, "ABX-discriminability measures and applications," Ph.D. dissertation, Université Pierre et Marie Curie, 2016.

[30] C. D. Manning, P. Raghavan, H. Schütze, *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[31] F. Murtagh and P. Contreras, "Methods of hierarchical clustering," *arXiv preprint arXiv:1105.0121*, 2011.

[32] P. D. Eimas, E. R. Siqueland, P. Jusczyk, and J. Vigorito, "Speech perception in infants," *Science*, vol. 171, no. 3968, pp. 303–306, 1971.

[33] I. Bushnell, "Mother's face recognition in newborn infants: Learning and memory," *Infant and Child Development*, vol. 10, no. 1-2, pp. 67–74, 2001.

[34] J. M. Cernoch and R. H. Porter, "Recognition of maternal axillary odors by infants," *Child development*, pp. 1593–1598, 1985.

[35] A. S. Walker-Andrews, L. E. Bahrick, S. S. Raglioni, and I. Diaz, "Infants' bimodal perception of gender," *Ecological Psychology*, vol. 3, no. 2, pp. 55–75, 1991.

[36] S. Barron-Hauwaert, *Language strategies for bilingual families: The one-parent-one-language approach*. Multilingual Matters, 2004, no. 7.

[37] K. Byers-Heinlein, "Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size," *Bilingualism: Language and Cognition*, vol. 16, no. 01, pp. 32–48, 2013.

2.3.6.1 Generalization to new speakers

In Experiment 6, we showed that the use of side information (in this case, speaker identity) could help a bilingual background model, which had otherwise no way of knowing of the existence of two languages, to disentangle the input. However, a limitation of this study is that we scored the ABX discrimination of the models on the same speech from which the LDA dimensions were learnt. To make sure that these results were not tied to these specific speakers, we need to prove that the model can generalize its discrimination ability to new speakers. We thus propose to select new English and Xitsonga speakers that the models have not seen before, neither during training of the background model nor during LDA, and to project their vectors onto the LDA dimensions learnt in Experiment 6. If the dimensions learnt from LDA are indeed relevant for language separation regardless of the speakers, then the models should be able to generalize to new speakers (i.e., showing good discriminability scores). We will compute the ABX score for the new utterances using the LDA-projected vectors.

Materials

As background models, we will use the same 3 UBMs (one English monolingual, one Xitsonga monolingual, and one mixed English-Xitsonga model) used in Experiment 6. For the test, we select 2 new speakers of each language (one male, one female) and extract the i-vectors of 5 utterances from each speaker, making a total of 20 test vectors.

Results

In Figure 2.12, we show a visualization of the old post-LDA utterances (from Experiment 6) and the new utterances, projected onto the same LDA dimensions learnt in Experiment 6. Based on this projection, it seems that the new utterances are overall well separated in the first dimension, each siding with the old utterances of the same language. However, the new utterances seem more disperse, suggesting that part (but not all) of what the LDA algorithm had learnt before was specific to that set of speakers. To quantify the separation of the two languages based on the LDA projection, we compute the ABX scores using the new test utterances (projected onto the 19 LDA dimensions) for each of the background models. We additionally compute an ABX score restricting the comparison to the first LDA dimension only. The results are show in Table 2.6.

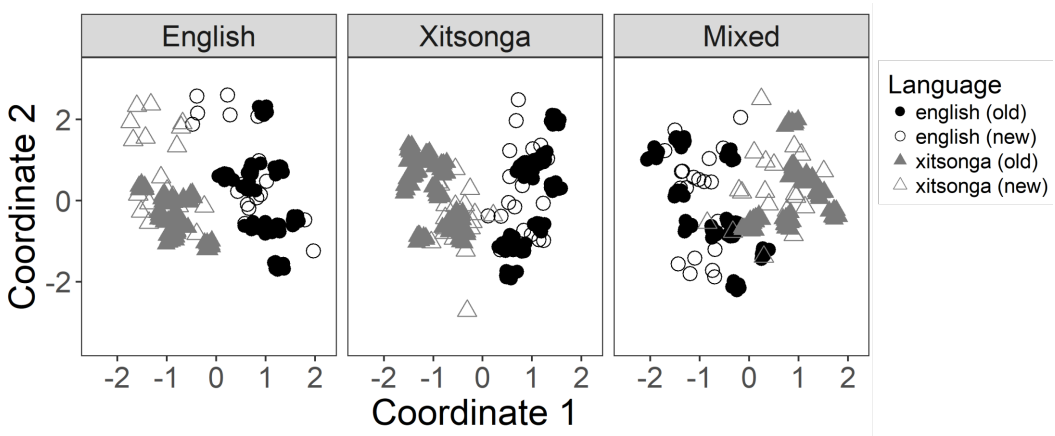


Figure 2.12: Multidimensional scaling of the post-LDA utterances from Experiment 6 (filled dots and triangles), and 20 new utterances (2 speakers from each language) projected onto the same LDA dimensions (empty dots and triangles).

Table 2.6: Mean ABX scores for new LDA-projected English and Xitsonga utterances.

Background	ABX (19 LDA dim.)	ABX (1st LDA dim.)
English	74%	93%
Xitsonga	79%	94%
Mixed	78%	86%

The ABX results show, first of all, a relatively good separation of the languages using the full 19 LDA dimensions learnt in Experiment 6. While lower than the scores reported in Carbajal, Dawud, Thiollière and Dupoux (2016), these results indicate that the model can generalize its discriminability to newly encountered speakers. Furthermore, when using only the first LDA dimension, the ABX scores show excellent discrimination performance. This means that, if infants could focus their attention on this specific dimension, they would be able to separate future input with little difficulty.

2.4 General discussion and future directions

In this chapter we have introduced a novel model of infant language discrimination, borrowing a state-of-the-art system from speech technologies: the i-vector model. Our goal was to revive the discussion on how young infants come to discover the presence of two languages in their input, without resorting to over-simplifications of the problem. In order to provide a computational account that could explain infant behavior in many different scenarios (e.g., hearing natural or filtered speech, by bilingual or multilingual speakers, etc.) it was necessary to work with a model that could incorporate different

levels of acoustic and prosodic information into a comparable ground. In a series of six experiments we explored the behavior of an unsupervised i-vector model when faced with multilingual speech under various conditions. While in LID technologies this model is typically trained with large amounts of multilingual data, here we have shown that a mere half an hour of speech from a single monolingual speaker suffices to reproduce young infants' ability to discriminate languages, succeeding and failing in similar scenarios.

In Experiments 1 and 2 we showed that, in the presence of a single bilingual speaker, our model (trained with a small French background) failed to discriminate languages with overlapping acoustic properties, such as Spanish-Catalan and English-Dutch, while it succeeded in languages with a larger acoustic distance. Furthermore, as has also been reported in infants, we found that the ability to discriminate distant languages did not depend on the familiarity of the background model with one of the test languages. In Experiment 3, we observed similar results (albeit with lower performance) when testing the model on filtered speech, indicating that the model was relying, at least to some degree, in prosodic information or other acoustic properties that are not degraded by low-pass filtering. These results seem to be in agreement with behavioral experiments in newborns, indicating that the i-vector model was able to capture similar properties of speech.

An interesting observation from Experiment 2 is that, when confronted with several bilingual speakers, within-language speaker distance was found to be larger than within-speaker language distance. If this is the case, then infants who encounter bilingual speakers may struggle to notice the presence of two languages, as this distinction could be hidden by speaker variability. However, as the speakers that we used in Experiment 2 were not perfect bilinguals (that is, they all had a moderate accent in their L2 English), within-speaker language distance may have been smaller than what would be expected from unaccented bilingual speakers. In a recent experiment with 5-month-old English-learning infants, Paquette-Smith and Johnson (2015) showed that Spanish-accented English was not discriminated from Spanish, while these same languages were successfully discriminated when produced by a perfect Spanish-English bilingual speaker. As accented speech is not rare in bilingual infants' input, future work should attempt to characterize how much variability is introduced by the accent of the bilingual speaker, and how this could impact language discriminability. For this, however, a special dataset would be needed, with speech from several bilingual speakers with varying degrees of non-native accent in each language.

In Experiment 4, we explored language discrimination in the presence of multiple monolingual speakers,

using a simulation inspired in the habituation model proposed by Ramus et al. (1999). Unlike in previous psycholinguistic models described at the beginning of this chapter, the i-vector model allowed us to make separate predictions for natural and filtered speech. We were thus able to show, for the first time, a computational replication of the surprising results found by Ramus (2002b), where two distant inter-class languages, Dutch and Japanese, were found to be confused when natural speech from several speakers was used, while they were discriminated when using filtered speech. These results highlight the importance of constructing models that can deal with multiple sources of variability. Overall, our simulation results coincided with Ramus et al.'s (1999) in those language pairs that had been tested in newborns. Other language pairs resulted in disagreements that cannot be resolved until further empirical evidence is collected.

Next, in Experiments 5 and 6 we explored the role that language experience plays in the extent to which languages can be discriminated. In Experiment 5 we confirmed that, given only 25 minutes of training speech (which could potentially be accumulated by a 5-day-old infant), language discrimination does not seem to depend on individual characteristics of the background. That is, the discrimination of distant language pairs, and the confusion of close language pairs, is overall the same regardless of the speaker and the language of the background model. These robust results are in agreement with the notion that language discrimination abilities, as many other speech processing capacities, are universal at birth (Werker, 2012).

Finally, in Experiment 6, we explored the impact of having two languages in the background model. Our results showed that, if infants were to blindly accumulate speech from two languages during the initial formation of the background, language discrimination would not come as easily as it does for those with a single language input. However, we found that simple attention to side information, such as speaker identity, would allow the child to overcome this initial confusion. This would mean that, for language pairs such as Spanish-Catalan or English-Dutch, which are not discriminated at birth, additional evidence supporting the presence of two languages would need to be discovered. The ability to identify and attend to useful dimensions of variability has been proposed as one of the key mechanisms that allows bilingual infants to learn their two languages (Curtin et al., 2011). Indeed, by the time they are 4 months old, infants learning Spanish and Catalan succeed in discriminating their languages (Bosch & Sebastián-Gallés, 2001). Which specific cues allow them to succeed, and how they discover them, are still open questions.

As, in Experiment 6, all speakers were monolinguals, the benefit of attending to speaker identity may

seem trivial. Nonetheless, we have shown that the dimensions resulting from learning to separate certain speakers in the environment can be generalized to new (monolingual) speakers, facilitating language discrimination. This would suggest that language separation by speaker, at least in the initial stages of development, may be a useful cue (although probably not the only one) to language discrimination. Once the two languages have been discovered, language separation by speaker may be less important. Whether these learnt dimensions would be sufficient to separate speech from bilingual speakers, however, remains to be tested.

A common factor throughout these six experiments has been the non-negligible variability that can be found within languages and speakers, which begins to paint a different picture of the difficulty of the challenge that bilingual infants must face during the first few months of life. It should be kept in mind that we have conducted these experiments with relatively controlled datasets; the variability present in real-world child-directed speech may far exceed what has been described in computational and even experimental studies. Far from the simplified statement that languages from different rhythmic classes are easy to discriminate, our computational observations suggest that the task is not at all trivial, and understanding how infants achieve this feat is still far from solved. Discovering the presence of two languages may not only depend on how different the languages sound, but also on how (and from whom) these languages are introduced to the child, and how much variability from many other sources beyond language rhythm are present in their input. Much research (both experimental and computational) thus remains to be done before we can offer a concrete account of how this fantastic achievement unfolds.

In summary, in this chapter we have explored the feasibility of this novel model as a representation of infant perception, finding in general good agreement with experimental results, and offering some new insights on the task of language discrimination. Its main interest, however, lies not in the fact that it replicates what is already known from infant experiments, but rather in its flexibility to explore where and why it succeeds or fails. Here, we have barely scratched the surface of what the i-vector models can offer. Firstly, it is possible to modify, redefine, or add different feature vectors to represent properties of the speech signal. For instance, the MFCCs could be replaced by the more organic Gammatone features, which are gaining popularity in cochlear implant applications due to their ability to model the frequency response of the cochlea. The feature vectors (or even directly the i-vectors) could be complemented with non-acoustic information, such as contextual information, or visual information. Models based on different sets of features could be compared to explore which

properties of the signal are most relevant for language separation. Furthermore, and perhaps most importantly, it is possible to examine the contents of the Total Variability space: which dimensions has it learnt? How do these impact language discrimination? Finally, this model can be trained with many different datasets, providing the possibility of training different "babies" with which to run experiments. Overall, language development research has much to gain from incorporating tools from artificial intelligence research and speech technologies.

2.A Appendix A: Pipeline details

2.A.1 Feature extraction

As in many GMM-based systems, the i-vector model requires transforming the raw speech signal into feature vectors capturing relevant linguistic properties. While many different sets of features can be used for this purpose, here we will use the popular MFCC-SDC features which, as discussed before, can capture both static and dynamic information. For every step in the feature extraction process (shown in Figure 2.13) we will discuss, when possible, the parallels between the algorithms and human auditory perception.

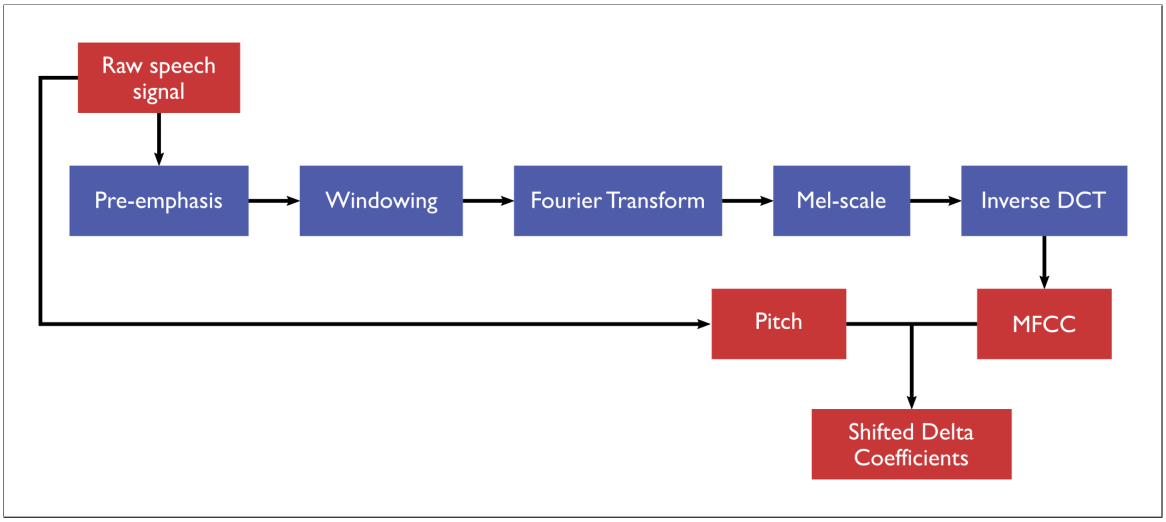


Figure 2.13: Pipeline for MFCC-SDC feature extraction.

2.A.1.1 Static spectral features (MFCCs)

The *Mel Frequency Cepstral Coefficients* (MFCC) are feature vectors that can capture frequency information of the speech signal over small time windows. Using this representation, the continuous waveform of any given utterance can be transformed into a sequence of static acoustic feature vectors.

Pre-emphasis

The first step in the pipeline is the application of a high-pass filter to the speech signal. This is done to compensate for the *spectral tilt* of the glottal source, which causes the energy of voiced segments

to decrease towards higher frequencies (Kreiman, Gerratt, & Antoñanzas-Barroso, 2007). By pre-emphasising the signal, we reduce the effect of the spectral tilt and thus obtain a better balance across the spectrum. It has been recently argued that this pre-emphasis step is compatible with experimental evidence that mid-range frequencies, i.e. between 500Hz to 4kHz, are perceived as equally loud as low-range frequencies with bigger amplitudes (H. Fletcher & Munson, 1933; Schatz, 2016; Suzuki & Takeshima, 2004). The pre-emphasized signal has the form:

$$s_{pre}[n] = s[n] - \alpha \cdot s[n-1] \tag{2.2}$$

where $n = \{0, \ldots, N\}$ are the time samples of the signal $s[n]$, and α is a constant. Typical values of α range between 0.9 and 1.0.

Windowing

As the spectral properties of speech vary throughout the utterance, in order to calculate these properties it is necessary to restrict the analysis to short segments during which the signal can be assumed approximately stationary. To achieve this, the waveform is sliced into a sequence of windows (often referred to as *frames*), with a typical width of about 20 ms to 40 ms. To avoid discontinuities at the borders of each frame resulting from this segmentation –which could cause artifacts in the spectrum– the signal is usually multiplied by a *window function* with smooth edges. A common function used in speech processing is the *Hamming* window:

$$w[n] = 0.54–0.46 \cos \frac{2\pi \cdot n}{N} \tag{2.3}$$

where $n = \{0, \ldots, N\}$ are the N time samples within a frame. The shape of this function is shown in Figure 2.14

Figure 2.15 shows an example of a speech frame (a short segment of the vowel [ɐ]), before and after multiplying the signal by the Hamming function: $s[n] \cdot w[n]$.

This windowing process is then repeated along the signal, shifting forward by a given time step (usually smaller than the window width), which results in a sequence of partially overlapping frames.

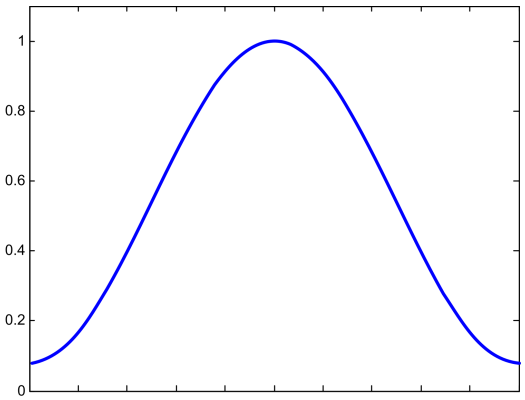


Figure 2.14: Hamming function.

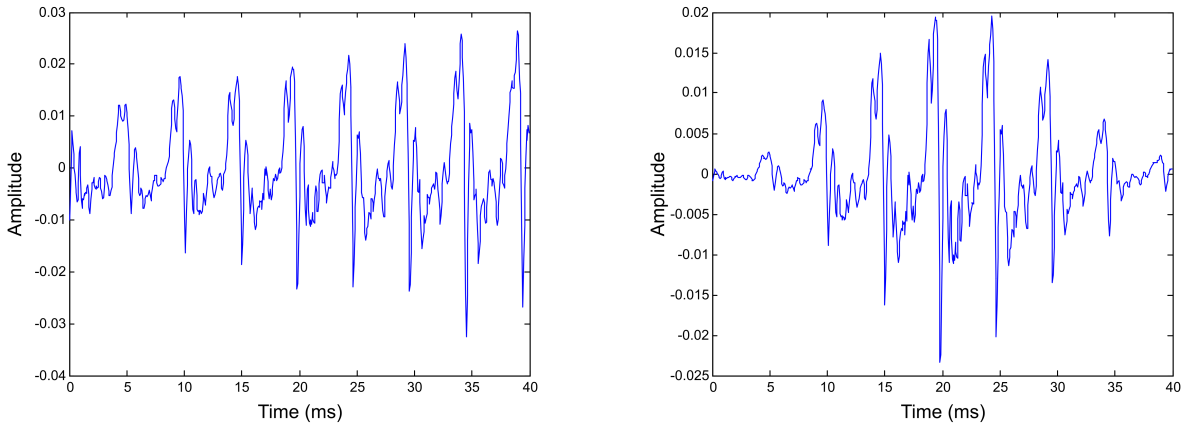


Figure 2.15: A 40 ms frame of the vowel [ɐ] segmented from a Catalan utterance, before (left) and after (right) applying a *Hamming window* function.

For instance, a 20 ms window with a 10 ms shift will result in 100 frames for 1 second of speech, with a 50% overlap between contiguous frames.

Fourier Transform

Next, a short-term Discrete Fourier Transform (DFT) is computed over each window to transform the speech signal from the time domain to the frequency domain,

$$X[k] = \sum_{n-0}^{N-1} (s[n] \cdot w[n]) e^{-i \cdot 2\pi \cdot k \cdot n/N} \tag{2.4}$$

where each k corresponds to a frequency $f(k) = k \cdot R/N$ with R being the sampling rate. In practice, the computation is done with a Fast Fourier Transform, and only the squared magnitude of the spectrum $|X[k]|^2$ is retained. Figure 2.16 shows the resulting magnitude of the frequency spectrum for the pre-emphasized and windowed speech segment.

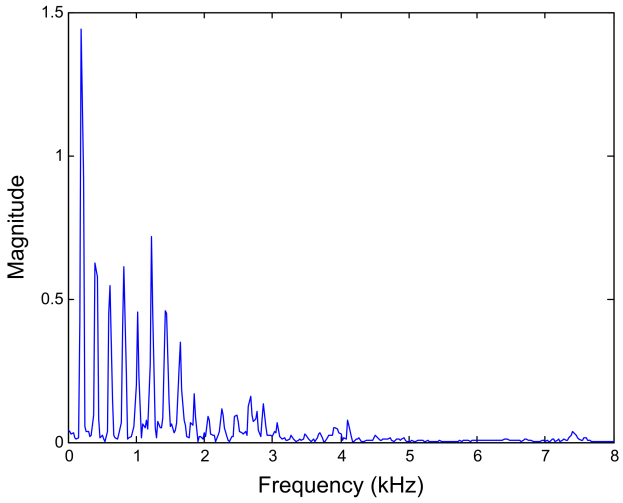


Figure 2.16: Magnitude of the frequency spectrum obtained with a DFT of the pre-emphasized and windowed segment of the vowel [ɐ] shown before.

Mel-frequency spectrum

Experiments on pitch perception have shown that humans are not equally sensitive to changes in frequency across the spectrum: low frequencies are better discriminated than higher frequencies (S. S. Stevens & Volkmann, 1940; S. S. Stevens, Volkmann, & Newman, 1937). To correct for this non-linear perception, the frequency spectrum obtained through the DFT must be passed through a series of non-uniformly distributed band-pass filters called *Mel filter banks*. The center frequencies of the Mel filter banks are separated linearly on the *Mel-scale*, i.e. a scale of perceived pitch distance based on Stevens et al.'s (1937) experiments. The relationship between frequencies in the Mel-scale, f_{mel}, and frequencies in the Hertz scale, f_{Hz}, is approximately linear within the first 1000 Hz, and follows a logarithmic relation above that frequency, given by the following formula:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{2.5}$$

The shape and number of filters can vary, but typically within 20 to 40 filters with a triangular shape are used. Figure 2.17 (left) shows a Mel filter bank with 20 triangular filters. The resulting filtered spectrum, shown in Figure 2.17 (right), is called the *Mel-frequency spectrum*. The new spectral coefficients MF_k are thus given by the application of the kth filter to the squared magnitude of the frequency spectrum.

Finally, the logarithm of the amplitude is computed for each Mel-frequency coefficient. This log-

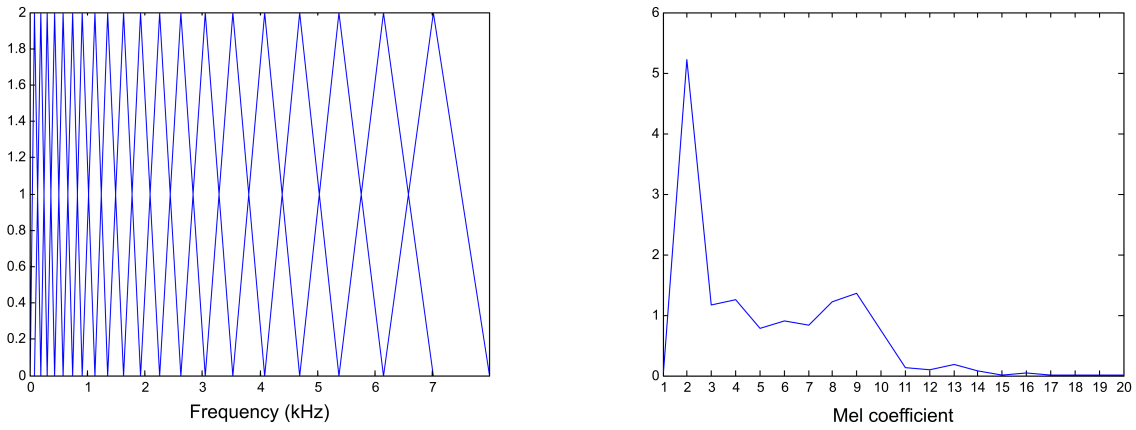


Figure 2.17: Triangular Mel filters (left), and magnitude of the Mel-frequency spectrum (right) after applying the filter bank to the DFT spectrum shown before.

compression is intended to compensate for the fact that sound loudness is not perceived linearly across the dynamic range.

Discrete Cosine Transform

The last step in the MFCC pipeline is the projection of the Mel-frequency spectrum onto a cosine basis, using an Inverse Discrete Cosine Transform. The result of this step will be a "spectrum of the spectrum", which is referred to as *Mel Frequency Cepstrum* (MFC), and its *cepstral coefficients* will be our final feature vectors, the MFCCs (i.e., Mel-Frequency Cepstral Coefficients). The coefficients are given by:

$$\mathrm{MFCC}_d = \sum_{k=1}^{K} \log_{10}(\mathrm{MF}_k) \cos\left(\frac{\pi \cdot k \cdot (d - 0.5)}{K}\right) \tag{2.6}$$

where $d = \{1, \ldots, D\}$ are the indexes of the D resulting cepstral coefficients (usually, $D = 12$, with energy often added as a 13th coefficient), and MF_k are the K Mel-frequency spectral coefficients obtained in the previous step. Figure 2.18 shows the magnitude of the first 12 MFCC for the speech segment shown before.

According to the source-filter model of speech production, the spectrum of the speech signal is the result of passing the glottal source spectrum through a filter generated by the specific configuration of the vocal tract (K. Stevens, 2000). In this context, the last step in the pipeline can be interpreted as a decorrelation of the contribution of the source and the filter, where only the spectral information

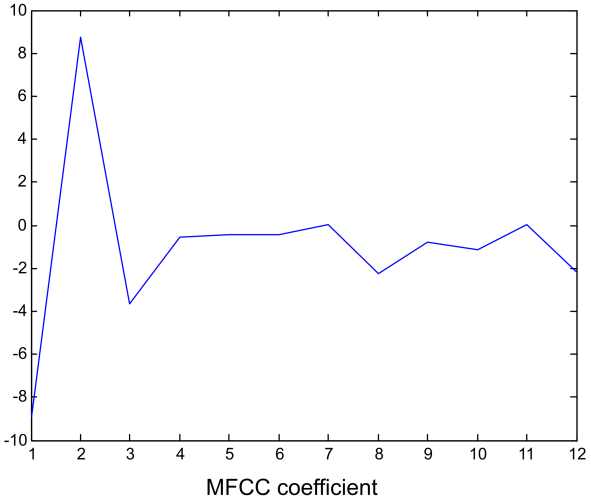


Figure 2.18: Magnitude of the first 12 Mel-Frequency Cepstral Coefficients.

coming from the filter (represented by the lowest cepstral coefficients) is retained, thus keeping mostly acoustic information pertaining to the phonemes and discarding information about the source[7]. To the best of our knowledge, there is no current evidence showing that the human auditory system performs a similar separation of the signal. However, we will keep this step in the pipeline as it is a standard procedure in feature extraction systems. Future work should investigate the effect of using different features in the performance of the system, in comparison with human perception.

The resulting MFCCs are D-dimensional feature vectors that capture short-term spectral information that is relevant for speech recognition. By repeating this process over every frame along the signal, any given utterance can be translated into an MFCC sequence or matrix, as shown in Figure 2.19.

2.A.1.2 Dynamic features (SDCs)

While MFCCs are widely used to capture spectral information of speech, a known limitation of this representation is that it only reflects instantaneous acoustic properties. Speech, however, is a rich signal whose properties evolve in time. In order to overcome this limitation, it is possible to incorporate dynamic features. Here, we will use *Shifted Delta Coefficients* (SDCs), which provide an approximation to the evolution of the features over a wide range of time frames (Li et al., 2013; Torres-Carrasquillo et al., 2002). The SDCs are obtained by calculating k local Δ coefficients (that is, first order derivatives of the feature vectors) at different time points around a center frame, each separated by a distance of P frames, as shown in Figure 2.20. These *shifted deltas* are then stacked together and appended to

[7]Source information, e.g. the pitch, can be calculated separately and added to the feature vector if it is necessary for a specific application

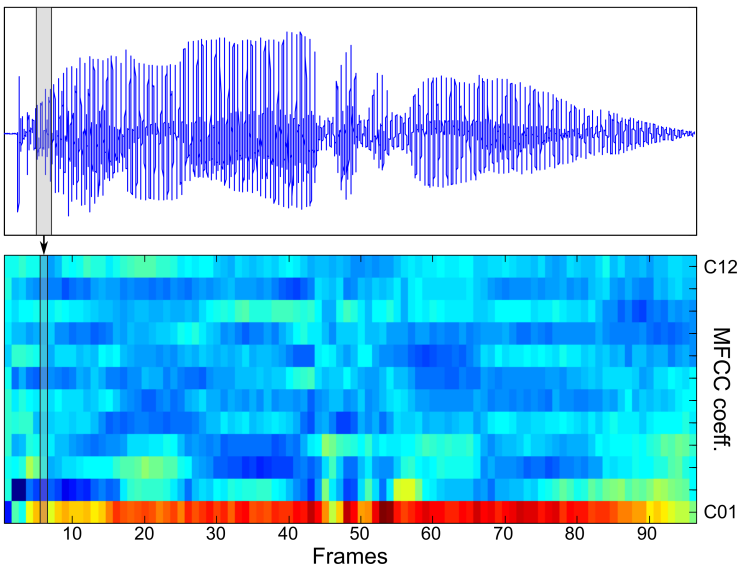


Figure 2.19: Speech signal (top) and magnitude of the first 12 Mel-Frequency Cepstral Coefficients at each time frame of the signal (bottom). Magnitude of the MFCCs is represented in a color scale.

the MFCCs to form an extended feature vector.

A common configuration of the SDCs is the *Z-d-P-k* = *7-1-3-7*, meaning that we retain the first $Z = 7$ MFCC coefficients (often including the *C0* coefficient, representing energy), and for each of them we calculate the Δs by computing the difference between frame $t + 1$ and frame $t - 1$, at $k = 7$ points around each time frame t, shifting by 3 (P) frames at a time. The result of this step is a feature vector of dimension $k \cdot Z = 7 \cdot 7 = 49$ (which is stacked with the 7 static MFCC features, for a total of 56 features) which represents the temporal evolution of the MFCCs over a span of $k \cdot P = 21$ consecutive frames. This means that with a typical 10 ms shift, we cover 210 ms of speech, which – depending on the language – corresponds approximately to the length of one syllable (Duanmu, 1994; Duez, 2006; J. Fletcher & McVeigh, 1993; Kuwabara, 1996; Yang, 1998).

Finally, to take into account additional prosodic information in our feature vector (besides the energy, already captured by MFCC coefficient *C0*), we can include a measure of pitch (*F0*). As with the static spectral coefficients, by computing SDCs over pitch along the signal, we can represent the evolution of this feature in time. The instantaneous pitch and its SDC features are stacked together with the MFCC-SDC features to form a 64-dimensional feature vector.

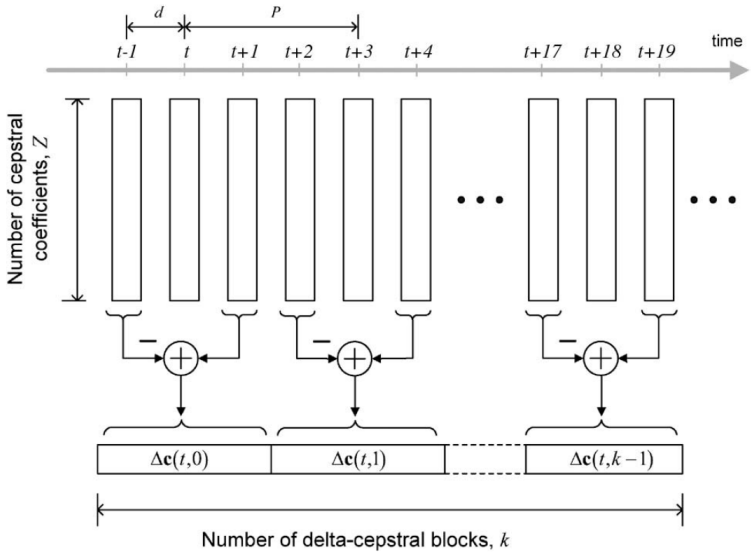


Figure 2.20: SDC feature extraction at frame time t. Reproduced from Li, Ma, and Lee, 2013.

2.A.2 Background Model

After feature extraction, the next step in the pipeline is the training of a background model. The goal of this model is to learn the regularities of the acoustic space of speech. This step can be taken to represent a child's prior exposure to speech, and will be crucial to assess the role of language experience in the system's ability to separate languages. As explained in Section 2.1.0.2, the background distribution of speech features is approximated with a Gaussian Mixture Model (GMM), that is, a linear combination of Gaussians $\mathcal{N}(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The superposition of the density functions of K Gaussian components, each weighted by a mixing coefficient ω_k, gives rise to a probability distribution of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.7}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and the covariance of each Gaussian component. A reduced example of a GMM is shown in Figure 2.21. Since the acoustic space is defined by the D-dimensional feature vectors, $\boldsymbol{\mu}_k$ will also be a vector of dimension D, and $\boldsymbol{\Sigma}_k$ will be a symmetric matrix of dimension $D \times D$. To reduce the cost of computation of the covariance matrix, diagonal covariances are often used instead of full-covariances. Given K components and D-dimensional features, we define a mean *supervector*, $\mathbf{m}$, as the stacked vectors of all the D-dimensional means. The resulting supervector will be of dimension $K \cdot D$.

As the real underlying categories (i.e., the Gaussian components) that generate the background dis-

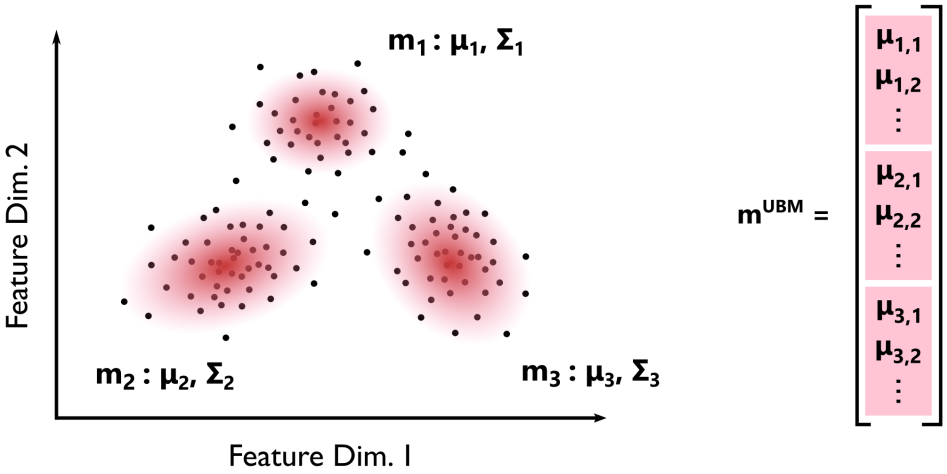


Figure 2.21: Example of a GMM with 3 Gaussian Mixtures, m_1, m_2, and m_3, defined by their means μ_k and covariances Σ_k. Each dot represents one feature vector from a single frame, plotted on the first two dimensions. Each Gaussian mean is a vector of dimension D (same size as the feature vector). The mean supervector, $\mathbf{m}^{\text{UBM}}$, is obtained by stacking the vectors of means of each Gaussian mixture, resulting in a large vector of dimension $K \cdot D$ (here, $K = 3$).

tribution are unknown, the parameters of each component must be estimated using unsupervised techniques. The method that is most commonly used to estimate means and covariances of the GMM is the Expectation-Maximization algorithm. For a description of this method see Section 2.B.

GMMs can be used to fit any complex distribution by adjusting the parameters and weights of a sufficient number of Gaussian components (Bishop, 2006), making them useful to model real-world data, and in particular the acoustic properties of speech. As we mentioned before, it is in the context of SID and LID that the concept of *Universal Background Model* (UBM) emerged. The UBM is a GMM trained on large amounts of data (usually in the order of hundreds of hours of speech) containing as much variability as possible, with the aim of capturing *universal* (i.e., speaker- and/or language-independent) properties of speech. For instance, in LID systems, UBMs are typically trained with speech from a large number of different languages, while in SID systems the model is trained with speech from as many different speakers as possible.

Here, we are interested in modelling the acoustic space that a very young infant could begin to form from a small amount of input. A baby's language background evidently cannot contain as many speakers or languages as typically used to train these models. However, nothing prevents the UBM from being trained with smaller datasets with reduced variability. For consistency with the terminology used in the i-vector framework, we will continue to call the background model "universal", although in practice it will capture speech properties of a rather small universe.

2.A.3 Total Variability space

Once the UBM has been trained, the next step in the pipeline is learning the Total Variability subspace. As mentioned before, we intend to model a given utterance $\boldsymbol{\mu}$ as a shift from the means of the background model, $\mathbf{m}^{\text{UBM}}$, restricted to a low-dimensional subspace:

$$\boldsymbol{\mu} = \mathbf{m}^{\text{UBM}} + \mathbf{T}\mathbf{w} \tag{2.8}$$

where $\boldsymbol{\mu}$ represents the supervector of Gaussian means for the given utterance. This process is in essence a *factor analysis*, which aims to describe the variability in the input in terms of a lower number of dimensions (Dehak, Kenny, et al., 2011). In order to learn the bases of the subspace (i.e. the matrix $\mathbf{T}$), we need to train the model with a large number of utterances, which are typically the same as used for training the UBM. The size of the TV subspace (and therefore of the i-vectors), F, is defined prior to training.

Given that the number of speech frames in a single utterance is scarce, it is not possible to train a GMM to compute the supervector $\boldsymbol{\mu}$. Thus, in order to train the TV space, we will instead rely on the posterior probabilities of the data using our pre-trained UBM as prior, and then use an Expectation-Maximization (EM) algorithm to estimate $\mathbf{T}$, using $\mathbf{w}$ as the latent variable (P. Kenny, Boulianne, & Dumouchel, 2005; P. Kenny, Ouellet, Dehak, Gupta, & Dumouchel, 2008). This process can be conceptually seen as an optimization of the subspace $\mathbf{T}$ to minimize the distance between the observed data, $\tilde{\mathbf{F}}$, and the approximated shift, $\mathbf{T}\mathbf{w}$:

$$\mathbf{T}, \mathbf{w} = \underset{\mathbf{T},\mathbf{w}}{\operatorname{argmin}} \|\tilde{\mathbf{F}} - \mathbf{T}\mathbf{w}\|^2 \tag{2.9}$$

The EM algorithm begins by initializing the matrix $\mathbf{T}$ randomly, and then iterating the following two steps until convergence:

- E-step: For each utterance u, given the UBM parameters and the current estimate of $\mathbf{T}$, calculate the posterior distribution of the latent variable $\mathbf{w}(u)$.
- M-step: Accumulate statistics over all the training utterances $u = \{1, \ldots, U\}$, and then re-estimate $\mathbf{T}$ by solving a system of linear equations that maximize the log-likelihood of the data.

Additional details on this step are given in P. Kenny et al. (2008). The result of this process is a basis defining the low-dimensional Total Variability subspace. The dimension of $\mathbf{T}$ is $KD \times F$ where F is the pre-defined number of factors (i.e., the size of the i-vectors). Thus, by restricting the shifts to the TV space, we reduced the dimensionality from KD (the size of the UBM supervector) to F. For a typical i-vector system with around 400 TV dimensions, this would mean a reduction of several orders of magnitude from the original size of the supervector. For instance, a UBM with 2048 Gaussians and 60-dimensional features would have supervectors of dimension 122880×1, much larger than the final 400 dimensions of the i-vectors.

2.A.4 I-vector extraction

Finally, given the matrix $\mathbf{T}$ obtained through the EM algorithm, for any new utterance u its posterior distribution $\mathbf{w}$(u) will be normal with covariance matrix $\boldsymbol{l}^{-1}(u)$ and mean:

$$\overline{\mathbf{w}}(u) = \boldsymbol{l}^{-1}(u).\mathbf{T}^t\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{F}}(u) \tag{2.10}$$

where $\boldsymbol{l}(u) = \mathbf{I} + \mathbf{T}^t\boldsymbol{\Sigma}^{-1}\mathbf{N}(u)\mathbf{T}$ (Dehak, Kenny, et al., 2011). The mean of the posterior is what we refer to as i-vector and can be seen as a projection of the utterance's shift from the background model in a low-dimensional space. These vectors are fixed-length, meaning that all utterances will be represented by the same number of dimensions, regardless of their length.

While in engineering applications further processing is usually done to improve performance (for instance, using a supervised Linear Discriminant Analysis to increase the separation between languages), we will use them on their own, removing all supervised training. We can thus interpret the resulting vectors as representing new unheard utterances as a deviation from the background speech distribution with no prior knowledge other than what was learnt during training.

2.B Appendix B: Expectation-Maximization algorithm for GMMs

EM is a recursive algorithm that allows to obtain maximum-likelihood estimates for the parameters of a model given incomplete data. In the case of GMMs, the goal of the EM algorithm is to estimate the weights, means and covariances of each component given an observed distribution of N data points, for which we do not know the underlying categories (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2009). In order to do this, the algorithm uses a latent variable z_{nk}, which in the context of a GMM corresponds to the underlying category of each observation, such that $z_{nk} = 1$ if observation n came from component k and $z_{nk} = 0$ otherwise. However, because the values z_{nk} are unknown, they are replaced by a "soft" class value $\gamma_n(k)$, which represents the probability of each data point n to have come from Gaussian component k.

The algorithm is initialized by setting arbitrary weights, means and covariances ω_k^0, μ_k^0, Σ_k^0 for each component (see Figure 2.22, Step 0).

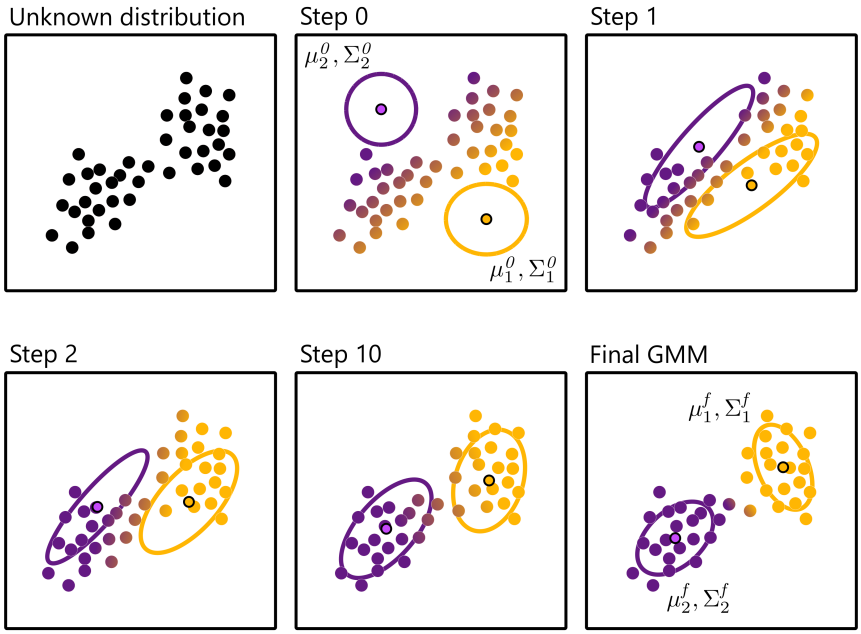


Figure 2.22: Illustrative example of the EM algorithm for estimation of parameters of a GMM with 2 Gaussian components.

In each successive iteration, the algorithm performs two steps: first, it computes the expected value of the latent variable, $\gamma_n(k)$, given the previous estimation of the parameters:

$$\gamma_n(k) = p(k|x_n) = \frac{\omega_k^{i-1}\mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}_k^{i-1}, \boldsymbol{\Sigma}_k^{i-1})}{\sum_{k=1}^{K}\omega_k^{i-1}\mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}_k^{i-1}, \boldsymbol{\Sigma}_k^{i-1})} \tag{2.11}$$

Then, it finds new parameters μ_k^i, Σ_k^i, that maximize the log-likelihood of the model given the expected values $\gamma_n(k)$. The updated parameters at step i are:

$$\boldsymbol{\mu}_k^i = \frac{\sum_{n=1}^{N}\gamma_n(k)\mathbf{x}_n}{\sum_{n=1}^{N}\gamma_n(k)} \tag{2.12}$$

$$\boldsymbol{\Sigma}_k^i = \frac{\sum_{n=1}^{N}\gamma_n(k)(\mathbf{x}_n - \boldsymbol{\mu}_k^i)(\mathbf{x}_n - \boldsymbol{\mu}_k^i)^{\mathrm{T}}}{\sum_{n=1}^{N}\gamma_n(k)} \tag{2.13}$$

Finally, new weights are computed:

$$\boldsymbol{\omega}_k^i = \frac{\sum_{n=1}^{N_k}\gamma_n(k)}{N} \tag{2.14}$$

These two steps are repeated until convergence.

2.C Appendix C: P-values from simulations in Exp. 4

Table 2.7: Median p-values obtained in i-vector-based simulations with filtered speech.

	italian	spanish	catalan	french	dutch	english	polish
spanish	0.023	•	•	•	•	•	•
catalan	0.0008	0.71	•	•	•	•	•
french	0.17	0.35	0.096	•	•	•	•
dutch	0.097	0.066	0.019	0.89	•	•	•
english	0.001	0.40	0.38	0.35	0.079	•	•
polish	0.004	0.16	0.002	0.86	0.62	0.12	•
japanese	$< .0001$	0.01	0.005	0.0001	$< .0001$	0.008	$< .0001$

Table 2.8: Median p-values obtained in i-vector-based simulations with natural speech.

	italian	spanish	catalan	french	dutch	english	polish
spanish	0.023	•	•	•	•	•	•
catalan	0.005	0.59	•	•	•	•	•
french	0.49	0.27	0.18	•	•	•	•
dutch	0.12	0.023	0.03	0.64	•	•	•
english	0.003	0.026	0.02	0.45	0.12	•	•
polish	0.008	0.0006	$< .0001$	0.35	0.14	0.02	•
japanese	0.15	0.007	0.001	0.65	0.18	0.43	0.001

Chapter 3

Dual language input and its impact on lexical development

3.1 Introduction

Quantity and quality of input, as well as environmental factors, have long been argued to play an important role in language acquisition (Hoff, 2006). For instance, word frequency and syntactic complexity in the speech that children hear, socioeconomic status, and maternal responsiveness, have all been found to influence language skills in monolingual children (Fernald, Marchman, & Weisleder, 2013; Hoff & Naigles, 2002; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Tamis-LeMonda, Bornstein, & Baumwell, 2001; Weisleder & Fernald, 2013). For infants growing up in a bilingual environment, language exposure may differ in many more aspects. Bilingual children might not only vary in the amount of input they receive in each language, but also in the contexts in which each of them is used, in the characteristics of language use of their communicative partners (such as language choice, or linguistic proficiency), and in the degree of similarity between the target languages, among many other properties. The large number of potential sources of variability makes describing the bilingual experience a difficult task. However, in a world where an estimated half of the population is bilingual (Grosjean, 2010), characterizing the bilingual input and its impact on language acquisition is of great importance.

So far, the most investigated aspect of bilingual input is the amount of exposure to each language (e.g., Cattani et al., 2014; David and Wei, 2008; Garcia-Sierra et al., 2011; Hoff et al., 2012; Hoff,

Welsh, Place, and Ribot, 2014; Marchman, Martínez, Hurtado, Grüter, and Fernald, 2017; Pearson et al., 1997; Place and Hoff, 2011, 2016; Poulin-Dubois et al., 2013; Thordardottir, 2011). While for monolingual children their entire input is in a single language, bilingual input is inevitably divided and varies from child to child in the proportion and context of each language. This means that - unless speakers in bilingual families are more talkative than in monolingual families - the amount of exposure to each language will be smaller than that of their monolingual peers. In a study of bilingual toddlers learning English and an additional language, Cattani et al. (2014) investigated the amount of exposure to English that is necessary to perform like monolingual English toddlers in several language tasks. Their findings showed that, in order to perform like their monolingual peers, English would need to comprise at least 60% of the bilingual's total input. Furthermore, the relative amount of exposure to a given language has been found to correlate with that language's phonological development (Garcia-Sierra et al, 2011), vocabulary size (Cattani et al., 2014; David & Wei, 2008; Hoff et al., 2012; Pearson et al., 1997; Place & Hoff, 2011, 2016; Poulin-Dubois et al., 2013; Thordardottir, 2011), and grammatical skills (Gathercole, 2002a, 2002b, 2002c; Hoff et al., 2012; Place & Hoff, 2016). Most studies investigating the amount of exposure in the bilingual's input have used parental questionnaires (such as the Language Exposure Questionnaire designed by Bosch and Sebastián-Gallés, 1997, or the Language Exposure Assessment Tool developed by DeAnda, Bosch, Poulin-Dubois, Zesiger, and Friend, 2016) to estimate the global percentage of each language throughout the child's life. However, a bilingual's language ratios may vary in time, due to life events such as moving from maternal care to attending a daycare center, or going on a trip to a country with a different language. In a longitudinal study of 13 French-English bilingual children from 1 to 3-years-old, David and Wei (2008) found that their relative productive vocabulary (e.g., the proportion of French words in their total lexicon) adapted to changes of language proportions in the bilingual's input throughout the months. Thus, linguistic abilities at a given time point in the child's life may be affected by recent changes in their input.

Other properties of the bilingual experience have been less explored, but a growing body of research suggests that many other factors beyond the relative amount of exposure can have an effect on language outcomes (De Houwer, 2018; Gathercole, 2014; Hoff & Core, 2013). One particular aspect that has been discussed since the beginnings of research on bilingualism is the impact of language choice of the primary caregivers. French linguist Maurice Grammont, as cited by Ronjat (1913), was perhaps the first to suggest that language separation by speaker - i.e., following a one-person-one-language (OPOL) approach - was necessary to guarantee successful bilingual development. For decades, this

has been a common advice given to parents raising bilingual children, but most research was based on case studies, often by linguists raising their own children (see reviews by Barron-Hauwaert, 2004 and Yamamoto, 2001).

Over the past 20 years, several studies have investigated the role of parental language separation with larger sample sizes and more systematic methods (Byers-Heinlein, 2013; De Houwer, 2007; Lyon, 1996; Place & Hoff, 2016; Yamamoto, 2001). In a large study with nearly 2000 bilingual families of school-aged children in a Dutch-dominant region of Belgium, De Houwer (2007) found that an OPOL separation of languages was neither sufficient nor necessary to guarantee that children would become actively bilingual. Indeed, the percentage of families where children only spoke the community's majority language was similar in families that followed an OPOL approach (26%) and those where both parents spoke both the majority and a minority language (21%). Unfortunately, the conclusions of this study are limited, as the data consisted of a short parental questionnaire, asking only which language(s) each parent used at home, and whether the child spoke only the majority or the minority language, or both. Furthermore, the study only examined children aged 6 to 10 years old, all of whom were attending Dutch-speaking schools. It is not uncommon for bilingual children to prefer using the language spoken at school (see for example Wong Fillmore, 1991); the role of language separation may thus be different at earlier developmental stages, when children spend substantially more time with their families. In a recent study of a heterogeneous bilingual population in Canada, Byers-Heinlein (2013) used a parental questionnaire to measure the frequency of language mixing produced by parents of bilingual toddlers. Their results revealed that intra-sentential mixing was negatively correlated with receptive vocabulary in 1.5-year-olds, and marginally so with productive vocabulary in 2-year-olds. Place and Hoff (2016) did not find clear evidence of a negative impact of parental language mixing using the same questionnaire in a group of Spanish-English bilingual 2.5-year-olds in the US. This may mean that language separation by speakers is only important during the first two years of life, but more studies are needed to reach conclusions on the impact of this variable.

Regardless of whether parents stick to a one-person-one-language rule or not, bilingual children may differ in how often both languages co-occur in time. Place & Hoff (2011, 2016) investigated this issue using a Language Diary method (originally designed by De Houwer and Bornstein, 2003) with Spanish-English bilingual families of 2-year-olds living in the US. In these studies, parents were asked to report every half-hour who spoke to the child and which languages were used. Time blocks were categorized as English-only, Spanish-only, or mixed (this last category indicated simply that both languages were

used, regardless of speaker). Seven diaries were collected for each child, each on a different day of the week, over the course of 7 weeks. The results from their first study (Place & Hoff, 2011), with a sample size of 29 children, revealed that the number of time blocks in which only one language was used was correlated with productive vocabulary size in that language.[1] Yet, the complementary measure of number of mixed blocks did not correlate with either language's lexical development. In a follow-up study with 90 toddlers (Place & Hoff, 2016), where mixed blocks were further categorized as English- or Spanish-dominant, a correlation was found between number of English-dominant blocks and several English language outcome measures,[2] while Spanish-dominant blocks were unrelated to Spanish language skills. As details about who was speaking each language during mixed blocks were not collected in either study, it is difficult to interpret which aspects of language separation may or may not influence development.

Some other properties of the bilingual environment that have been found to influence language development are the number of speakers of each language (Gollan, Starr, & Ferreira, 2015; Place & Hoff, 2011), the presence of siblings (Bridges & Hoff, 2014; Silven, Voeten, Kouvo, & Lunden, 2014), parental strategies and attitudes towards bilingualism (Juan-Garau & Perez-Vidal, 2001; Nakamura, 2016) and the use and status of each language in the community (Gathercole & Thomas, 2009).

Because of the large amount of variability within and across bilingual populations, it is hard to draw general conclusions about the impact of specific factors on bilingual language development from any specific study. What is true for a given population may not be true for another one. For instance, as argued by Place and Hoff (2016), the lack of an effect of language mixing in their Spanish-English bilinguals may be due to the fact that mixing (and specifically code-switching) is a common behavior in the community where the study was conducted. This behavior, however, may not be common in the heterogeneous bilingual population studied by Byers-Heinlein (2013). Thus, despite great progress made so far in identifying potential factors that could impact bilingual development, much work remains to be done. As more studies continue to explore these and other properties of language input in new groups of bilingual children from diverse communities and at different developmental stages, the bigger picture will begin to form.

The goal of the present study is to further the characterization of the bilingual exposure, by exploring

[1] Additionally, English-only blocks correlated with grammatical complexity in English, but Spanish-only blocks did not correlate with the equivalent measure in Spanish.

[2] In Place and Hoff (2016), five different measures of language skills were used: productive vocabulary size, MLU3, grammatical complexity (all these based on the CDI), auditory comprehension (PLS-4) and picture naming (EOWPVT).

different properties of the dual input in a group of 11-month-old infants who are regularly exposed to French and an additional language. In order to capture different quantitative and qualitative aspects of their input, we designed a modified version of the Language Diaries previously used by Place & Hoff (2011, 2016), and complemented it with a language environment questionnaire. As in previous diary studies, we asked parents to report every half-hour all the people who spoke to the child. However, in contrast to the diaries used by Place & Hoff (2011, 2016), we asked parents to specify, for each speaker and each time block, the language used to talk to the child and the language used to talk to other people. This modification allows us to disentangle effects of co-occurrence of the two languages in time from within-speaker effects of dual language use. Using various measures of bilingual input derived from the diaries, we explore the sources of variability that characterize the dual exposure in this heterogeneous population, and investigate how these may influence lexical development.

3.2 Methods

3.2.1 Subjects

Fifty-nine families with bilingual 11-month-old infants participated in this study (26 girls, 33 boys; mean age: 338 days, range: 319 - 356 days). Four additional families were excluded from analysis due to incorrect completion of the diaries, and 20 additional families did not send one or both diaries back. All infants heard both French and an additional language (L2)[3] on a regular basis, and were being raised in the Paris area. The additional languages that the infants were exposed to were the following: Spanish (n = 13), English (n = 10), Italian (n = 9), German (n = 7), Polish (n = 3), Arabic, Catalan, Portuguese, Romanian, Russian (n = 2 each), Bulgarian, Greek, Hungarian, Japanese, Mandarin, Swedish, and Wolof (n = 1 each). Out of the 59 participants, 39 heard their L2 mainly from their mothers, 11 from their fathers, 6 from both parents, and 3 from a nanny. At home, the percentage of exposure to L2 ranged from 15% to 98% of their input, with a mean exposure of 52%. Outside of home, however, most infants heard a majority of French, as it is the dominant language in the region they were being raised in. French comprised an average of 84% of their outside of home input (range: 40% - 100%).

[3]We will refer to the additional language as L2, but this is not intended to imply that the L2 was learnt after French, nor that it is any less important in the child's input. Alternative labels for the additional language in the literature include AL (for Additional Language) and Language Alpha.

3.2.2 Materials and procedures

The families were invited to come to the babylab prior to their participation in the Language Diary study. During their visit, the child participated in a laboratory experiment, and the primary caregiver was given the materials and instructions for the completion of the diaries. Additionally, while at the babylab, each family completed a custom-made Language Environment Questionnaire to collect complementary information regarding the general language background of the child, as well as a short vocabulary questionnaire in French and, when available, the adaptation of this questionnaire in the child's additional language. We describe each of these assessment tools below.

3.2.2.1 Language Diaries

Each diary was constructed as a booklet containing one page per half-hour slot, beginning at 7:00 in the morning and finishing at 20:30 in the evening. Each page contained five rows, which served to annotate each of the speakers that interacted with the child and their language use (see a sample page in Appendix 3.A). If more than five people were present at a given time, the four people who interacted the most with the child would be noted in the first four rows, and a summary of the remaining people would be noted in the fifth row (number of people and average language use). This was done to simplify the task of the annotator, who would have otherwise been required to keep track of the language use of a large number of people, possibly leading to inaccuracies in their report.

In each row, speakers were identified by their roles in the child's life, i.e., as "mother" or "aunt # 1"[4]. For each speaker, two columns were used to annotate their language use, the first one corresponding to the language(s) used to speak to the child, and the second one to the language(s) spoken to other people in the presence of the child. A 5-point scale was given as options of language use. To avoid any potential ambiguities, the booklets were adapted to each language pair by writing the name of the L2 explicitly, e.g., *only French*, *mostly French*, *both equally*, *mostly English*, *only English*. Two additional options were given to cover alternative scenarios, *none* (i.e., no language was spoken) and *other (specify)*. Finally, two boxes at the bottom of each page were provided to indicate the location and activity (e.g., "In the kitchen, having breakfast"), as well as any additional comments. The diaries were written in French, and translations were made available in English and Spanish.

[4]This procedure was done to guarantee that we could correctly identify individual speakers while preserving their anonymity. Likewise, children were identified by an ID noted at the front of the booklet.

The parents were asked to complete two diaries, one on a week day and one during a weekend day of their choice, within the month following their visit. The chosen days should be as typical as possible, in order to capture the child's daily routine. They were given two booklets which were to be sent back to the lab by mail as soon as they were completed.

3.2.2.2 Language Environment Questionnaire

A detailed Language Environment Questionnaire was designed to collect information regarding the child's general language background (see full questionnaire in Appendix 3.A). For abbreviation, we will refer to this questionnaire as LEQ, but it should not be confused with the *Language Exposure Questionnaire* (also known as LEQ) designed by Bosch and Sebastián-Gallés (1997).

Our questionnaire contained questions regarding the family composition and the languages used by parents, siblings and other caregivers. Additionally, we collected information about the 4 adults who most regularly interacted with the child. This included a measure of their proficiency in each of the two languages (noted in a 6-point scale ranging from *doesn't speak the language* to *native speaker*), an estimation of the hours per day spent with the child during week and during weekend days, and a measure of their language use when talking to the child and when talking to other people, using the same scale provided in the diaries.

3.2.2.3 Vocabulary questionnaire

To assess the vocabulary of the infants, we used a modified version[5] of the European French short-form adaptation of the MacArthur CDI (Kern, Langue, Zesiger, & Bovet, 2010). Parents were instructed to answer whether the child understands (and, additionally, uses) each of the words in the list. In the case of English, Spanish and Portuguese bilinguals, families were also given the corresponding short-form adaptations[6] in those languages (Fenson et al., 2000; Frota et al., 2015; Jackson-Maldonado, Marchman, & Fernald, 2013). Each form was filled in by the main caregiver who spoke the respective language to the child.

[5]The original French short-form for assessing 12-month-olds designed by Kern et al. (2010) contains 81 words. In this modified version we have included all 81 words plus 10 additional words that were being used in a behavioral experiment the infants participated in.

[6]As with the French form, these forms were also extended to include the translation of some of the French words used in a behavioral experiment. The number of words in each language (and, in parenthesis, the original number of words in each form) was: 93 (90) words in English, 92 (90) in Portuguese, and 113 (104) in Spanish.

3.2.3 Coding and pre-processing

Diaries were coded in long format, with one entry per time slot and per speaker. Each speaker was assigned a 3 letter code, e.g., MOT for mother and FAT for father. Given the big inconsistencies in the way the number of daycare workers present during daycare hours was reported, we decided to group them all into one single row per time slot. Note that this does not affect calculations of language proportions during those hours, as French was the only language spoken at the daycare centers that the participants attended.

Speakers' language use was translated into French and L2 percentages. While the exact proportions of each language used in a given time block are not known, we estimated them as follows: *only French* is translated to 100% French - 0% L2, *mostly French* is translated to 75% French - 25% L2, *both equally* is translated to 50% French - 50% L2, and so on. Note that these values, although not accurate, will never be off by more than 25% with regards to the real percentages. Additionally, for each speaker in each time block, we defined a measure of *within-speaker language purity* (WSLP) as the percentage of the most spoken language[7]. For instance, if a person spoke 75% French and 25% L2 (or 25% French and 75% L2), then their WSLP value would be 75%. Thus, WSLP ranges from 50% to 100%. This measure intends to capture how often speakers use both languages within half an hour. Finally, time blocks were classified as at-home or out-of-home based on the description of the place and the activity provided in the diaries.

Before processing the data, we filtered the diaries by keeping only close relatives and caregivers of the child, i.e., parents, siblings, grandparents, aunts/uncles, nannies and daycare staff. From each diary, daily averages were computed as follows: First, an average of language use (% French, % L2, and WSLP) over all speakers was obtained for each half an hour. Here we have made the assumption that the amount of speech is divided between all speakers within a given period of time, thus keeping the amount of speech per half-hour constant. Additionally, for each block we defined a measure of *within-block language purity* (WBLP) as the percentage of the most spoken language, regardless of speaker. Next, daily averages were calculated by averaging over all time blocks. These calculations were done separately for direct speech and indirect speech. An example of this process is shown in Figure 3.1.

[7]Note that this measure is very similar to the measure of purity typically used in clustering analyses, such as the one we reported in Carbajal, Dawud, Thiollière & Dupoux (2016).

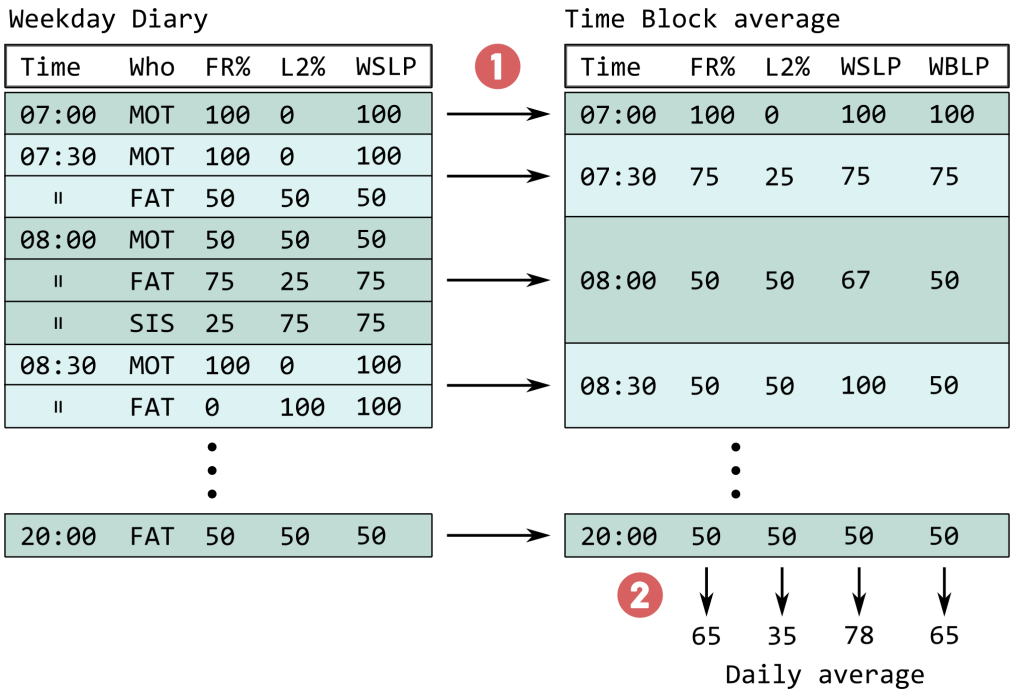


Figure 3.1: Example of computation of daily averages from direct speech during a weekday. (1) Mean percentage of French (here noted as FR), L2, within-speaker language purity (WSLP) and within-block language purity (WBLP) are computed for each time block (i.e. 30-min period) by averaging over all speakers. (2) Daily averages of percentage of French, L2, WSLP and WBLP are obtained by averaging over all time blocks.

Finally, weekly averages of all four measures were estimated by weighting the daily averages of weekdays $\times$ 5 and of weekend days $\times$ 2. Here we have made the assumption that infants' routine reported on the weekday is likely to be repeated throughout the 5 days of the working week (and analogously for the weekend). Alternatively, if no assumptions regarding the routine were made, both diaries (week and weekend days) could be given equal weights. A correlation analysis of the average exposure computed with and without applying different weights shows that the two measures are highly correlated ($r = 0.95$, $p < .0001$, see Figure 3.15 in Appendix 3.C). Thus, the non-weighted measure would likely yield similar results in subsequent analyses.

Additionally, separate averages were obtained by speaker (i.e., by computing the mean FR%, L2% and WSLP over all rows of a given speaker over 2 days, using the same weights as for the weekly averages) and by at-home or out-of-home location. The latter was done by first computing the time-block averages as explained in the first step of the daily averages, and then calculating averages over the blocks where the child was at home and out of home, separately.

3.3 Results and discussion

In this section, we will first describe properties of the bilingual environment based on the Language Environment Questionnaire, followed by a characterization of their input based on the Language Diaries. Finally, we will examine potential correlations between their vocabulary and their exposure (derived both from the LEQ and from the diaries) in each language.

3.3.1 Language Environment Questionnaire

The LEQ provides a first overview of the language background of the bilingual infants. As mentioned in the Methods section, the percentage of exposure to each language at home covered a wide range, from 15% L2 – 85% French, to 98% L2 – 2% French. For the great majority of the infants (49 out of 59), their parents followed roughly a one-parent-one-language approach (OPOL), that is, one parent spoke mostly or only French, and the other spoke mostly or only L2. Out of the 10 remaining cases, 8 correspond to families where both parents spoke the same language (3 French families whose children learnt L2 from a nanny, and 5 L2 families whose children learnt French from a nanny or at daycare), and finally only 2 had one parent who spoke both languages equally often. Table 3.1 shows the reported language behavior of the parents when talking to the child, depending on their proficiency in each language. As can be seen in the table, bilingual parents generally chose one language to communicate with the child, most often the language not spoken by the other parent. In the case of both bilingual parents, each one chose a different language.

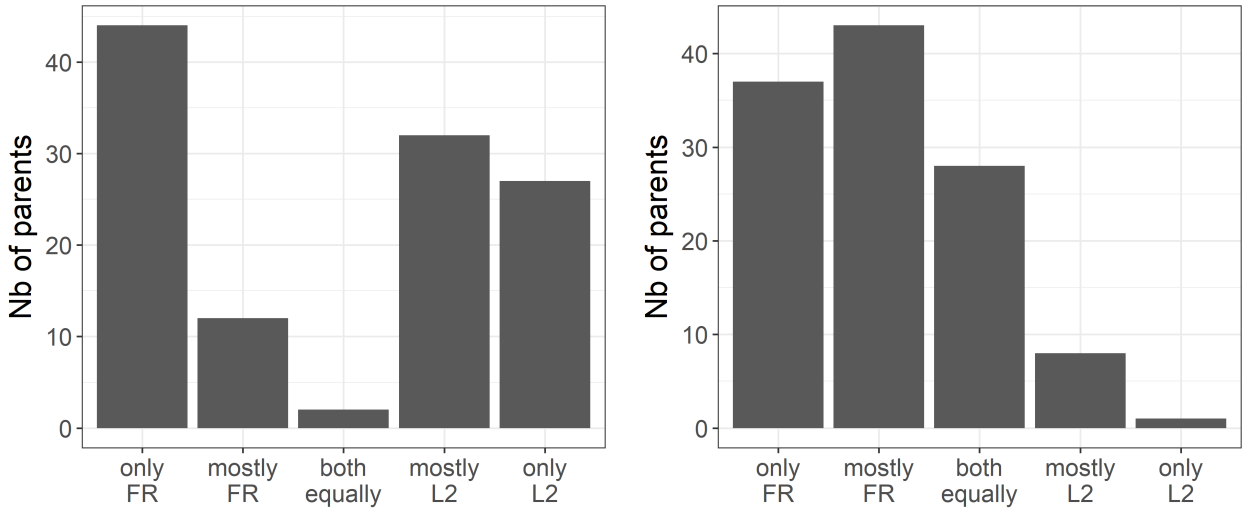
We then examined differences in the way parents used the languages with the child and with other people. Figure 3.2 shows a histogram of the languages used by fathers and mothers when addressing their infants (Fig. 3.2a) and when talking to other people in front of the child (Fig. 3.2b) as reported in the LEQ. As can be seen in the figure on the left, in spite of generally adhering to an OPOL division of languages, some parents reported also speaking a small amount of the other language when talking to their child (cases *mostly-FR*, *mostly-L2*). This behavior was more common in L2 speakers than in French speakers, which may be partly due to the influence of the community language, and partly due to the fact that many L2-speaking parents were actually bilinguals (53%, compared to only 12% of French-speaking parents). Moreover, as shown in the figure on the right, parents used both languages more often when talking to other people in front of the child, as can be seen by a prominent

Table 3.1: Summary of parents' reported language use.

Parents' languages	N	Main language used with the child	
		Mother	*Father*
Both French monolinguals	3	French	French
French mother, L2 father	3	French	L2
French mother, Bilingual father	5	French	L2
Bilingual mother, L2 father	5	French ($n = 3$), L2 ($n = 2$)	L2
Bilingual mother, French father	20	L2	French
L2 mother, French father	14	L2	French
L2 mother, Bilingual father	1	L2	Both equally
Bilingual mother	1	Both equally	-
Both bilinguals	4	L2 ($n = 3$), French ($n = 1$)	French ($n = 3$), L2 ($n = 1$)
Both L2 monolinguals	3	L2	L2

Note: We have coded parents as speakers of a given language if they self-reported a native-like or native level of proficiency in that specific language. If both languages had at least native-like proficiency, we coded them as bilinguals, otherwise they were coded as monolinguals.

increase in *both-equally* responses, and a decrease in *only-FR* and *only-L2* responses. Finally, when communicating to other people, French was used much more often than L2. This is not surprising given that French is the language spoken by the community.



(a) Infant-directed speech.

(b) Indirect speech.

Figure 3.2: Histograms of language use by parents as reported in the LEQ (FR = French, L2 = additional language).

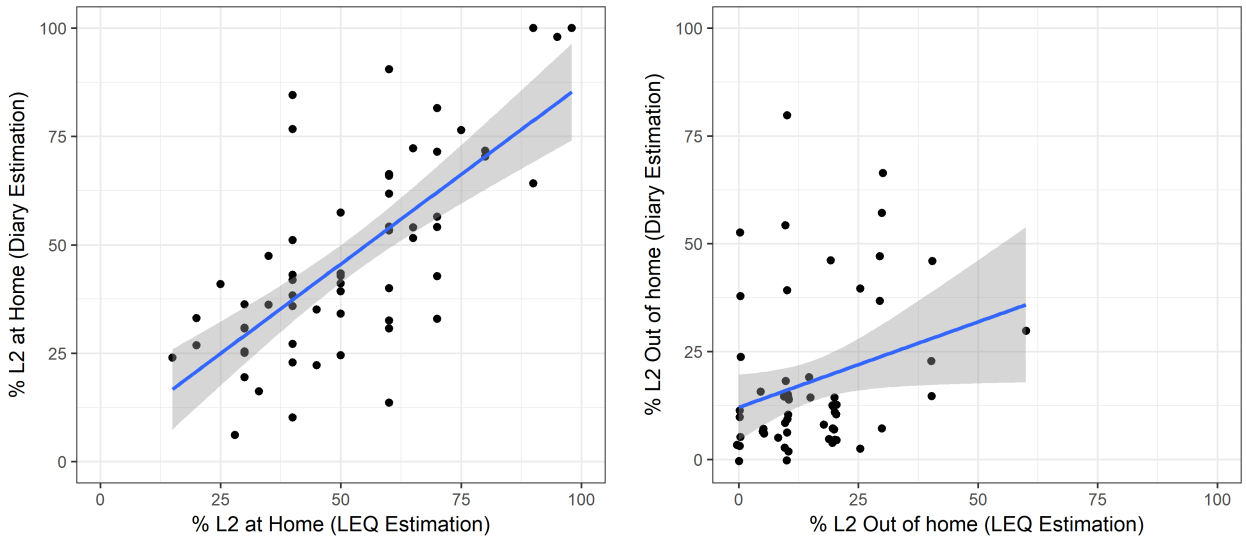
Next, we looked at the household composition. The great majority of the infants lived alone with their parents and siblings (if any). Only 2 families did not follow this pattern: one child lived with her mother and an aunt, and another one lived with her mother and grandparents. Out of 59 children,

43 were first-borns, 11 had one older sibling, and the rest had 2 or more older siblings. In general, siblings spoke more French than L2, with 62% speaking mostly or only French, and the rest speaking both languages equally. No siblings were reported to speak mostly or only L2.

Finally, we examined secondary caregivers (nannies and daycare centers). Out of 59 children, 23 attended a daycare center regularly and had no nanny, 21 had a nanny but did not attend daycare, 8 attended daycare and also had a nanny, and 7 did not have any secondary caregivers. French was the only language spoken by daycare staff. On the other hand, out of 29 nannies, 22 spoke French to the child, 6 spoke L2, and only one spoke both languages. Thus, with the exception of these 7 infants with L2 or bilingual nannies, most children heard their L2 primarily from their close family.

3.3.2 Language Diaries

First, in order to investigate the validity of the diaries as a way of estimating the infant's language input, we compared the exposure to L2 both at home and out of home as calculated from the diary data, against the percentages reported by the parents in the LEQ. Since in the questionnaire we asked parents to consider all the speech the child may have heard when estimating the exposure to each language, here we computed the diary estimates by pooling direct and indirect input. Figure 3.3 shows a comparison of both estimates.



(a) Speech heard at home.

(b) Speech heard out of home.

Figure 3.3: Correlation between parental report (LEQ) and Language Diary estimation of infants' exposure to L2. A small horizontal jitter (width = 0.5) was added to the out-of-home figure (b) to show overlapping points.

For at-home exposure (Fig. 3.3a), we found a correlation of $r = 0.70$ ($p < .0001$) between the parental and the diary estimations, indicating good agreement between both methods. For out-of-home input (Fig. 3.3b), a weak correlation was found ($r = 0.26$, $p = .046$). However, this is likely due to the fact that most infants had a majority of French exposure outside their homes, as can be seen both in the parental estimation and in the diary averages (for 71% of the children, French comprised 75% or more of their exposure out of home according to both measures simultaneously), thus the range of possible values is quite narrow. Furthermore, when estimating how much French and L2 infants hear out of home, parents probably took into consideration a great amount of indirect French input from speakers who do not regularly interact with the child (such as people on the street and in shops), while we only kept speech from regular speakers in our estimation. Thus it is both possible that parents overestimated the weight of indirect French input and that we underestimated it. However, since all infants live in the same community, they are likely to have similar amounts of indirect French input from strangers out of home, and so our estimations will all be affected by a similar offset. Finally, since we will analyze direct and indirect input separately in the remainder of the study, this difference should not be problematic. We conclude that the diary estimates are overall in good agreement with parental reports and are thus a reliable method to explore bilingual infant's exposure.

Next, we computed the average exposure to each language separating direct from indirect input (regardless of whether it occurred at home or out of home). The distribution of exposure percentages for each language in direct speech is shown in Figure 3.4, and for indirect speech in Figure 3.5. While overall the range of language exposure in direct speech covers a wide spectrum (L2 min: 12%, max: 90%), on average infants heard more French than L2 (mean FR: 60%, mean L2: 40%). Three infants also heard a small amount of a third language, with a maximum of 18% of their direct input.

Moving on to indirect speech (Fig. 3.5), it is clear that infants heard a great majority of French (mean FR: 77%, mean L2: 20%). For 37 out of 59 children, the percentage of French in their indirect input was higher than in their direct input by at least 5 percentage points. For 14 others there was a small increase of less than 5 percentage points, and finally for 8 children there was a decrease in the amount of French (2 of which were only caused by the presence of a third language, and not by an increase in L2). In total, only 6 children had higher L2 in indirect input. However, infants' direct and indirect input showed a moderate correlation ($r = 0.49$, $p < .0001$), indicating that the highest values of indirect L2 input correspond to infants who also had relatively high amounts of direct L2 input.

Infants' language exposure is not only characterized by how much of each language they hear, but

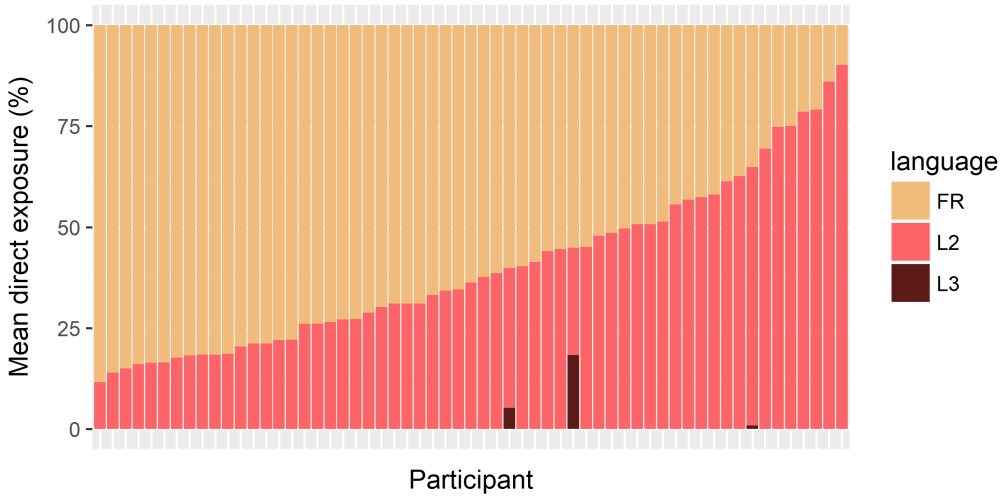


Figure 3.4: Estimates of infants' weekly average percentages of French (FR), L2 and L3 in direct input based on their diary data.

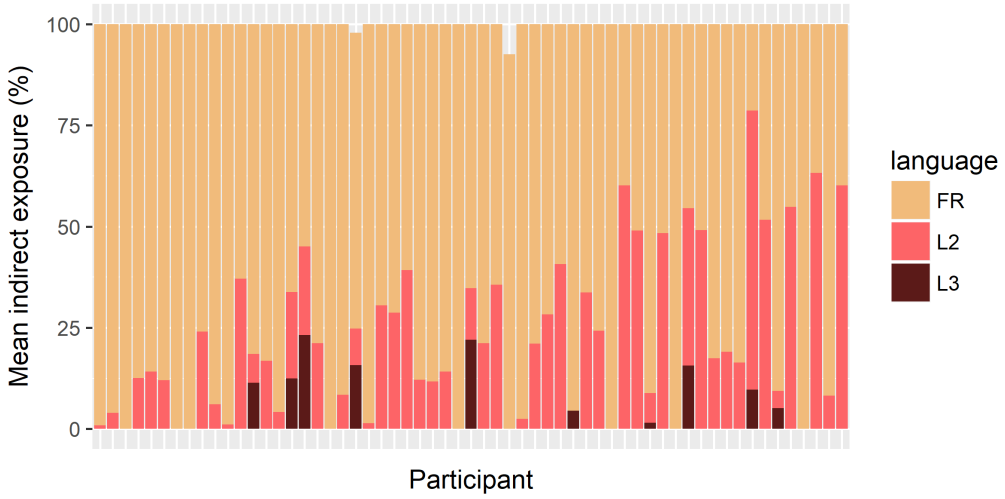


Figure 3.5: Estimates of infants' weekly average percentages of French (FR), L2 and L3 in indirect input based on their diary data. Gaps at the top of the bars for two infants correspond to small percentages of a 4th language.

also by how often the languages co-occur throughout the day. We thus examined the frequency of co-occurrence of both languages throughout the two days. Figures 3.6 and 3.7 show the percentage of direct French and L2 for each half-hour block throughout week and weekend days for two infants with very similar average exposure to each language (infant *BB016*, Fig. 3.6, Mean FR: 43%, Mean L2: 57%; and infant *BB026*, Fig. 3.7, Mean FR: 49%, Mean L2: 51%).

In spite of the similarity of their amount of exposure to French and L2, these infants have drastically different experiences. In the case of *BB016*, the two languages are well separated in time: at home, L2 is the only language spoken, while at daycare (from 9am to 6pm on weekdays), only French is used. On the other hand, *BB026* often hears both languages used within the same half an hour. To

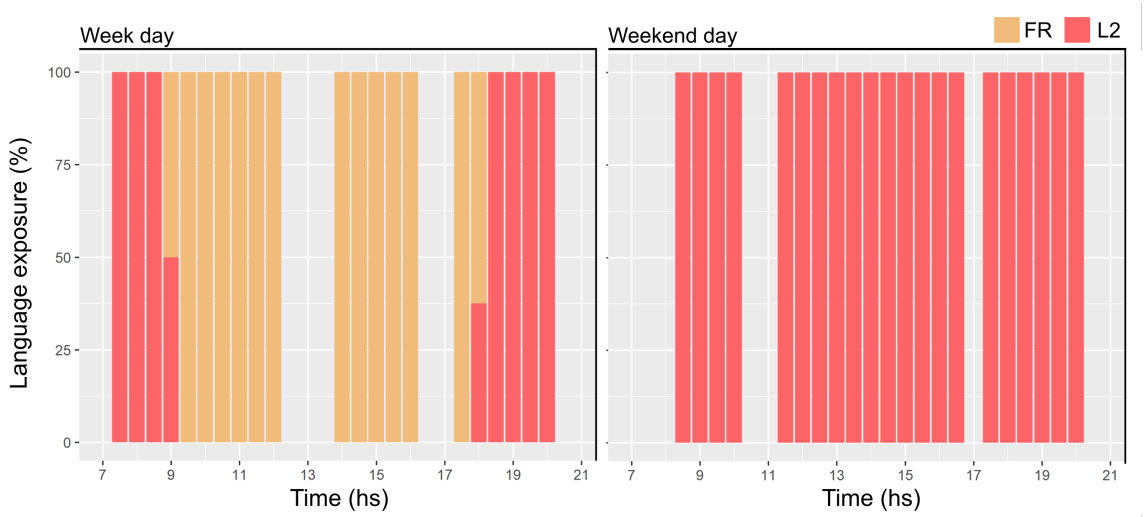


Figure 3.6: Proportions of French and L2 exposure (direct speech) throughout the weekday and weekend day for baby *BB016*. Empty time blocks represent time during which the child did not receive any input (e.g., during naps).

quantify this difference, we computed the weekly average of within-block language purity (WBLP). Infant *BB016* had a 97% average WBLP, while infant *BB026* had a much lower block purity of 77%.

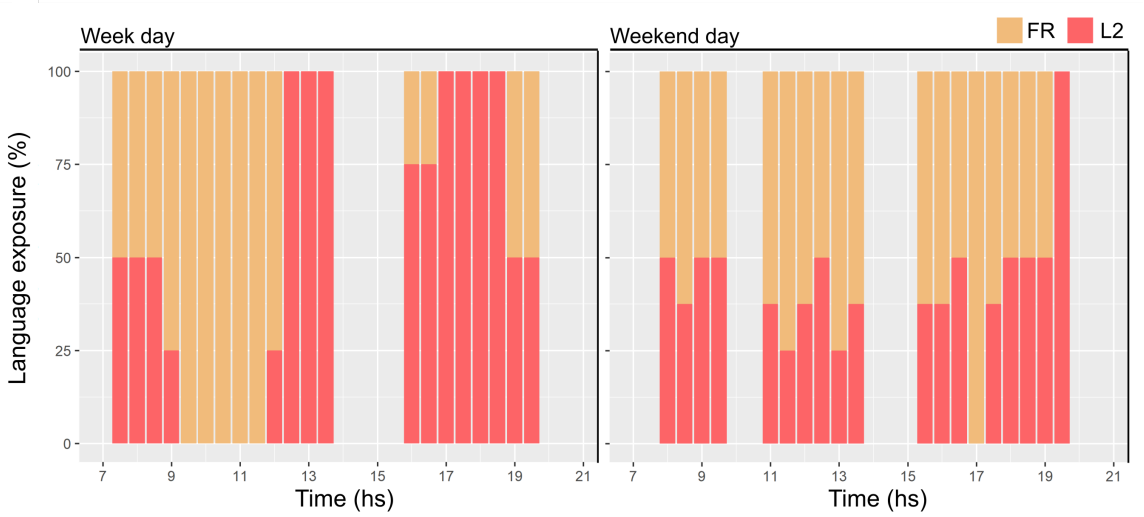


Figure 3.7: Proportions of French and L2 exposure (direct speech) throughout the weekday and weekend day for baby *BB026*. Empty time blocks represent time during which the child did not receive any input (e.g., during naps).

In Figure 3.8 we show a histogram of average WBLP (direct speech) for all infants in our study. On average, infants had a global WBLP of 84% (SD = 7%) in their direct speech. Only 5 out of 59 children had an average language purity above 90%, meaning that the great majority of the children often heard both languages spoken within the same half-hour. The average WBLP in indirect speech was only slightly higher, with a mean of 88% (SD = 11%).

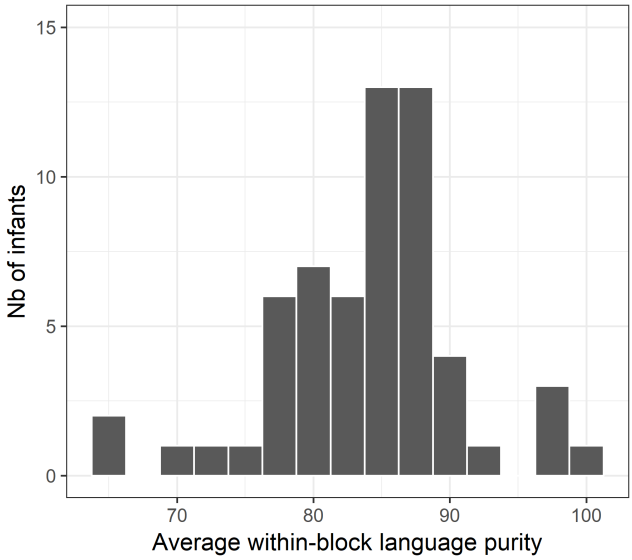


Figure 3.8: Histogram of weekly average within-block language purity (WBLP) in direct input.

The co-occurrence of both languages within the same 30-minute block may be due to the presence of bilingual speakers, or to the presence of monolingual speakers of two languages. To untangle these two scenarios, we compute the average within-speaker language purity (WSLP). In Figure 3.9 we show the distribution of average WSLP that infants were exposed to. To illustrate the difference between within-block language purity and within-speaker language purity, let's compare two infants who have similar averages of WBLP: *BB082* (77%), and*BB002* (79%). In the environment of infant *BB082*, speakers rarely used both languages when talking to the child. The average WSLP produced by the three most frequent speakers when talking to the child was 100% from his father and daycare, and 99.3% from his mother. Overall, across two days and averaging all frequent speakers, this infant had a mean WSLP of 99.6%. On the other hand, in the environment of infant *BB002*, speakers sometimes used both languages. While this child's father never used both languages in the same half-hour block (average WSLP of 100%), his mother and sister often did, with an average WSLP of 84% and 73%, respectively. Across all frequent speakers throughout the two days, *BB002* heard an average WSLP of 85%.

Finally, it should be noted that within-block language purity and within-speaker language purity are moderately correlated ($r = 0.49$, $p < .0001$), as infants whose caregivers frequently use both languages will inevitably encounter both languages co-occurring in time more often.

Next, we compared parents' language use estimated from the diaries against what had been reported in the LEQ (which we have shown previously in Figure 3.2). Figure 3.10 shows the relationship

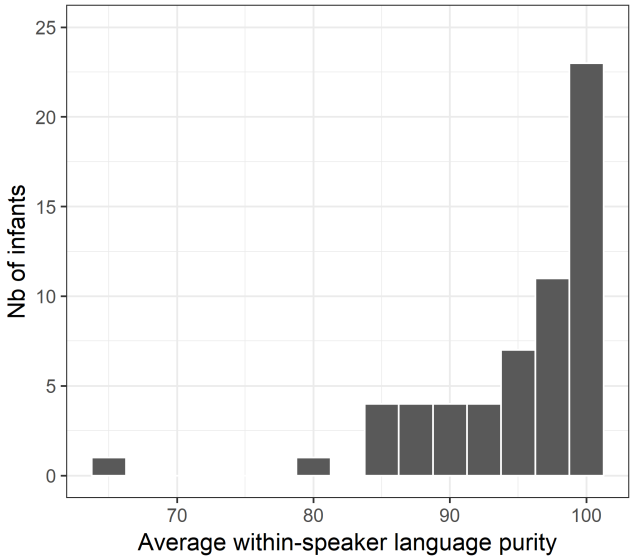
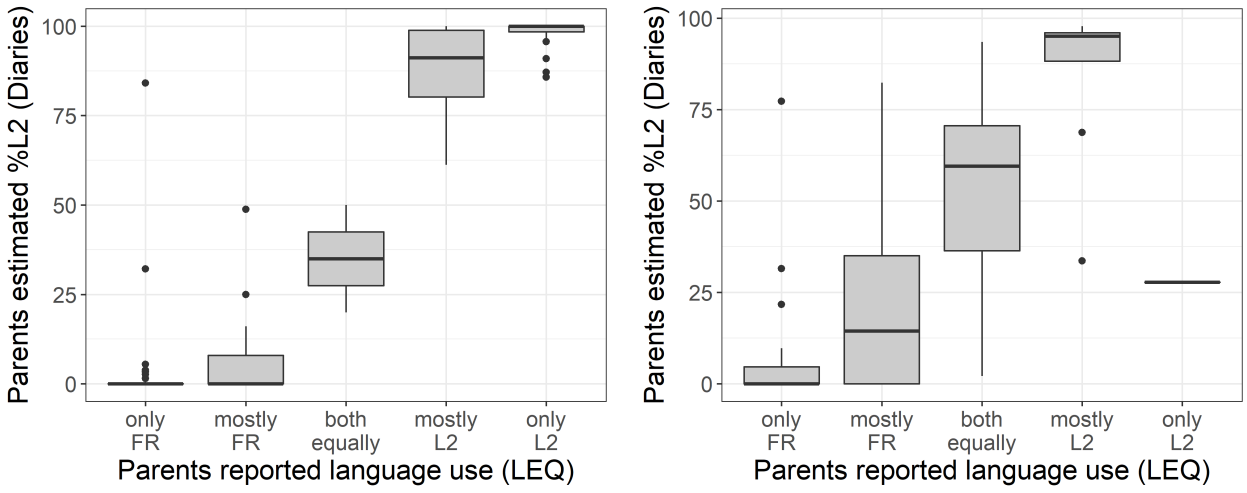


Figure 3.9: Histogram of weekly average within-speaker language purity (WSLP) in direct input.

between these two estimations for direct (3.10a) and indirect (3.10b) speech, respectively. The diary estimations of parental language use when talking to the child (left panel) were overall similar to what parents reported in the LEQ, as confirmed by a very high correlation between both measures ($r = 0.95$, $p < .0001$). Particularly, parents who reported in the LEQ using mostly or only one language were indeed found to use mainly that language in the diaries, with very few exceptions. In the case of indirect speech (right panel), parental estimations also showed good agreement with the diary estimations, ($r = 0.70$, $p < .0001$), with the exception of one parent who reported using only L2 to talk with other people, while he was observed to use a majority of French in the diaries. This difference could be due to an error in the LEQ report, as it is indeed unlikely that an adult speaks L2 exclusively while living in France.

Finally, we computed the number of speakers of each language that infants encountered throughout the two days. Table 3.2 shows the number of speakers who spoke on average a majority of a given language (i.e., >50%) when talking directly to the child, taking into account all speakers encountered across two days, as well as restricted to relatives and caregivers only. As can be seen, regardless of whether all speakers or only relatives and caregivers were counted, infants encountered significantly more French speakers than L2 speakers. Only 12 out of 59 infants encountered additional L2 speakers who were not included in the relatives and caregivers list mentioned before, while 35 of them encountered additional French speakers. This indicates that, for the majority of the infants, their most frequent speakers are the only source of L2, while they are (unsurprisingly) more likely to find new French speakers in their environment. We found weak correlations between the number of frequent speakers of a given



(a) Infant-directed speech.

(b) Indirect speech.

Figure 3.10: Correlation between parental report (LEQ) and Language Diary estimation of parental language use.

language and the percentage of exposure to that language ($r = 0.37$, $p = 0.004$ for French; $r = 0.25$, $p = 0.06$ for L2).

Table 3.2: Number of speakers who spoke a majority of French or L2.

	N speakers >50% FR	N speakers >50% L2	T-test (paired samples)
All speakers	$M = 4.7$ (range: 1–21)	$M = 1.8$ (range: 0–12)	$t(58) = 5.77$ ($p < .0001$)
Relatives & caregivers	$M = 2.7$ (range: 1–8)	$M = 1.4$ (range: 0–7)	$t(58) = 5.19$ ($p < .0001$)

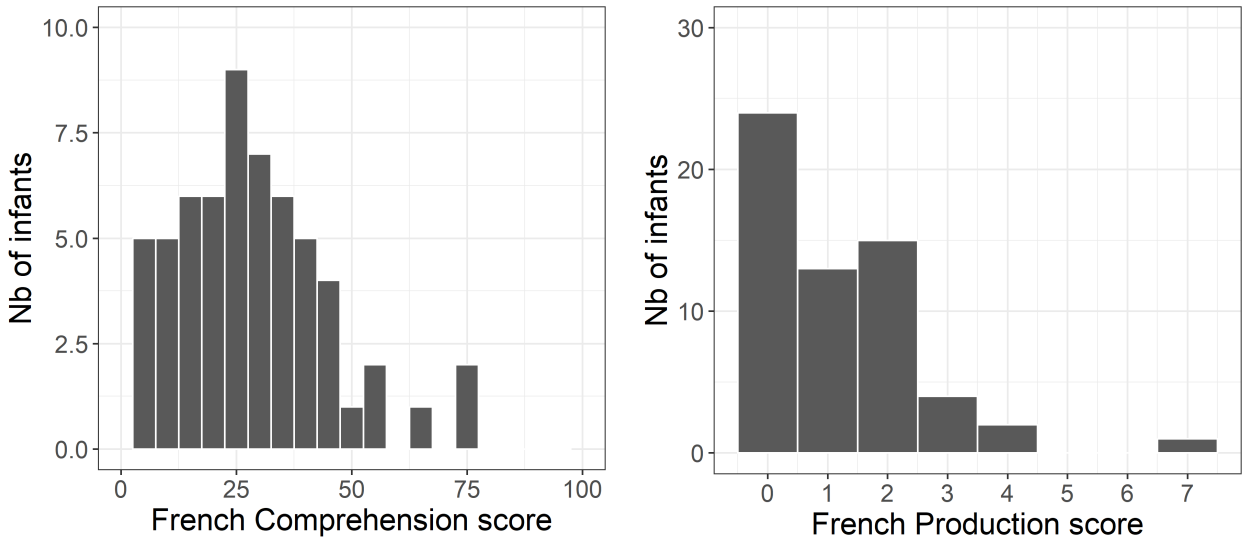
In conclusion, so far, we have shown that bilinguals' language experience can vary greatly between infants, despite most of them being raised under an OPOL pattern in a mainly monolingual community. In particular, children differ in the amount of exposure to each language, in the frequency of language overlap within time blocks and within speakers, and finally in the number of speakers that provide input in each language. In the following two subsections we will examine their vocabulary scores, and the potential impact of the factors described so far on this outcome measure.

3.3.3 Vocabulary scores

For each infant, we computed a French comprehension score as the sum of all words that parents reported the infant understands[8], and a production score counting only the words the infant produces. Figures 3.11a and 3.11b show the distribution of comprehension and production scores, respectively.

[8]Words counted as being understood both if parents reported them as only understood (not yet produced) and also if they were reported as produced by the child.

On average, infants were reported to understand 28 out of 91 words (range: 4 – 75). Production scores were significantly lower, with a mean of 1 word (range: 0 – 7), and a mode of 0 words produced.



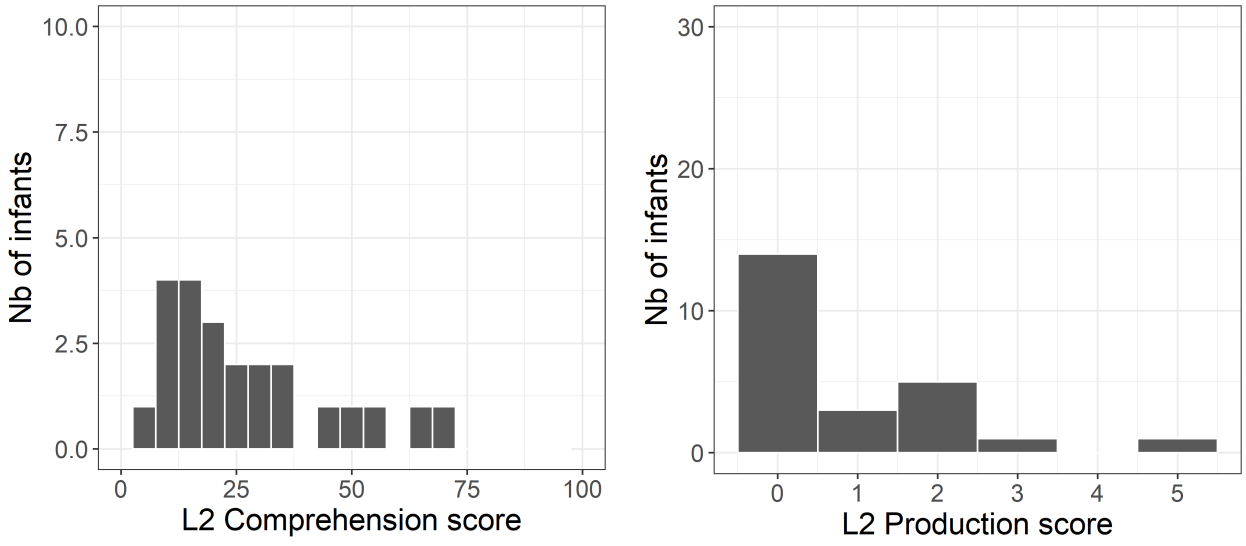
(a) (b)

Figure 3.11: Histograms of French vocabulary size. (a) Receptive vocabulary. (b) Productive vocabulary.

For comparison, we looked at the comprehension and production scores of 64 11-month-old monolinguals (mean age: 337 days) who participated in a laboratory study in our babylab during the same period. Their mean comprehension (M = 28 words, range: 7 – 83) and production (M = 2 words, range: 0 – 11) scores were not significantly different from the vocabulary scores of our bilingual infants (comprehension: $t(120.8) = 0.18$, $p > 0.1$, production: $t(106.9) = 1.40$, $p > 0.1$). This comparison suggests that, overall, our bilinguals are within normal ranges of vocabulary for their age, at least in one of their two languages.

For 24 out of 59 infants, we also obtained their L2 vocabulary questionnaire (namely in Spanish, English, and Portuguese). As these questionnaires differ in their total number of words, in order to compare them we normalized the vocabulary scores as follows: first, we divided each score by the total number of words in the corresponding L2 questionnaire, and then we multiplied it by 91 (the number of words in the French questionnaire). In Figures 3.12a and 3.12b we show the distribution of comprehension and production L2 scores in this normalized scale. On average, infants were reported to know 26 L2 words (range: 0 – 71) and to produce one L2 word (range: 0 – 5). A t-test revealed a marginally smaller receptive vocabulary in L2 compared to French for this subset of infants ($t(23) = -1.96$, $p = 0.06$), but no difference in their production scores ($t(23) < 1$, $p > 0.1$). Furthermore, we found a high correlation of comprehension scores across their two languages ($r = 0.81$, $p < .0001$), but

no correlation of their production scores ($r = 0.34$, $p > 0.1$).



(a) (b)

Figure 3.12: Histograms of L2 vocabulary size. (a) Receptive vocabulary. (b) Productive vocabulary. For comparability, vocabulary scores were normalized by the number of words in each language's form and multiplied by the number of words in the French form.

Given the low variability in infants' production scores, we will focus our analysis (mainly) on comprehension scores. For all subsequent analyses, we exclude infants whose French comprehension scores are considered outliers following Tukey's criterion[9]. This criterion resulted in two exclusions.

3.3.4 Effects of Language Exposure on Vocabulary Scores

To explore possible effects of bilingual language exposure on language development, we computed the correlation between infants' French and L2 vocabulary scores on the one hand, and the different measures of bilingual exposure that we have defined so far on the other hand, i.e., total percentage of exposure to French and L2, average within-block language purity, average within-speaker language purity[10], number of frequent speakers of each language, and maternal speaking behavior. Additionally, we include a measure of number of blocks with only French or only L2 direct input, as it was found to have a significant effect on vocabulary size in Place & Hoff (2011, 2016). It should be noted that these correlations are intended as an exploratory analysis, and any observed effects should be confirmed in future studies. Table 3.3 shows the results of the correlations for French, and Table 3.4 shows the correlations for L2. These correlations based on measures derived from direct speech. The

[9]Tukey's criterion defines outliers as points falling outside the range [Q1 – 1.5*IQR; Q3 + 1.5*IQR], where Q1 is the first quartile of the distribution, Q3 is the third quartile, and IQR = Q3 – Q1.

[10]We exclude from this correlation analysis one child with an outlier value of WSLP.

corresponding tables for indirect speech measures can be found in Appendix 3.D.

Table 3.3: Correlations between measures of French direct exposure and French vocabulary scores.

Language Exposure Measure	Correlation with French vocabulary (N=57)			
	Receptive		Productive	
Total percentage of French exposure	−0.04	($p > .10$)	0.01	($p > .10$)
Number of French-only blocks	0.00	($p > .10$)	0.03	($p > .10$)
Percentage of French in mother's input	−0.18	($p > .10$)	**0.27**	**($p = .04$)**
Number of French speakers (relatives & caregivers)	−0.10	($p > .10$)	0.11	($p > .10$)
Average within-block language purity	0.11	($p > .10$)	−0.06	($p > .10$)
Average within-speaker language purity	**0.36**	**($p = .007$)**	0.12	($p > .10$)

Table 3.4: Correlations between measures of L2 direct exposure and L2 vocabulary scores.

Language Exposure Measure	Correlation with L2 vocabulary (N=24)			
	Receptive		Productive	
Total percentage of L2 exposure	0.21	($p > .10$)	0.30	($p > .10$)
Number of L2-only blocks	**0.40**	**($p = 0.05$)**	0.26	($p > .10$)
Percentage of L2 in mother's input	**0.57**	**($p = .004$)**	0.06	($p > .10$)
Number of L2 speakers (relatives & caregivers)	0.20	($p > .10$)	−0.04	($p > .10$)
Average within-block language purity	0.07	($p > .10$)	0.07	($p > .10$)
Average within-speaker language purity	**0.35**	**($p = .09$)**	0.09	($p > .10$)

The correlation analyses showed overall few significant results, especially on French vocabulary scores. Most surprisingly, the total percentage of direct input in a given language did not seem to affect the vocabulary size in that language. However, we found a positive correlation between vocabulary size and the percentage of each language spoken by the mother: the percentage of French maternal input correlated with French production scores, and the percentage of L2 maternal input correlated with L2 comprehension scores. Furthermore, the number of blocks where only L2 was used revealed a moderate, marginally significant, correlation with L2 receptive vocabulary.

Interestingly, we found a positive effect of within-speaker language purity, but not of within-block language purity, affecting significantly French comprehension scores, and marginally L2 production scores. This effect suggests that the frequency with which speakers use both languages in the same time block may affect infants' vocabulary, with higher language overlap resulting in lower vocabulary scores. This observation is in line with the results presented by Byers-Heinlein (2013) showing that

parental language mixing had a negative effect on comprehension scores of 18-month-old infants.

Overall, the correlation results suggest that properties of the bilingual exposure may affect vocabulary development in 11-month-olds, in particular for the minority language. As this is an exploratory analysis, these effects remain to be confirmed in a new study. A larger sample size (especially regarding the L2 vocabulary scores, for which we had only half of the participants) will allow for further investigation of the relative contribution of each effect to bilinguals' vocabulary in each language.

3.4 Discussion

In this study we examined properties of the input and the environment that characterize bilingual exposure in 11-month-old infants, and their possible effects on vocabulary size. In order to capture these properties, we used a Language Diary method (De Houwer & Bornstein, 2003). Caregivers kept a record of their children's language input every half hour throughout two days. In our sample of 59 infants with a regular exposure to French and an additional language, we found that a great majority of the participating families adhered (either strictly or flexibly) to the one-parent-one-language approach. Despite this, we observed great variability in the proportions of each language in their input, in how often children encountered both languages within the same half-hour, in how frequently speakers in their environment used both languages within the same time block, and in the number of speakers who provided input in each language. In contrast with parallel studies in a homogeneous bilingual population (Spanish-English bilinguals studied by Place & Hoff, 2011, 2016), our bilinguals growing up in a French-dominant community encountered significantly more speakers of French than of their additional language. The minority language was provided almost exclusively by members of their close family, and in a few cases, additionally or exclusively by a nanny.

Using these measures, we examined potential effects of the dual input on infants' French and L2 vocabulary scores. The correlation results revealed some effects of language exposure on the minority language, namely the number of L2-only time blocks, the proportion of the maternal input in L2, and the frequency of within-speaker language purity (trend). In the case of French vocabulary, correlations were only found for the proportion of the maternal input in French and for within-speaker language purity. Evidence for an impact of language separation by speakers, as reflected by the within-speaker language purity correlations, is in line with previous results using a parental language mixing questionnaire by Byers-Heinlein (2013). As discussed in the introduction, contradictory evidence of the

impact of by-speaker language separation on language development may reflect a change in the relative importance of this variable throughout the child's life. Our results thus provide further evidence for this hypothesis.

The absence of other expected effects often found in the literature (such as the overall proportion of the input in each language, or the number of speakers), especially on French vocabulary, may have several explanations. Firstly, some measures of input defined in this study may not be the most relevant factors in language development for this specific population. Similar measures of language exposure have been found to affect vocabulary size in Spanish-English bilingual 2-year-olds (Place & Hoff, 2011, 2016). In particular, the number of speakers of a given language, and the total number of blocks where only one language was used, were found to positively affect productive vocabulary size in each language. However, those studies were conducted in a region with a large number of Spanish, English, and Spanish-English bilingual speakers (South Florida, USA), meaning that children growing up in that area were more likely to be exposed to a large range of speakers of each language and of one-language-only situations for both languages. Thus, this factor may not reveal differences in the linguistic outcomes of French-L2 bilingual infants living in Paris.

Secondly, for our population, French vocabulary may be relatively easier to acquire than L2 vocabulary – regardless of the amount of input in French from regular speakers – given that it is the majority language spoken in the region. Indeed, previous studies with pre-school and school-aged children have found that the community's dominant language has an advantage over the minority language, with higher success rates in acquisition (e.g., De Houwer, 2007; Gathercole & Thomas, 2009; Yamamoto, 2001). In contrast with French, our bilinguals' L2 development depends solely on the input coming from the few L2 speakers in their environment, which are coincidentally the main caregivers, making the measures derived from the Language Diaries more relevant for L2 than for French vocabulary acquisition.

Thirdly, the measures of input (i.e., the diaries) may have been too noisy to observe certain effects. While we found our diary estimates of language exposure to be in good agreement with parental reports, it is possible that more fine-grained differences between infants' backgrounds are not observable with only two diaries. Furthermore, the real contribution of each speaker to the total input may have been misrepresented, as diary reports were separated by speaker and then weighted equally across speakers in a given time block, thus possibly misjudging the amount of speech for some speakers. As suggested by De Houwer (2018), complementing diary informations with audio recordings may give a

better estimation of the contribution of each speaker to the child's input.

Fourthly, the outcome measure, which was limited to a list of less than 100 words (short CDI), may be inappropriate to measure bilingual infants' vocabulary. Although mean vocabulary size was far from the upper limit (and no infant was reported to know all words in the questionnaire), it is possible that the range of words was not sufficient to capture small differences between infants. As a bilingual's vocabulary could be tied to different contexts for each language, using the full CDI may be necessary to tackle these differences. However, parents might be unable to reliably estimate the receptive vocabulary of pre-verbal infants, a problem which would not be solved (and might even be amplified) by using a larger CDI. In previous studies (Place & Hoff, 2011, 2016), only productive vocabulary was used as outcome measure, which is arguably easier to estimate than receptive vocabulary, especially when children are 2-years-old. At 11 months, whether a child knows a word or not may be up to subjective impressions.

Last but not least, it should be noted that our population was heterogeneous, with infants being exposed to a large range of languages with varying cross-linguistic similarities with French, while in Place & Hoff (2011, 2016) all infants were exposed to the same language pair. Language distance has been suggested to modulate language acquisition, affecting, for instance, phonological development (Bosch & Sebastián-Gallés, 2003b; Sundara & Scutellaro, 2011), acquisition of translation equivalents (Bosch & Ramon-Casas, 2014) and grammatical structures (Döpke, 2000; Hulk & Müller, 2000; Müller & Hulk, 2001). However, it is rarely taken into account in studies of lexical acquisition with heterogeneous populations. When investigating properties of the bilingual exposure and their impact on vocabulary development, an oversight of this factor may end up obscuring other effects, especially with small sample sizes. Cross-linguistic research of the "language pair effect" in this kind of study is only recently emerging (Floccia et al., 2018; O'Toole et al., 2017). In a very recent study with nearly 400 toddlers learning English and an additional language in the UK, Floccia et al. (2018) investigated the effect of language distance on vocabulary development. Based on measures of phonological overlap, morphological complexity and word order typology, they showed, for the first time, a cross-linguistic effect of language distance on vocabulary size: close language pairs resulted in higher vocabulary sizes in the additional language.

While an investigation of the effect of language distance was not part of the original plan of this study, in light of these recent results we did a post-hoc examination of differences in vocabulary size across the language pairs included in our sample. Given the small sample sizes for each language, we will

not perform statistical analyses, but we will provide a qualitative description. In Figure 3.13 we show the French comprehension scores obtained for each of the L2 languages in our population, separated by language family. While for most language pairs we only had one or two participants, those for which we had several infants show large cross-linguistic differences. Particularly, infants exposed to French plus a Germanic language (English, German, Swedish) seemed to have overall lower vocabulary scores (and less variance) in French than infants in the Romance group. Out of all language pairs, French-Spanish had the highest vocabulary scores. Although we cannot draw conclusions from these observations, these general differences are in line with previous studies suggesting that the specific language pair may play a role in bilingual language development. Particularly, language pairs from the same family (here, the Romance languages), which are likely to share a large number of cognates and structural similarities, resulted in higher French vocabulary scores. It should be kept in mind, however, that these results may be due not only to cross-linguistic differences (or similarities), but also to cultural differences. In conclusion, in studies comparing bilingual infants with a variety of language pairs, it might be necessary to include a larger sample that allows one to adequately measure language pair effects.

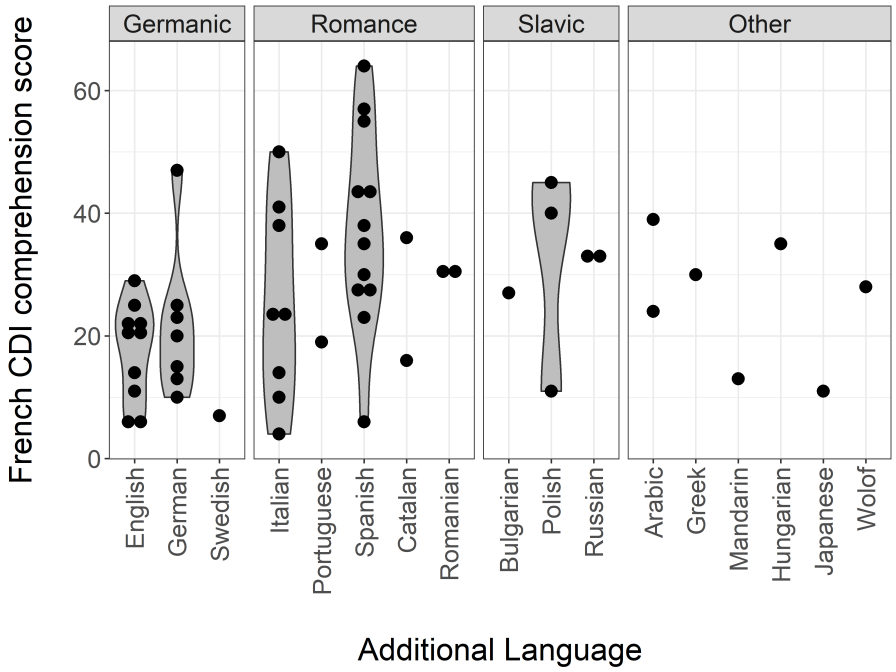
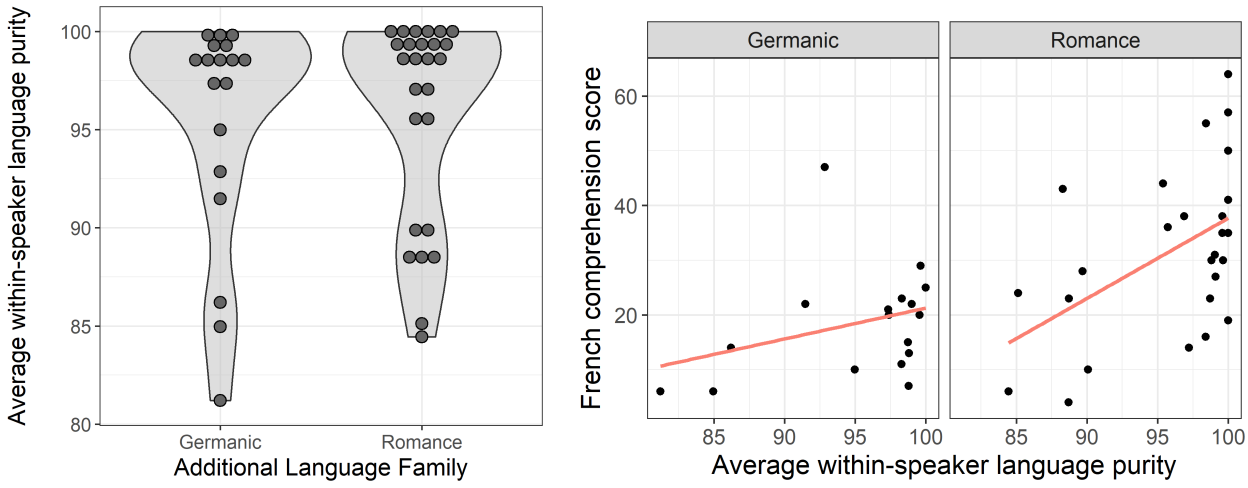


Figure 3.13: French vocabulary size separated by language and language family.

Given these large differences in vocabulary between language families, we wondered if the strong effect that we had observed of within-speaker language separation would still hold within each language group. We thus compared the two groups for which we had a sufficient number of infants, that is, the

French + Romance group, and the French + Germanic group. First, we checked that both groups had similar distributions of WSLP values. This is shown in Figure 3.14a. Then, we recalculated the correlation of WSLP with their French vocabulary score, independently for each language group. As can be seen in Figure 3.14b, while both groups have a positive tendency, only the French + Romance group shows a significant correlation (Romance: $r = 0.50$, $p = 0.01$; Germanic: $r = 0.33$, $p = 0.19$). Again, as these are post-hoc analyses with small sample sizes, this finding calls for a replication study. However, if confirmed, this might mean that by-speaker language separation is only relevant to infants learning close language pairs, as it is precisely in those cases that language discrimination has been claimed to be difficult (as we have discussed in Chapter 1). Further research should thus address the interaction of these two factors, that is, language distance and environmental language separation.



(a) (b)

Figure 3.14: Language group differences in (a) distribution of within-speaker language purity (WSLP), and (b) correlation of WSLP with French vocabulary score.

A final remark should be made regarding the methodology used in this study. The Language Diary technique has several advantages and disadvantages over other methods of examining language input. On the one hand, in comparison with parental questionnaires that estimate the global percentage of exposure to each language, the Language Diaries offer more local estimates, which do not reflect their whole experience since birth. On the other hand, as parents report language use as it takes place, the diaries are less affected by parents' memory and biases, therefore giving a more accurate and detailed picture of their children's language exposure. Furthermore, as discussed earlier, bilinguals' language skills may adapt to changes in the amount of each language's input (David & Wei, 2008; De Houwer, 2009). The diaries thus offer the possibility to analyse the role of current exposure on language outcomes. While in this study we have only asked parents to fill in the diaries over 2 days,

this method can be used over longer periods of time, as has been done before by Place & Hoff, 2011, 2016 (one diary per week for 7 weeks) and by De Houwer, 2011 (one diary per week over 15 months).

An alternative technique to examine language exposure is the use of naturalistic audio or video recordings (e.g., De Houwer, 2014; De Houwer and Bornstein, 2016; Marchman et al., 2017; Ramírez-Esparza, García-Sierra, and Kuhl, 2017). While recordings can provide an even more accurate measure of language input (it allows, for instance, to compute the number of words per hour in each language), their processing is highly resource- and time-consuming. Existing semi-automatic systems, such as LENA^{TM}, have been mainly designed for monolingual speech, and may not cope well with multilingual speakers. Current advances in speech technologies (such as the i-vectors presented in Chapter 2) may provide, in the near future, a better automatised solution to this problem. Until then, the Language Diaries offer the best cost-benefit ratio.

3.A Appendix A: Language Diary sample page

Time range	Person or people who were with the infant	Languages spoken to the infant	Languages spoken to other people in the room
7:00 to 7:30	Person #1: Who?	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____
	Person #2: Who?	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____
	Person #3: Who?	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____
	Person #4: Who?	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____
	Other people: How many? Who?	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ None ☐ Other: _____
Place and activity:			
Comments:			

3.B Appendix B: Language Environment Questionnaire

Babylab
Laboratoire de Sciences Cognitives et Psycholinguistique

Subject's ID:
(to be filled in by the researcher)

Language Environment Questionnaire

Date:

The purpose of this questionnaire is to study the variability in speakers and contexts in a bilingual setting. All your answers will be completely anonymized in all further processing of the data, and any personal information you provide will remain strictly confidential and kept separate from this questionnaire. The following questions concern the language environment of your child, from their birth to this day. Please read each question carefully before providing an answer. Remember that each child has a unique combination of languages, so there are no right or wrong answers.

1. Where was the child's mother born? (Please indicate country and region)

   _____

2. What is her native language? (If more than one, please specify)

   _____

3. Where was the child's father born? (Please indicate country and region)

   _____

4. What is his native language? (If more than one, please specify)

   _____

5. Have the child and family moved from city or country since he/she was born? Yes/No: _____

   a) If so, where did the child use to live, and when did the family move?

      _____

6. Who lives with the child? (Please specify all adults and children that are **currently** living in the house, for example: *mother, father, grandmother, two brothers and one sister.* Their names are not required.)

   _____

   a) Has this changed since the infant was born? If so, when, and who lived with the child before?

      _____

7. Which languages are currently spoken **at home**? (Specify all languages)

   _____

   a) Has this changed since the infant was born? If so, when, and which languages were spoken before?

   _____

8. Which languages does the child hear **out of home**? (Specify all languages)

   _____

   a) Has this changed since the infant was born? If so, which languages did he/she often hear before?

   _____

9. Who is currently the main caregiver of the child? (For example: you, your partner, a relative, a nanny, etc.)

   _____

   a) In what language(s) does he/she usually speak to the child?

      _____

b) If not the mother or the father, what is the caregiver's native language?

_____

10. Does the child attend a nursery? If so, since when, and how many hours a week?

_____

a) What language is usually spoken at the nursery? (If more than one, please specify)

_____

11. Does the child have a nanny? If so, since when, and how many hours a week?

_____

a) What is the nanny's native language? (If more than one, please specify)

_____

b) In what language(s) does he/she usually speak to the child?

_____

12. Does the child have older siblings? If so, how many and how old are they?

_____

a) In what language(s) do the siblings usually speak to the child?

_____

b) In what language(s) do the main caregivers speak to the child's siblings?

_____

13. Does anyone else help take care of the child? (For example: *grandparents*, *family friends*, etc.)
If so, who? (No names are required).

_____

a) What is (or are) their native language(s)? (Specify for each person)

_____

b) In what language(s) do they usually speak to the child? (Specify for each person)

_____

14. Please estimate the percentage of each language that the child heard **at home in the past year**
(For example, French 40% - English 60%)

French: _____ % English: _____ % Other: _____ % (Specify language: _____)

15. Please estimate the percentage of each language that the child heard **out of home in the past year**
(For example, French 80% - English 20%)

French: _____ % English: _____ % Other: _____ % (Specify language: _____)

16. Do you adopt a particular strategy to teach each language to your child? Yes/No: _____ If so, which?

☐ One parent – one language ☐ Other: _____

_____

17. Does your child watch cartoons on TV/tablet/computer? In which language? (Tick **all** options that apply):

☐ French ☐ English ☐ Other language: _____ How many hours a day? _____

If you would like to clarify an answer or add more information, please write your comments in the following box, specifying which question it concerns.

Comments (optional)

18. The following questions concern the adults that often spend time with your baby. Please identify the **4 adults that most interact with your child during a typical week**. For each of them, give their **initials** and their **relationship with the child** (for example, *B. mother*, *A.C. family friend*, etc.) and fill in the following information. Their initials will help us identify how many different speakers there are, especially if there are two people with the same relationship to the child (for example, two grandmothers). These initials will be removed and replaced by numbers (for example: *grandmother 1* and *grandmother 2*) in all further processing of the data.

Adult (initials and relationship with the child)	Level of proficiency in French	Level of proficiency in English	Language(s) used when speaking to the child	Language(s) used to speak with other adults in presence of the child	Amount of hours a day spent with the child
Person #1:	☐ Doesn't speak the language ☐ Basic skills ☐ Intermediate ☐ Advanced ☐ Native-like ☐ Native	☐ Doesn't speak the language ☐ Basic skills ☐ Intermediate ☐ Advanced ☐ Native-like ☐ Native	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ Other (please specify) _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ Other (please specify) _____	During the week: On weekends:
Person #2:	☐ Doesn't speak the language ☐ Basic skills ☐ Intermediate ☐ Advanced ☐ Native-like ☐ Native	☐ Doesn't speak the language ☐ Basic skills ☐ Intermediate ☐ Advanced ☐ Native-like ☐ Native	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ Other (please specify) _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ Other (please specify) _____	During the week: On weekends:
Person #3:	☐ Doesn't speak the language ☐ Basic skills ☐ Intermediate ☐ Advanced ☐ Native-like ☐ Native	☐ Doesn't speak the language ☐ Basic skills ☐ Intermediate ☐ Advanced ☐ Native-like ☐ Native	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ Other (please specify) _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ Other (please specify) _____	During the week: On weekends:
Person #4:	☐ Doesn't speak the language ☐ Basic skills ☐ Intermediate ☐ Advanced ☐ Native-like ☐ Native	☐ Doesn't speak the language ☐ Basic skills ☐ Intermediate ☐ Advanced ☐ Native-like ☐ Native	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ Other (please specify) _____	☐ Only French ☐ Mostly French ☐ Both equally ☐ Mostly English ☐ Only English ☐ Other (please specify) _____	During the week: On weekends:

If you would like to clarify an answer or add more information, please write your comments in the following box, specifying which of the 4 adults it concerns.

Comments (optional)

3.C Appendix C: Comparison of weighting options

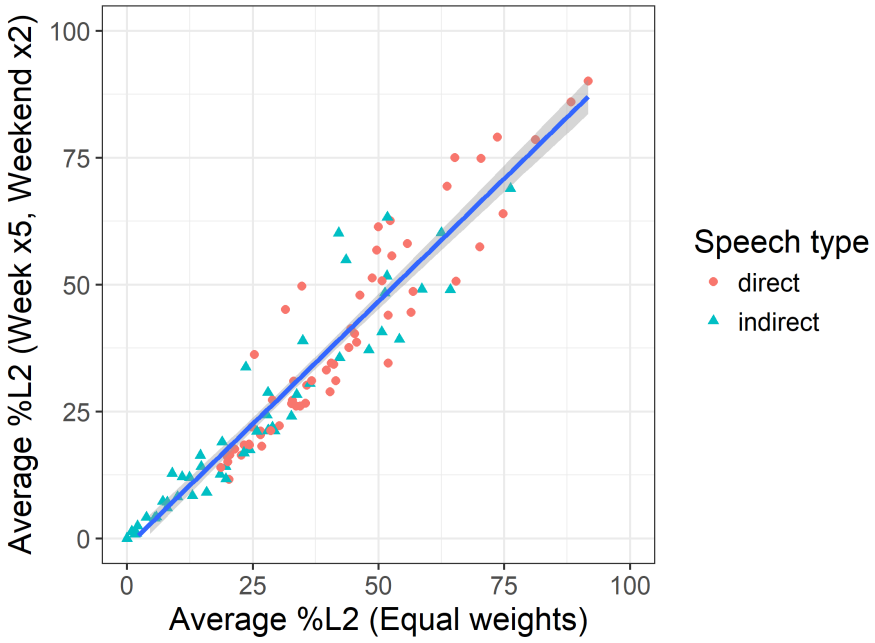


Figure 3.15: Correlation of average percentage exposure estimates for L2 using equal vs. unequal week/weekend weights ($r = 0.95$, $p < .0001$).

3.D Appendix D: Correlation results for indirect input measures

Table 3.5: Correlations between measures of French indirect exposure and French vocabulary scores.

Language Exposure Measure	Correlation with French vocabulary			
	Receptive		Productive	
Total percentage of French exposure	0.02	($p > .10$)	0.04	($p > .10$)
Number of French-only blocks	0.12	($p > .10$)	−0.02	($p > .10$)
Percentage of French in mother's input	−0.02	($p > .10$)	0.10	($p > .10$)
Number of French speakers (relatives & caregivers)	−0.11	($p > .10$)	0.16	($p > .10$)
Average within-block language purity	0.16	($p > .10$)	−0.07	($p > .10$)
Average within-speaker language purity	0.18	($p > .10$)	−0.22	($p > .10$)

Table 3.6: Correlations between measures of L2 indirect exposure and L2 vocabulary scores.

Language Exposure Measure	Correlation with L2 vocabulary			
	Receptive		Productive	
Total percentage of L2 exposure	0.13	($p > .10$)	−0.11	($p > .10$)
Number of L2-only blocks	0.23	($p > .10$)	−0.27	($p > .10$)
Percentage of L2 in mother's input	0.26	($p > .10$)	0.16	($p > .10$)
Number of L2 speakers (relatives & caregivers)	0.09	($p > .10$)	−0.04	($p > .10$)
Average within-block language purity	0.25	($p > .10$)	-0.14	($p > .10$)
Average within-speaker language purity	**0.38**	**($p = .08$)**	0.16	($p > .10$)

Chapter 4

Perception of language-specific phonological rules

In this chapter we will explore the perception of language-specific phonological rules in French-English bilingual children. The first half of this chapter was submitted for publication: Carbajal, M.J., Chartofylaka, L., Hamilton, M., Fiévet, A.C. and Peperkamp, S. (in revision) *Compensation for phonological assimilation in bilingual children.*

In the second half of this chapter, we describe a series of pilot experiments that were conducted in order to find an appropriate paradigm for testing our within-subject design.

4.1 Article

Compensation for Phonological Assimilation in Bilingual Children

Abstract

We investigate bilingual children's perception of assimilations, i.e. phonological rules by which a consonant at a word edge adopts a phonological feature of a neighboring consonant. For instance, English has place assimilation (e.g., *green* is pronounced with a final [m] in *green pen*), while French has voicing assimilation (e.g., *sac* is pronounced with a final [g] in *sac vert* 'green bag'). Previous research has shown that French and English monolingual toddlers compensate for the assimilation rule of their language, correctly recovering the intended words, but not for a rule that does not exist in their language. Using a word recognition videogame with French sentences, we show that French-English bilingual 6-year-olds' perform exactly like French monolinguals of the same age: they compensate for voicing but not for place assimilation. Thus, despite their dual language input they have acquired French voicing assimilation and show no interference from English place assimilation.

Introduction

Bilingual children need to acquire two distinct phonological systems, each composed of a variety of aspects, such as sound categories, syllable structure, and phonological rules. As in monolinguals, phonological acquisition occurs during the first years of life, but not necessarily at the same speed. For instance, depending on the language pair and the overlap between their sound categories, the acquisition of these categories and the ability to use them for word learning can take longer to develop when infants are exposed to a dual language input (Bosch & Sebastián-Gallés, 2003a, 2003b; Fennell, Byers-Heinlein & Werker, 2007; Havy, Bouchon & Nazzi, 2015; Liu & Kager, 2015; Ramon-Casas, Swingley, Sebastián-Gallés & Bosch, 2009; Sundara, Polka & Genesee, 2006; Sundara & Scutellaro, 2011).

Languages differ not only in their composing sounds, but also in the way these sounds are produced in different phonological contexts, and how they organize to form words and sentences. Children exposed to a dual speech input are thus faced with the task of disentangling and learning two such sets of language-specific phonological grammars. This challenging task may produce different patterns of acquisition compared to what has been documented for monolingual children. Research on the development of phonological structures in young bilinguals beyond the acquisition of sound categories, however, remains scarce, and has so far focused on children's productions (Fabiano-Smith, Oglivie, Maievski & Schertz, 2015; MacLeod & Fabiano-Smith, 2015; Nicoladis & Paradis, 2011; Paradis, 2001). For instance, in a study using a nonword repetition task, Paradis (2001) investigated French-English bilingual toddlers' truncations in nonce words. Truncations (e.g., *nana* for *banana*) are typical of toddlers' productions and are known to be influenced by language-specific word-prosodic properties (Allen & Hawkins, 1980). While the truncation patterns of the bilingual toddlers in each of their languages were found to generally match those of the corresponding monolingual French and English toddlers, there was also some evidence of cross-linguistic transfer. Thus, their phonological systems appeared to be differentiated, yet not fully independent. In another study using a picture naming task, Nicoladis and Paradis (2011) explored French-English bilingual children's production of liaison, a complex phonological rule in French that causes word-final silent consonants to be pronounced before a vowel-initial word (e.g., *petit* [pəti] 'small' is pronounced [pətit] in *petit ours* 'small bear'). They found that, in general, 3- to 5-year-old children's French vocabulary –but not their age– correlated positively with their production of liaison. Interestingly, when matched by vocabulary, bilinguals applied liaison less often than their monolingual peers', but only in low-frequency collocation frames, suggesting that their acquisition of liaison lagged behind that of monolinguals. However, the sample size of this matched comparison was very small (6 monolinguals and 6 bilinguals), making it difficult to draw conclusions.

Here, we investigate the perception of a phonological process that alters the surface form of words in specific contexts. *Assimilations* are phonological rules that cause certain sounds to adopt specific features from adjacent sounds. While common across the world's languages, the sounds that undergo a change, the specific features that change, and the contexts that trigger them vary from one language to the other. For example, in English, the word-final alveolar consonants /t/, /d/, and /n/ can adopt the place of articulation of a following labial (/b/, /m/, /p/) or velar (/g/, /k/) consonant. Thus, a word like *green* [gɹin] may be produced as *greem* [gɹim] if followed by a word beginning with a labial sound, such as the consonant /b/ in the phrase *green ball* [gɹimbɔl]. While *place* assimilation rules are present in many languages, assimilation may also involve other features. This is the case in French, where place assimilation does not exist; instead, it is the *voicing* feature (i.e., the vibration of the vocal folds) of certain consonants that may spread to neighboring sounds. More specifically, voiceless obstruents (/p/, /t/, /k/, /f/, /s/, /ʃ/) may adopt the voicing value of a following voiced obstruent (/b/, /d/, /g/, /v/, /z/, /ʒ/), and vice versa: voiced obstruents may turn into their voiceless counterparts if followed by an unvoiced obstruent. For instance, the French word *robe* [ʁɔb] 'dress' may be pronounced as [ʁɔp] if followed by a word beginning with a voiceless obstruent, as in the phrase *robe sale* [ʁɔpsal] 'dirty dress'. The application of these rules is context-specific, such that, for instance, the word *robe* does not undergo assimilation if followed by a consonant other than a voiceless obstruent, as in the phrase *robe noire* [ʁɔbnwaʁ] 'black dress'.

There is ample evidence that listeners have detailed knowledge of language-specific assimilation, and compensate for its effect in order to retrieve the intended word (Coenen, Zwitserlood & Bölte, 2001; Darcy, Ramus, Christophe, Kinzler & Dupoux, 2009; Gaskell & Marslen-Wilson, 1996; Gaskell & Snoeren, 2008; Mitterer & Blomert, 2003). For instance, Darcy et al. (2009) showed that French listeners tend to recognize the word *robe* [ʁɔb] when it is pronounced with a final [p] in *robe sale*, where voicing assimilation is viable, but not in *robe noire*, where it is unviable. Thus,

they compensate for their native rule of voicing assimilation in a context-sensitive manner. By contrast, they hardly compensate for a hypothetical rule of place assimilation; that is, they generally fail to recognize the word *lune* [lyn] 'moon' when it is pronounced with final [m] in *lune pale* 'pale moon', even though it presents a viable context for place assimilation.[1] English listeners who were tested on English sentences showed the reverse pattern; that is, they compensated more for their native rule of place assimilation than for a hypothetical rule of voicing assimilation.

This language-specific compensation for assimilation has also been found in toddlers (Skoruppa, Mani & Peperkamp, 2013; Skoruppa, Mani, Plunkett, Cabrol & Peperkamp, 2013). For instance, Skoruppa, Mani and Peperkamp (2013) used a picture pointing task with French and English 2½- to 3-year-old toddlers. French toddlers were tested on both their native voicing assimilation and the non-native place assimilation rule. In each trial, they were first presented with two pictures, one of a familiar and one of an unknown object. Each picture was presented with a short labeling sentence, where the label for the unknown object differed from that of the familiar one only in either the place or the voicing feature of the final consonant. For instance, the familiar object chair - in French: *chaise* [ʃɛz] - would be paired with an unknown object called [ʃɛs]. Both pictures then reappeared on screen side by side, accompanied by a phrase containing the novel word embedded in one of two possible phonological contexts: either followed by a consonant that allows the corresponding assimilation (*viable* condition), such as the voiceless obstruent /p/ in *Montre la [ʃɛs] par ici !* 'Show the [ʃɛs] over here!', or followed by a context that does not produce the respective assimilation (*unviable* condition), such as the liquid consonant /l/ in *Montre la [ʃɛs] là-bas !* 'Show the [ʃɛs] over there!'. In voicing assimilation trials, French toddlers pointed at the familiar object more often upon hearing the altered word form in a viable than in an unviable context for assimilation. Furthermore, when tested on non-native place assimilation, they failed to recognize the familiar

1. They recognize it even less in an unviable context for place assimilation, such as *lune rousse* 'red moon', suggesting a small language-independent compensation effect.

words independently of the context. Additionally, English toddlers were tested, but only on their native place assimilation rule, for which they showed context-sensitive compensation. In a second study, however, Skoruppa, Mani, Plunkett, et al. (2013) showed that like French toddlers, English toddlers fail to compensate for a non-native assimilation rule. Here, 24-month-olds were tested in an intermodal preferential looking paradigm. The design of the experiments was similar to the one for the picture pointing task. In particular, pictures of familiar objects were paired with pictures of unfamiliar objects whose label differed from that of the familiar one only in the last consonant. Following the presentation of the two pictures and their labels, the pictures were shown side by side and toddlers were asked to look at one of them. Both French and English toddlers were tested on voicing assimilation. French toddlers increased their looks to the familiar object in the post-naming phase when they heard an assimilated form in a viable but not in an unviable context. English toddlers, by contrast, showed no such increase, regardless of the context in which the assimilated form occurred.

Taken together, these studies thus show that like adults, French and English toddlers show language-specific knowledge of assimilation: they compensate for their native but not for a non-native assimilation rule. This is quite remarkable, since the frequency with which assimilation applies in spontaneous speech tends to be low. For instance, Dilley & Pitt (2007) found that in an English corpus of spontaneous speech, 3.2% of words ended in a consonant that can undergo place assimilation given its following context, of which 9% were effectively assimilated. Note, though, that a higher assimilation rate, 22%, has been reported for English infant-directed speech (Buckler, Goy & Johnson, 2018). For French, no systematic analysis of voicing assimilation rates in spontaneous speech has yet been conducted. However, in a corpus of journalistic speech, Adda-Decker & Hallé (2007) found that 1.8% of all word boundaries contained a viable context for voicing assimilation, and assimilation rates were slightly above 20%.

The aim of the present study is to examine how the presence of a second language during early childhood affects acquisition. Would bilingual children acquire their language-specific rules and hence behave like monolinguals in both of their languages? To start investigating this question, we examine how French-English bilingual children who have heard both languages regularly from their first year of life, perceive voicing and place assimilation when listening to French. While place assimilation is not a native rule in French, French-English bilinguals may be familiar with this rule from their English input. Since bilingual children have a reduced exposure to each language compared to monolinguals of the same age, we test 6-year-old children, who should have had experience with both rules and have learnt a sufficient number of assimilable words to be used in the experiment.

In the studies with monolingual toddlers mentioned above (Skoruppa, Mani & Peperkamp, 2013; Skoruppa, Mani, Plunkett, et al., 2013), native and non-native rules were tested in separate groups of participants. Here, we want to test all children on both voicing and place assimilation in a single experiment, allowing us to directly compare their processing of these rules. We therefore implemented a child-friendly touchpad videogame that allows us to gather sufficient data on both rules. In this game, children are presented with pictures of familiar objects. For each item they hear a phrase containing the object's name, always pronounced with either a place or a voicing change in its final consonant in either a viable or an unviable context for the corresponding assimilation rule. If children recognize the altered word form as a good pronunciation of the familiar object, they are to click on its picture, while if they reject it as a valid pronunciation of the target word, they are instructed to click on a red cross presented on its side. How often children choose the picture of the familiar object – that is, how often they accept the altered word form as a good pronunciation – in the viable and unviable contexts for each assimilation rule is thus informative of their ability to compensate for assimilation.

In Experiment 1, we test French monolingual 6-year-olds. Based on previous results with monolingual adults (Darcy et al., 2009) and toddlers (Skoruppa, Mani &

Peperkamp, 2013; Skoruppa, Mani, Plunkett, et al., 2013), we expect them to show compensation for voicing (native) but little or no compensation for place (non-native) assimilation. Crucially, we also expect to observe a significant difference in their response patterns for these two rules, thus providing clear evidence that native and non-native assimilation are treated differently. In the between-participant designs for toddlers of Skoruppa, Mani and Peperkamp (2013) and Skoruppa, Mani, Plunkett, et al. (2013), this difference was not tested, but it was observed in adults with the within-participants design of Darcy et al. (2009).

In Experiment 2, we test French-English bilingual children of the same age, for whom there are several plausible outcomes. On the one hand, 6-year-old bilinguals may have already acquired French voicing assimilation, with no interference from English. If this is the case, then we should observe the same response pattern as that in their monolingual peers in Experiment 1, i.e. compensation for voicing but not or only a little for place assimilation. On the other hand, it is possible that bilinguals show signs of delay and/or cross-linguistic influence in their compensation patterns. For instance, due to their familiarity with place assimilation in their English input, they may show context-specific compensation for both voicing and place assimilation in French. Or, given that their input is more variable than that of monolinguals, they may be more flexible regarding mispronunciations and thus accept all word alterations as valid, regardless of context; in that case, they should show high acceptance rates for both voicing- and place-assimilated forms in both the viable and the unviable contexts. Alternatively, due to their more variable input they may show a lack of compensation by rejecting all assimilated forms, even those presented in valid contexts.

Experiment 1: Monolinguals

Methods

Participants

Twenty-one French monolingual 6-year-olds (13 girls, 8 boys, mean age: 70.01 months, age range: 64.73 – 75.33 months) participated. An additional two children were tested but not included in the analysis due to failure to pass the training criteria (see *Exclusion criteria* section below). Written consent was obtained from the parents of all participating children prior to testing.

Materials

A set of twenty-four monosyllabic French nouns and matching color pictures were selected as test items. Some of the pictures were taken from Rossion and Pourtois' (2014) color version of the Snodgrass pictures set; the others were drawn by the first author. All items were judged to be generally known by French children based on children's picture books and vocabulary questionnaires collected during a pilot study. Three of the nouns had an English cognate that differed, however, in at least one phoneme.

Half of the nouns were selected to test voicing assimilation, the other half to test place assimilation (see Appendix A). The nouns for voicing assimilation ended in either a voiced or a voiceless obstruent, e.g., *robe* [ʁɔb] "dress", *tasse* [tas] "cup". From each of these nouns, its assimilated form was constructed by changing the voicing value of the final consonant, thus transforming voiced obstruents into their voiceless counterparts, and vice-versa (e.g., [ʁɔb] → [ʁɔp], [tas] → [taz]). The nouns for place assimilation ended in one of the alveolar consonants [t, d, n], e.g., *botte* [bɔt] "boot", *viande* [vjɑ̃d] "meat", *lune* [lyn] "moon". From each of these nouns, an assimilated form was constructed by changing the place of articulation of the final consonant to bilabial (e.g. [bɔt] → [bɔp], [vjɑ̃d] → [vjɑ̃b], [lyn] → [lym]). All

assimilated forms were non-words or infrequent real words not known to 6-year-old children.

Each of the 24 assimilated forms was embedded in two short sentences with a touching request. One of the sentences provided a *viable* context for assimilation, and one an *unviable* context. Examples are shown in Table 1). Thus, for voicing assimilation, the final obstruent of the assimilated form was followed by an obstruent with the same voicing value (viable context), or by any other consonant (i.e., an obstruent with the opposite voicing value, or a liquid or nasal consonant; unviable context). Similarly, for place assimilation, the final, labial, consonant of the assimilated form would be followed by a labial consonant (viable context), or by a consonant with another place of articulation (unviable context). Note that for this rule, the terms viable and unviable refer to the context's status according to the place assimilation rule in English.

Table 1

Sample Sentences with Viable and Unviable Contexts for Voicing and Place Assimilation.

Rule	Target word	Context	Example	Translation
Voicing	*tasse* [tas] "cup"	Viable	*Touche la [taz] devant toi !*	"Touch the # in front of you."
		Unviable	*Touche la [taz] maintenant !*	"Touch the # now."
Place	*lune* [lyn] "moon"	Viable	*Touche la [lym] par ici.*	"Touch the # over here."
		Unviable	*Touche la [lym] là-devant.*	"Touch the # there up front."

Nine additional color pictures denoting familiar nouns were selected for pre-training ($n = 3$) and training ($n = 6$). The items for pre-training were a ball, a heart, and a hen. The ball was only matched with its correct name (in French, *balle* [bal]), and the other two items were only matched with a non-word, differing from the object's name on either the entire rhyme (**kime* [kim] for *coeur* [kœʁ] "heart") or on its final consonant (**pouke* [puk] for *poule* [pul] "hen"). The items for training were -

like the test items - each matched with both its correct pronunciation and a non-word, differing only in either voicing or place of articulation of the final consonant (e.g., *glace* [glas] "ice cream" - **glaze* [glaz]). All pre-training and training items, as well as the matched non-words, were embedded in the final position of a short sentence of the form *Touche le/la # !* "Touch the #".

All sentences were recorded by a female native French speaker. She was instructed to read them in child-directed speech and without pauses. Minor editing and intensity normalization (70 dB) were done using the software *Praat* (Boersma & Weenink, 2015). To verify that the target consonants in the test sentences were always produced in their fully assimilated form, we asked 12 adult monolingual French speakers to listen to the final V(C)C segments of each assimilated word (e.g., [az] from *Touche la /taz/ maintenant*) in both viable and unviable conditions, and to categorize the final consonant. In each trial, they were given two options to choose from: either the assimilated form (in this example, [z]) or the unassimilated form (here, [s]). To avoid a bias for the assimilated form, control samples extracted from the unassimilated words produced in sentence-medial position were included as distractors (e.g., [as] from *Touche la tasse maintenant*). The order of presentation of the two rules (place, voicing) was alternating, while items and contexts (viable, unviable) were fully randomized. Consonants from voicing items were correctly identified in 97.9% of the cases in the viable condition and 95.8% in the unviable condition. A paired-samples t-test revealed no significant difference in voicing classification accuracy between conditions ($t(11) = 0.67$, $p = 0.52$). Consonants from place items were correctly identified in 98.6% of the cases in the viable condition and 92.4% in the unviable condition. No significant difference in place classification accuracy was found between conditions ($t(11) = -1.62$, $p = 0.13$). Finally, no difference was found in overall accuracy for place and voicing samples ($t(67.2) = -0.26$, $p = 0.79$).

Procedure

The experiment was implemented as a videogame app for tablets (a modified version of Cristia, 2016's app was used). An iPad Air 2 (model A1566) with a 9.7 inch screen and running on iOS 9 was used. The study took place either in a quiet room at a kindergarten school in Paris ($n = 13$), or in our babylab ($n = 8$). During the experiment, children sat down at a low table accompanied by the experimenter. Parents of children tested in the babylab were allowed to observe the experimental session through a monitor in a separated room, or by staying in the same room in silence and out of the child's sight.

Before the game began, the experimenter explained the task and motivated the child to play. Children were told that during the game, they would see an object on one side of the screen and a red cross on the opposite side (Figure 1), and that a cartoon girl would appear and ask them to touch the object. Children were then told that the girl sometimes made mistakes while saying the name of the object, and if she did, they should click on the red cross instead of the object. The exact sequence of events in each trial was as follows: first, the picture of the object and the red cross appeared each on one side of the screen. Following a 3 second pause, the girl appeared in the middle of the screen and waited until the child clicked on her. This second pause, controlled by the children, gave them sufficient time to look at the picture and recognize the object before hearing the phrase. After clicking, the touch response became blocked while the character produced the request phrase. The touch response then reactivated, allowing the child to give an answer by clicking either on the picture of the object or on the red cross. No time limit was given to produce an answer, however, if the child took more than 6 seconds to respond, the character would make a sound to remind the child to make a choice.

The experiment lasted approximately 10 minutes and was composed of three stages:

Pre-training. The game began with three trials of low difficulty that allowed the child to get familiarized with the game and the touchpad. In all three trials, the target word or non-word appeared sentence-finally (e.g., *Touche la balle !* "Touch the ball!"). The order of pre-training trials was identical for all children, always beginning with a ball paired with its correct pronunciation, and continuing with a heart and then a hen, for which only non-words were used (the first one differing on the entire rhyme, the second one only on the last consonant). During this first stage, the experimenter offered help when necessary, and incorrect trials were repeated until the child answered correctly. Each time a correct answer was given, a cheerful chime played and a progress bar located at the top of the screen increased in size. After completion of pre-training, a star appeared on screen and the child was congratulated before moving on to the next stage.

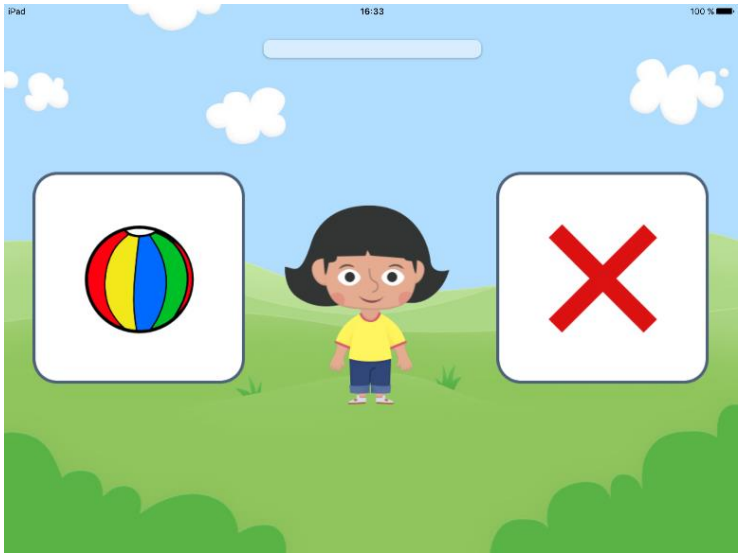


Figure 1. Screenshot of the game, showing the picture of a familiar object (here, a ball) on the left and a red cross on the right, with the cartoon character in the middle.

Training. Children were presented with 6 training trials, half of which contained the correct pronunciation of the target word, and the other half a mispronunciation formed by a change in either voicing or place of articulation of the final consonant. As in pre-training, words appeared sentence-finally and trials were repeated until the correct answer was given, but the experimenter remained silent

until the child made a choice. If the answer was correct, they heard a cheerful chime and the experimenter gave positive feedback, while if it was incorrect, she encouraged the child to try again without offering help. As a reward, a star appeared on screen after the third and sixth successful trial.

Test. Children were told that they had won the first part of the game and would play a second part without any help from the experimenter, who would stay in the room but turn her back and look away. To motivate them to keep playing, the experimenter explained that there were four stars to win, and that they would receive a sticker if they won them all. Unknowingly to them, trials were never repeated, and all children would get to see the four stars and win the sticker.

Children were presented with sentences containing the assimilated form of the target word in sentence-medial position, followed by either a viable or an unviable context for the corresponding assimilation rule. There were six trials of each of four experimental conditions (i.e., *voicing viable*, *voicing unviable*, *place viable*, *place unviable*), for a total of 24 trials. The context in which each test item appeared was counterbalanced across two lists, thus presenting each item only once to each child. Trials were divided into four blocks of six, containing equal numbers of voicing and place trials, as well as equal numbers of viable and unviable trials. Odd-numbered blocks contained two voicing viable trials and one place viable trial; in even-numbered blocks it was the opposite. All four test conditions were completely balanced every two blocks, and the side on which the picture of the target object appeared (left or right) was balanced within every block. No feedback was given during the test, except for the progress bar included in the game. At the end of each block, a star appeared.

Exclusion criteria

Although the game was expected to be easily learnt, a limit on the number of errors accepted in the training phase was imposed to make sure that all children included in the analysis had understood the task. Specifically, children were excluded from

analysis if they *a*) made mistakes in more than two items, or *b*) made more than two mistakes on the same item.

Results and discussion

Children's responses were automatically collected by the app and coded as a binary dependent variable representing whether the child clicked on the picture or the cross in each trial. Figure 2 shows the proportion of trials where the picture of the familiar object was chosen, split by condition. Responses were analyzed with a generalized linear mixed model (GLMM) with binomial family and logit link using package *lme4* (Bates, Maechler, Bolker & Walker, 2015) in the *R* environment (R Core Team, 2017). The model included fixed effects of assimilation rule (*voicing*, *place*) and context (*viable*, *unviable*), as well as the interaction between them. A maximal random effects structure was used (Barr, Levy, Scheepers & Tily, 2013), including an intercept as well as slopes for rule, context and their interaction by participant, and an intercept and slope for context by item. The variables *rule* and *context* were treatment-coded, with *rule* = *voicing* and *context* = *viable* as baseline levels.[2] P-values were obtained for all fixed effects using the *car* package (Fox & Weisberg, 2011).[3] A summary of the results is shown in Table 2.

2. In treatment coding, the estimate of the intercept corresponds to the mean of the baseline level, and the estimates for the independent variables correspond to *simple* effects (as opposed to *main* effects) of these variables with respect to the baseline (e.g., the estimate for context given *voicing viable* as baseline corresponds to the difference between voicing viable and voicing unviable trials).

3. In order to evaluate a potential effect of trial order, the model originally contained block as an additional fixed effect. This factor did not yield a significant effect (β = -0.20, SE = 0.22, n.s.), and was therefore excluded from the model reported in the main text. We also checked that the overall model fit was not better when this factor was included (likelihood-ratio test: $\chi^2(1) = 0.87$, $p = 0.35$).

Table 2

Summary of the generalized linear mixed model for monolinguals.

Fixed effect	*β*	*SE*	*z*	*p*
Intercept (*voicing, viable*)	1.82	0.97	1.86	0.06
Context (*unviable*)	-3.30	0.80	-4.11	<.0001
Rule (*place*)	-3.48	0.97	-3.59	0.0003
Interaction: Rule x Context	2.94	1.19	2.47	0.01

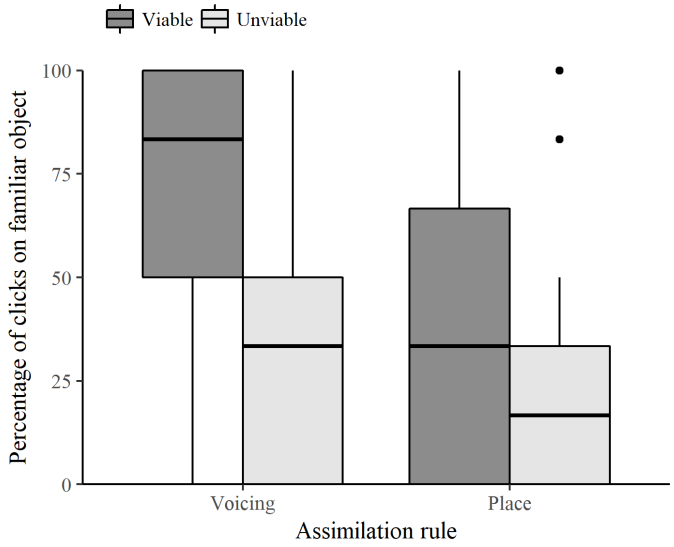


Figure 2. Boxplots representing the percentage of clicks on the picture of the familiar object per condition for monolingual children in Experiment 1. The bold horizontal line inside each box represents the median, while the bottom and top boundaries represent the first and third quartiles, respectively. Vertical lines extend to the highest and lowest values within 1.5 times the inter-quartile range above and below the boundaries. Individual dots represent outliers.

A likelihood ratio test of the resulting GLMM against the null model showed an overall good fit of the data ($\chi^2(3) = 16.3$, $p = 0.001$). As shown in Table 2, the estimated intercept shows a marginally significant preference for the familiar object over the cross in *voicing viable* trials. Crucially, the analysis revealed significant

effects of rule and context, as well as an interaction of rule by context. The negative estimate for rule indicates that the picture of the familiar object was more often selected when hearing *voicing viable* than *place viable* trials. The negative estimate for context indicates that children selected the picture of the familiar object more often when hearing voicing assimilation in a viable than in an unviable context. As to the interaction, it suggests that the effect of context is smaller for place than for voicing assimilation. In order to examine whether there is an effect of context in place assimilation trials, the rule variable was releveled with *place* as the baseline level. This releveling revealed no significant difference between viable and unviable contexts for place assimilation (β = -0.37, SE = 0.82, n.s.).

In line with previous findings with toddlers (Skoruppa, Mani & Peperkamp, 2013; Skoruppa, Mani, Plunkett, et al., 2013) these results show that French monolingual children compensate for voicing assimilation – a phonological rule that applies in their language – in a context-specific manner, but not for place assimilation, a rule that applies in English but not in French. Furthermore, this study successfully allowed for testing a native and a non-native rule within the same group of children, and showed for the first time a significant difference between compensation for the native rule and lack of compensation for the non-native rule, similarly to what has been reported for adults (Darcy et al., 2009). Most importantly, this provides a suitable method for testing bilingual children, as it allows us to simultaneously assess their interpretation of native and non-native rules

The next experiment assesses compensation for voicing and place assimilation in French-English bilingual children, using the same design and stimuli as in Experiment 1.

Experiment 2: Bilinguals

Methods

Preregistration

This experiment was pre-registered – after completion of Experiment 1 – through the Open Science Framework platform (available at https://osf.io/52z9g/). At the moment of pre-registration, two bilingual children had already been recruited but their data had not yet been downloaded from the application nor examined in any way. The document specified the number of participants, exclusion criteria, and data analysis plan. With the exception of a small modification to the exclusion criteria (see below), the study was conducted as pre-registered.

Participants

Twenty-one French-English bilingual 6-year-olds (10 girls, 11 boys, mean age: 71.44 months, age range: 66.66 – 77.40 months) participated in this study. An additional five children were tested but not included in the analysis due to failure to pass the training criteria ($n = 3$), lack of knowledge of some of the words used during training according to parental report ($n = 1$, see details of exclusion criteria below), or because they were more than 3SD above the average age of the monolingual group ($n = 1$).

Bilinguals' language background was assessed through a parental questionnaire, in which parents estimated the percentage of exposure to each language both since birth and in the past 6 months, as well as the percentage of the child's output in each language. They were also asked to rate their child's comprehension and expression skills in each language on a scale from 1 (doesn't speak the language) to 5 (comparable to a monolingual of the same age). Finally, they provided information regarding their own native language(s) and language use with their child. The questionnaire data are summarized in Table 3. All children had

Table 3

Summary of Bilinguals' Language Background

		Main language used with the child	
Parents' native languages	N	*Mother*	*Father*
Both French monolinguals*	1	French	French
Both English monolinguals	1	English	English
French mother, English father	1	French	English
English mother, French father**	11	English (*n* = 11)	French (*n* = 11)
French-English bilingual mother, French father	5	English (*n* = 5)	French (*n* = 5)
English mother, French-English bilingual father	1	English	French
Both French-English bilinguals	1	French	English

	French		*English*		
Language input	Mean	Range	Mean	Range	T-test
Percentage of exposure (since birth)	60%	(40% – 75%)	40%	(25% – 60%)	$t(20) = 4.76$, $p = .0001$
Percentage of exposure (current)	62%	(40% – 80%)	38%	(20% – 60%)	$t(20) = 5.12$, $p < .0001$

	French		*English*		
Language use	Mean	Range	Mean	Range	T-test
Percentage of output (current)	71%	(50% – 100%)	29%	(0% – 50%)	$t(20) = 5.60$, $p < .0001$
Comprehension rating (1 - 5)	4.95	(4 – 5)	4.55	(3.5 – 5)	$t(20) = 3.44$, $p = .003$
Expression rating (1 - 5)	4.95	(4 – 5)	3.79	(1 – 5)	$t(20) = 4.06$, $p = .0006$

** This family lived in Singapore. The child heard French at home and English from his nanny as well as from staff and classmates at an English kindergarten.*

*** In two cases one of the parents was bilingual in a second language that was neither English nor French, but never spoke the second language to the child.*

received regular exposure to both French and English since their first year of life (20 children had begun their bilingual exposure at birth, the remaining one had heard French since birth and English from the age of 6 months), with exposure to each language comprising between 25% and 75% of their total exposure. The majority of the participants had a native English speaking mother who talked to them mostly or only in English since birth ($n = 18$), but overall children heard more French than English. English-speaking parents spoke a variety of English dialects (British, North American, Australian and South-African).

Materials

All materials were the same as those used in Experiment 1.

Procedure

The experiment took place either in our babylab ($n = 13$), or in a quiet room at the participant's home located in the Paris region ($n = 8$). About half of the children were tested on the same iPad Air 2 as those in Experiment 1 (model A1566); the others were tested on a 5th-generation iPad (model MP2F2NF/A) running on iOS 10. Both iPads had the same screen size of 9.7 inch.

All procedures were the same as those used in Experiment 1, except that - as bilingual children's vocabulary could be smaller than their monolingual peers' due to their reduced exposure to each language - the main French-speaking caregiver filled in a vocabulary questionnaire which was used to check the child's knowledge of the training and test items.

Exclusion criteria

In addition to the exclusion criteria defined in Experiment 1, two other conditions were imposed on the bilingual group. As pre-registered, children were excluded from analysis if *a*) any of the words used during training was reported by the parents as not known to the child, or *b*) more than two words in any of the four experimental

conditions was reported as not known to the child. As we had initially selected and recorded a few extra training items, we were sometimes able to replace an unknown training item with an alternative known word with the same final contrast, thus preventing participant exclusion. This procedure, which was not pre-registered, was applied on two occasions.

Finally, individual test trials were excluded from analysis if the item was reported as not known by the child in the vocabulary questionnaire.

Results and discussion

Bilinguals' data was analyzed using the same generalized linear mixed model as used for monolinguals in Experiment 1. Four trials, each from a different child, were excluded from analysis based on their vocabulary questionnaires. A summary of the results is shown in Table 4. Figure 3 shows the proportion of trials where the picture of the familiar object was chosen, split by condition.

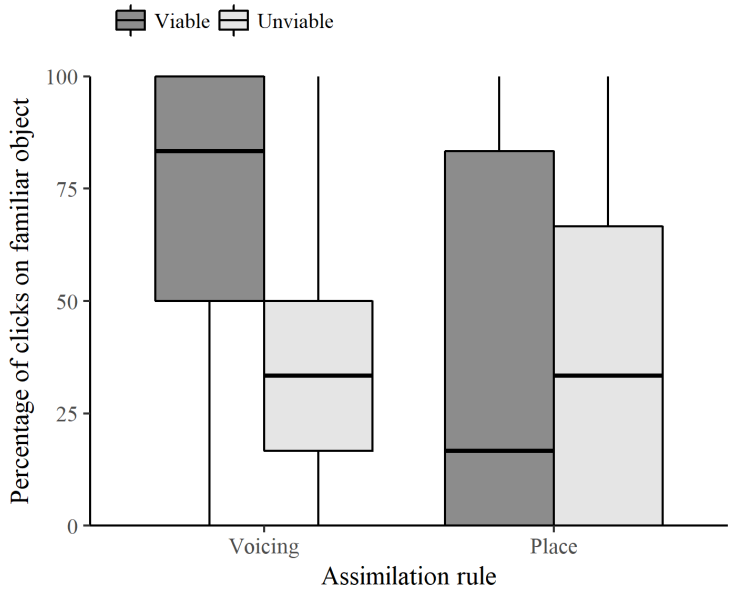


Figure 3. Boxplots representing the percentage of clicks on the picture of the familiar object per condition for bilingual children in Experiment 2. The bold horizontal line inside each box represents the median, while the bottom and top boundaries represent the first and third quartiles, respectively. Vertical lines extend to the highest and lowest values within 1.5 times the inter-quartile range above and below the boundaries.

Table 4

Summary of the generalized linear mixed model for bilinguals.

Fixed effect	β	*SE*	*z*	*p*
Intercept (*voicing, viable*)	2.77	1.29	2.15	0.03
Context (*unviable*)	-3.98	0.96	-4.14	<.0001
Rule (*place*)	-3.65	1.19	-3.06	0.002
Interaction: Rule x Context	4.34	1.03	4.20	<.0001

The resulting GLMM was tested against the null model through a likelihood ratio test, revealing a good fit of the data ($\chi^2(3) = 16.8$, $p = 0.0008$). As monolinguals, bilinguals showed significant effects of rule and context, as well as an interaction of rule by context. The intercept, corresponding to the mean in *voicing viable* trials, revealed a significant preference for the picture of the familiar object over the red cross. The direction of the effects indicates, as in the monolingual group, a higher preference for the familiar object when hearing *voicing viable* phrases compared to both *voicing unviable* and *place viable* phrases. A releveling of the rule variable with *place* as the baseline revealed no significant effect of context for place assimilation ($\beta = 0.36$, $SE = 0.57$, n.s.), that is, bilingual children did not compensate for this rule. To directly compare the responses of both groups, a post-hoc analysis - not included in the preregistration - was performed, pooling data from both experiments, and including *group* (monolingual, bilingual) as a fixed factor in a triple interaction with rule and context. The resulting model revealed no significant effect of group, neither as a simple effect, nor in interaction with the other fixed factors. Thus, bilingual children's results do not differ from those of their monolingual peers.

Further post-hoc explorations of the relation between bilingual children's compensation patterns and their language background revealed no correlation between their voicing difference score, defined as the difference between the percentage of clicks on the familiar object in the viable minus the unviable conditions

in voicing trials, and their percentage of French exposure (since birth: r = -0.40, p = 0.07; current: r = 0.10, $p > 0.1$). Similarly, there was no correlation between their place difference score and their percentage of English exposure (since birth: r = 0.37, p = 0.1; current: r = -0.13, $p > 0.1$).

Overall, these results show that, like monolingual French children, French-English bilingual children compensate for voicing but not for place assimilation when listening to French. The absence of a correlation between their language input and the magnitude of their compensation for either rule is not surprising given the high overlap between the compensation patterns of monolinguals and bilinguals. However, the number of trials per condition may have been insufficient to observe individual differences. It should also be noted that the bilingual population in this study was relatively homogeneous, as most children had similar language backgrounds (for instance, most of them had an English speaking mother and a French speaking father), and the observed range of children's French exposure was narrow (mean exposure to French: 60%, range: 40% - 75%). Different results might thus be observed in a more heterogeneous bilingual sample.

General discussion

The present study investigated monolingual and bilingual 6-year-olds' compensation for voicing and place assimilation in French sentences. Using a word recognition task in a tablet-based experiment, we first showed that monolingual French children compensate for a native voicing assimilation rule, but not for a non-native place assimilation rule. That is, they were more likely to recognize the name of a familiar object when presented with a word-final voicing change in a viable context than in an unviable context for voicing assimilation, while they failed to recognize the name when presented with a word-final place change, irrespective of the following context. These results are similar to previous findings with monolingual toddlers (Skoruppa, Mani & Peperkamp, 2013; Skoruppa, Mani, Plunkett, et al., 2013), but with a critical

addition: We simultaneously tested children's processing of native and non-native rules, showing for the first time a significant difference as indicated by the interaction between rule (voicing vs. place assimilation) and context (viable vs. unviable). This interaction was previously reported for adults (Darcy et al., 2009), and it is crucial for our research question on compensation for assimilation in bilinguals.

In our second experiment, we tested French-English bilinguals using the same game, and showed that their compensation pattern in French is not different from that of their monolingual peers. That is, they also compensated for voicing but not for place assimilation, and showed a significant interaction between rule and context. Furthermore, their performance did not correlate with the amount of exposure (whether counted from birth or over the last 6 months) to their two languages. These results indicate that, in spite of their reduced exposure to French, bilingual 6-year-olds have successfully developed a compensation mechanism for a complex phonological rule, without any apparent interference from their second language.

It would be interesting to know how our group of bilingual children would perform on English sentences. Despite the fact that 80% of them had an English-speaking mother, they were on average slightly dominant in French, both in terms of language input (mean: 60% French) and in terms of language use (mean: 71% French, and higher ratings for both comprehension and expression in French). This leaves open the possibility that their performance on English would not be native-like but show interference from French, with compensation not only for place but also for voicing assimilation. At a younger age, French-dominant bilinguals might even compensate only for voicing assimilation when listening to English. Conversely, testing in an Anglophone country would probably allow us to obtain an English-dominant sample, in which we might observe native-like performance on English but not on French sentences. To the best of our knowledge, cross-linguistic transfer in bilingual children's phonology has so far only been shown in production (Paradis, 2001). Evidence from second language learners, however, suggests that it could

likewise occur in the acquisition of compensation for assimilation. Darcy et al. (2007) tested native speakers of French learning English and native speakers of English learning French, in both their native and their second language (L2). The results revealed that beginner learners tend to transfer their native-rule perception to the newly learnt language: English learners of French compensated for place but not for voicing assimilation not only in English but also in French sentences, and vice versa for French learners of English. Advanced learners, however, correctly compensated for their native rule in their native language and for their L2 rule in their L2 language, thus showing full separation of the phonological rules. Future work should thus assess children with a fully crossed design at different ages, testing each child not only on both phonological rules but also in both languages. However, given that bilingual children often have smaller vocabularies in each language in comparison to monolinguals of the same age (Bialystok, Luk, Peets, & Yang, 2010; Marchman, Fernald & Hurtado, 2010; Hoff et al., 2012), finding a sufficient number of known assimilable words in both languages and testing children in such a demanding task may prove difficult, especially at younger ages.

More generally, even for monolinguals questions remain open as to when and how assimilation rules are acquired. Recall that Skoruppa, Mani, Plunkett, et al. (2013) found adult-like compensation for assimilation in 24 month-olds. It is difficult to test much younger infants with a lexical task, as they do not know enough words yet. However, using a mismatch paradigm with EEG recordings, Fort, Brusini, Carbajal, Sun and Peperkamp (2017) found that even at 14 months of age infants have some knowledge of native assimilation rules. That is, 14-month-old French-learning infants - like French adults (Sun et al., 2015) - failed to discriminate a voicing contrast in a viable assimilation context (e.g., [ofbe] vs. [ovbe]: no mismatch response), while they successfully detected it in an unviable assimilation context (e.g., [ofne] vs. [ovne]: mismatch response). Thus, they appeared to already have acquired voicing assimilation and perceptually compensate for this language-specific rule. Fort et al. speculate that infants' acquisition of voicing assimilation is triggered

by the tendency of words to be repeated during conversations, and hence for assimilated and non-assimilated forms of words to cluster together within short stretches of speech. Thus, even without having access to word meanings, infants could infer that word-final voicing differences reflect systematic variation induced by voicing assimilation. Still, we cannot rule out an alternative possibility: Fort et al. (2017) did not test French infants on a non-native rule, and to the extent that assimilation is phonetically motivated - it is rooted in coarticulation - it would not be completely unexpected if they likewise compensated for English place assimilation. This would be evidence that rather than acquiring their native assimilation rules, infants have to unlearn non-native rules. Such a scenario would fit well with findings in adults showing different amounts of compensation even for non-native assimilation rules (Gow & Im, 2004; Mitterer, Csépe & Blomert, 2006; Mitterer, Csépe, Honbolygo & Blomert, 2006; Darcy et al., 2009).[4]

Whether native assimilation rules must be acquired or non-native ones unlearned, bilinguals must be able to fully separate their two languages' phonologies in order to attain native-like compensation in both. This is easier for some language pairs than for others. Sundara & Scutellaro (2011) specifically mention the role of rhythmic properties in the speed of separation and, consequently, of phonological acquisition in bilingual infants. Focusing on the acquisition of sound inventories, they found that an acoustically similar contrast that is phonemic in only one of the languages is acquired faster by infants exposed to Spanish and English, which belong to different rhythmic classes, than by infants exposed to Spanish and Catalan, which belong to the same rhythmic class. French and English, the languages under scrutiny in our study, belong to different rhythmic classes (Ramus, Nespor & Mehler, 1999).

[4] The fact that neither we nor Skoruppa, Mani and Peperkamp (2013) and Skoruppa, Mani, Plunkett, et al. (2013) found some small amount of compensation for non-native assimilation might be due to a lack of power. Indeed, experiments with toddlers and young children typically have a small number of trials (e.g., our experiment had 6 trials per condition, compared to 16 for adults in Darcy et al., 2009).

Moreover, their phoneme inventories are very different, and while they share many stop consonants - which constitute about half of the consonants that undergo either voicing or place assimilation - these consonants have different phonetic implementations (French contrasts voiceless unaspirated with prevoiced stops, while English contrasts aspirated with voiced stops). We therefore expect that the acquisition of the respective assimilation rules by French-English bilingual children should be relatively easy. An example of a harder case is provided by Spanish and Catalan, which not only are rhythmically similar but whose phoneme inventories are largely overlapping, with similar phonetic implementations of the shared phonemes. Bilingual Spanish-Catalan children must learn that while both of their languages have nasal place assimilation, Catalan additionally has voicing assimilation in word-final obstruents (Wheeler, 2005; Recasens & Mira, 2012). Thus, we expect bilingual Spanish-Catalan to be delayed compared to bilingual French-English children as far as compensation for voicing assimilation in one but not the other language is concerned.

We conclude with a few methodological considerations: From a conceptual point of view, our design differed from those of Skoruppa, Mani and Peperkamp (2013) and Skoruppa, Mani, Plunkett, et al. (2013), which were based on minimal pairs of known and novel words. That is, toddlers in these studies were introduced to unknown objects (say, a spinning wheel), whose labels (e.g., [byz]) differed from that of known objects (a bus, in French: [bys]) in just the final consonant. In the viable condition, the sentences were therefore ambiguous as they could refer both to the known and to the novel object. Piloting showed that the use of such minimal pairs did not work well with older children (nor with adults).[5] In our design, known

[5] On the one hand, if both the familiar and the novel object were introduced by their names prior to the clicking request, 5-year old children (as well as adults) performed the task at an acoustic level; that is, they showed a bias for the novel object. On the other hand, if neither of the objects was named prior to the clicking request, they showed a bias for the familiar object, regardless of context. A detailed report of our pilot studies is available at the OSF project page, https://osf.io/52z9g/.

objects were therefore always contrasted with a red cross; children were asked to touch the object if its label was pronounced correctly and the red cross otherwise. Thus, there was no ambiguity; if they had perfect knowledge of assimilation, children should touch the known object in the viable condition and the red cross in the unviable condition. This is akin to the word recognition task used in the studies with adults by Darcy and colleagues (Darcy et al. 2007; Darcy et al. 2009), in which adults first heard a target word and then had to decide whether it was present and correctly pronounced in a following sentence. It would be interesting to use the present design with younger children; given the absence of ambiguity in the viable condition, we expect the compensation effects to be larger than those found previously in toddlers (Skoruppa, Mani & Peperkamp, 2013; Skoruppa, Mani, Plunkett, et al., 2013).

Our design also differed from those used previously in that we used a video-game for tablets, adapted from Cristia (2016), rather than a traditional setting in which the child sits in front of a computer screen. A wide range of tasks can be implemented on tablets to study cognitive development (Semmelmann et al., 2016). Moreover, in a direct comparison with 1- to 4-year-old children on a word-recognition task, Frank, Sugarman, Horowitz, Lewis, and Yurovsky (2016) found that a tablet-based paradigm compared favorably both with an eye tracking paradigm and an in-person storybook paradigm. Our use of a tablet facilitated the recruitment of participants in different locations, and allowed us to keep children engaged in the task for the whole duration of the experiment, with none of them abandoning before the end of the game. During pilot tests, children as young as 4 years likewise showed low drop-out rates. This fun and portable low-cost setup can thus be used in further research investigating the acquisition of assimilation and other phonological rules in diverse populations of young children around the world.

Acknowledgments

We would like to thank all the parents and children that participated, as well as the kindergarten school that kindly allowed us to test monolingual children in their facilities, and Mr. Derek Ferguson from the Roaming School House in Paris for his help with the recruitment of bilingual children. We are also grateful to Alex Cristia for lending us the tablet, and to Anne Christophe for making the recordings. This work was supported by the Agence Nationale pour la Recherche under grants ANR-17-CE28-0007-01, ANR-10-LABX-0087 IEC, and ANR-10-IDEX-0001-02 PSL; and the Ecole des Neurosciences de Paris Ile-de-France (ENP Graduate Program).

References

Adda-Decker, M., & Hallé, P. (2007). Bayesian framework for voicing alternation and assimilation studies on large corpora in French. In *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 613-616). Saarbrücken, Germany.

Allen, G. D., & Hawkins, S. (1980). Phonological rhythm: Definition and development. In *Child phonology* (pp. 227-256).

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, *13*(4), 525-531.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer (Version: 5.3.86) [Computer software]. Retrieved 17th March, 2015 from http://www.praat.org/.

Bosch, L., & Sebastián-Gallés, N. (2003a). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and speech*, *46*(2-3), 217-243.

Bosch, L., & Sebastián-Gallés, N. (2003b). Language experience and the perception of a voicing contrast in fricatives: Infant and adult data. In *Proceedings of the 15th*

International Congress of Phonetic Sciences (pp. 1987-1990). Barcelona, Spain: UAB/Casual Prods.

Buckler, H., Goy, H., & Johnson, E. K. (2018). What infant-directed speech tells us about the development of compensation for assimilation. *Journal of Phonetics*, *66*, 45-62.

Coenen, E., Zwitserlood, P., & Bölte, J. (2001). Variation and assimilation in German: consequences for lexical access and representation. *Language and Cognitive Processes*, *16*(5-6), 535-564.

Cristia, A. (2016). LSCP iDevXXI App [Computer software]. Retrieved 30th November, 2016 from https://github.com/alecristia/mandy_ipadvocabtest/.

Darcy, I., Peperkamp, S., & Dupoux, E. (2007). Bilinguals play by the rules: perceptual compensation for assimilation in late L2-learners. In: J. Cole & J. Hualde (eds.) *Laboratory Phonology 9*. Berlin: Mouton de Gruyter, 411-442.

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. *Variation and gradience in phonetics and phonology*, *14*, 265-309.

Dilley, L. C., & Pitt, M. A. (2007). A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. *The Journal of the Acoustical Society of America*, *122*(4), 2340-2353.

Fabiano-Smith, L., Oglivie, T., Maiefski, O., & Schertz, J. (2015). Acquisition of the stop-spirant alternation in bilingual Mexican Spanish–English speaking children: Theoretical and clinical implications. *Clinical linguistics & phonetics*, *29*(1), 1-26.

Fennell, C. T., Byers-Heinlein, K., & Werker, J. F. (2007). Using speech sounds to guide word learning: The case of bilingual infants. *Child development*, *78*(5), 1510-1525.

Fort, M., Brusini, P., Carbajal, M. J., Sun, Y., & Peperkamp, S. (2017). A novel form of perceptual attunement: Context-dependent perception of a native contrast in 14-month-old infants. *Developmental cognitive neuroscience*, *26*, 45-51.

Fox, J., & Weisberg, S. (2011). An *R* companion to applied regression, second edition [Computer software]. Thousand Oaks, CA: Sage.

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development*, *17*(1), 1-17.

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human perception and performance*, *22*(1), 144.

Gaskell, M. G., & Snoeren, N. D. (2008). The impact of strong assimilation on the perception of connected speech. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1632.

Gow Jr, D. W., & Im, A. M. (2004). A cross-linguistic examination of assimilation context effects. *Journal of Memory and Language*, *51*(2), 279-296.

Havy, M., Bouchon, C., & Nazzi, T. (2016). Phonetic processing when learning words: The case of bilingual infants. *International Journal of Behavioral Development*, *40*(1), 41-52.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of child language*, *39*(1), 1-27.

Liu, L., & Kager, R. (2015). Bilingual exposure influences infant VOT perception. *Infant Behavior and Development*, *38*, 27-36.

MacLeod, A. A., & Fabiano-Smith, L. (2015). The acquisition of allophones among bilingual Spanish–English and French–English 3-year-old children. *Clinical linguistics & phonetics*, *29*(3), 167-184.

Marchman, V. A., Fernald, A., & Hurtado, N. (2010). How vocabulary size in two languages relates to efficiency in spoken word recognition by young Spanish–English bilinguals. *Journal of Child Language*, *37*(4), 817-840.

Mitterer, H., & Blomert, L. (2003). Coping with phonological assimilation in speech perception: Evidence for early compensation. *Perception & Psychophysics*, *65*(6), 956-969.

Mitterer, H., Csépe, V., & Blomert, L. (2006). The role of perceptual integration in the recognition of assimilated word forms. *Quarterly Journal of Experimental Psychology*, *59*(8), 1395-1424.

Mitterer, H., Csépe, V., Honbolygo, F., & Blomert, L. (2006). The recognition of phonologically assimilated words does not depend on specific language experience. *Cognitive Science*, *30*(3), 451-479.

Nicoladis, E., & Paradis, J. (2011). Learning to liaise and elide *comme il faut*: evidence from bilingual children. *Journal of Child Language*, *38*(4), 701-730.

Paradis, J. (2001). Do bilingual two-year-olds have separate phonological systems? *International journal of bilingualism*, *5*(1), 19-38.

R Core Team (2017). R: A language and environment for statistical computing (Version 3.3.3) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive psychology*, *59*(1), 96-121.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265-292.

Recasens, D., & Mira, M. (2012). Voicing assimilation in Catalan two-consonant clusters. *Journal of Phonetics*, *40*(5), 639-654.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object set: The role of surface detail in basic-level object recognition. Perception, 33, 217-236.

Semmelmann, K., Nordt, M., Sommer, K., Röhnke, R., Mount, L., Prüfer, H., Terwiel, S., Meissner, T.W., Koldewyn, K. and Weigelt, S. (2016). U can touch this: how tablets can be used to study cognitive development. *Frontiers in psychology*, *7*, 1021.

Skoruppa, K., Mani, N., & Peperkamp, S. (2013). Toddlers' processing of phonological rules: Early compensation for assimilation in English and French. *Child Development*, 84, 313-330.

Skoruppa, K., Mani, N., Plunkett, K., Cabrol, D., & Peperkamp, S. (2013). Early word recognition in sentence context: French and English 2-year-olds' sensitivity to sentence-medial mispronunciations and assimilations. *Infancy*, 18, 1007-1029.

Sun, Y., Giavazzi, M., Adda-Decker, M., Barbosa, L. S., Kouider, S., Bachoud-Lévi, A. C., Jacquemot, C., & Peperkamp, S. (2015). Complex linguistic rules modulate early auditory brain responses. *Brain and language*, *149*, 55-65.

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of/d-/in monolingual and bilingual acquisition of English. *Cognition*, *100*(2), 369-388.

Sundara, M., & Scutellaro, A. (2011). Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*, *39*(4), 505-513.

Wheeler, M. W. (2005). *The phonology of Catalan*, Oxford: Oxford University Press.

Appendix A

List of Test Items Used in Experiments 1 & 2

Voicing items			Altered form	Viable context	Unviable context
brosse	[bʁɔs]	"brush"	[bʁɔz]	*devant toi* "in front of you"	*là-bas* "over there"
chaise	[ʃɛz]	"chair"	[ʃɛs]	*par ici* "over here"	*maintenant* "now"
couche	[kuʃ]	"nappy"	[kuʒ]	*devant toi* "in front of you"	*là-devant* "there in front"
crêpe	[kʁɛp]	"pancake"	[kʁɛb]	*juste ici* "just here"	*là-devant* "there in front"
douche	[duʃ]	"shower"	[duʒ]	*devant toi* "in front of you"	*là-devant* "there in front"
fraise	[fʁɛz]	"strawberry"	[fʁɛs]	*par ici* "over here"	*là-devant* "there in front"
langue	[lɑ̃g]	"tongue"	[lɑ̃k]	*s'il te plaît* "please"	*là-devant* "there in front"
porte	[pɔʁt]	"door"	[pɔʁd]	*juste ici* "just here"	*maintenant* "now"
robe	[ʁɔb]	"dress"	[ʁɔp]	*s'il te plaît* "please"	*là-devant* "there in front"
singe	[sɛ̃ʒ]	"monkey"	[sɛ̃ʃ]	*s'il te plaît* "please"	*là-bas* "over there"
tasse	[tas]	"cup"	[taz]	*devant toi* "in front of you"	*maintenant* "now"
vache	[vaʃ]	"cow"	[vaʒ]	*devant toi* "in front of you"	*là-bas* "over there"
Place items			Altered form	Viable context	Unviable context
boîte	[bwat]	"box"	[bwap]	*maintenant* "now"	*devant toi* "in front of you"
botte	[bɔt]	"boot"	[bɔp]	*maintenant* "now"	*devant toi* "in front of you"
chaîne	[ʃɛn]	"chain"	[ʃɛm]	*par ici* "over here"	*devant toi* "in front of you"
clown	[klun]	"clown"	[klum]	*par ici* "over here"	*s'il te plaît* "please"
corde	[kɔʁd]	"rope"	[kɔʁb]	*par ici* "over here"	*là-devant* "there in front"
crotte	[kʁɔt]	"poop"	[kʁɔp]	*maintenant* "now"	*là-devant* "there in front"
goutte	[gut]	"water drop"	[gup]	*maintenant* "now"	*là-devant* "there in front"
lune	[lyn]	"moon"	[lym]	*par ici* "over here"	*là-devant* "there in front"
monde	[mɔ̃d]	"world"	[mɔ̃b]	*par ici* "over here"	*là-devant* "there in front"
pâtes	[pat]	"pasta"	[pap]	*maintenant* "now"	*là-devant* "there in front"
reine	[ʁɛn]	"queen"	[ʁɛm]	*par ici* "over here"	*s'il te plaît* "please"
viande	[vjɑ̃d]	"meat"	[vjɑ̃b]	*par ici* "over here"	*là-devant* "there in front"

4.2 Additional studies

In Carbajal, Chartofylaka, Hamilton, Fiévet and Peperkamp (in revision), we implemented a modified version of the experimental design used by Skoruppa, Mani and Peperkamp (2013). In this section, we describe a series of five pilot studies conducted with children and adults which allowed us to reach the final experimental design. We further discuss the implications of the pilot results for the interpretation of this phonological process.

4.2.1 Pilot experiment 1

4.2.1.1 Introduction

Using an interactive pointing task, Skoruppa, Mani and Peperkamp (2013) tested three groups of monolingual 2.5- to 3-year-old children on their perception of native and non-native assimilation rules. English toddlers were tested on their perception of a native place assimilation rule, while two groups of French toddlers were tested on their compensation for a native voicing rule and a non-native place assimilation rule, separately.

Their experimental design was composed of three stages – pre-training, training, and test – similarly to the three stages described in our study (Carbajal et al., in revision). Each trial began with the presentation of a familiar object and a novel object, and was followed by a pointing request (see trial sequence in Figure 4.1). Crucially, in test trials, the label of the novel object differed from the familiar word only in the relevant one-feature change (voicing or place, depending on the rule) on its final consonant. For instance, in the place assimilation experiment with English toddlers, the item *moon* was matched with a novel object[1] that was labelled as *moom*, as shown in the example in Figure 4.1. After the sequential presentation of the two objects, children saw both items side by side, and heard one of four possible test sentences. In control trials, either the familiar or the novel words appeared sentence finally, as in *Can you find the moon?* (familiar condition) or *Can you find the moom?* (novel condition). In experimental trials, the novel word (here, *moom*) was produced in sentence-medial position, followed by either a viable (e.g., *Can you find the moom please?*) or an unviable (e.g., *Can you find the moom dear?*) context for assimilation.

[1]Some of the novel objects were real items that are generally unknown to toddlers.

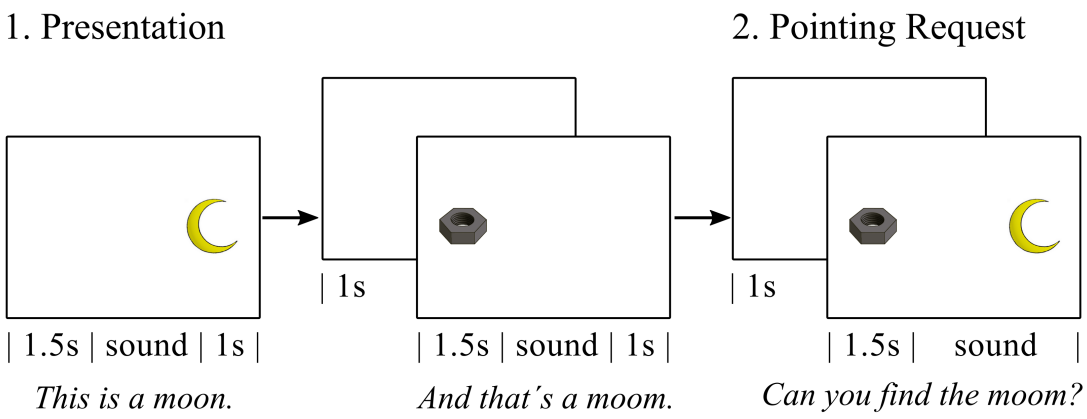


Figure 4.1: Example of a trial sequence in Skoruppa, Mani and Peperkamp (2013).

Skoruppa, Mani and Peperkamp (2013) found that, in experimental trials, toddlers compensated for their native assimilation rules (English toddlers compensated for place assimilation, and French toddlers compensated for voicing assimilation), that is, they pointed more often at the familiar object when the novel word was followed by a viable context for assimilation than when it was followed by an unviable context. However, children tested on a non-native rule (here, French toddlers tested on place assimilation) did not show this compensation pattern, and instead pointed at the novel object more often than the familiar object, regardless of the context. In control trials, all toddlers were able to discriminate the familiar word from the novel word (i.e., they pointed more at the familiar object than the novel object in the familiar condition, and vice versa in the novel condition).

With the intention of developing an experiment that could be used to test bilingual children, in this first pilot we implemented an adaptation of Skoruppa, Mani and Peperkamp's (2013) paradigm for an iPad application (Cristia, 2016), and extended the design to test both native and non-native rules in a single test session.

4.2.1.2 Methods

Participants

Thirty-two French monolingual 4- to 5-year-old children participated in the first pilot study (18 girls, mean age: 54.72 months, age range: 49.03 - 59.51 months). An additional 20 children participated but were excluded from analysis due to failing the training criteria (see Exclusion criteria below).

Materials

The set of twenty-four familiar nouns and pictures used as test items in this pilot were the same as those described in Carbajal et al. (in revision). However, as in Skoruppa, Mani and Peperkamp (2013), each familiar object was additionally paired with a picture of a novel, unfamiliar object. The pictures were chosen matching visual saliency, number and animacy of the corresponding familiar object as closely as possible. For instance, pictures of animals or characters were paired with rare or novel creatures, and pictures of plural nouns (e.g., *pâtes* "pasta") were paired with plural unfamiliar items. Each novel object was assigned a label using the same assimilated forms described in our final study (Carbajal et al., in revision). For example, for the item *tasse* [tas] "cup", its matching novel object was labelled as **taze* [taz]. As in Skoruppa, Mani and Peperkamp (2013), introduction phrases were recorded for each familiar word and each novel word, such as *Voici une tasse* "Here's a cup", and *Et voilà une taze* "And there's a [taz]".

Each of the 24 assimilated forms (12 for voicing and 12 for place assimilation) was embedded in two short sentences with a touching request, creating a total of 48 test phrases. One of the two sentences provided a *viable* context for assimilation (e.g., *Touche la [taz] devant toi !* "Touch the [taz] in front of you"), and the other one an *unviable* context (e.g., *Touche la [taz] maintenant !* "Touch the [taz] now"). Note that here, unlike in our final study, the phrases in the *viable* context have an ambiguous interpretation, as the label produced in test sentences (here, [taz]) could be interpreted both as the assimilated form of the familiar word, or as the label of the novel object. All sentences used in this pilot were identical to those used in Carbajal et al. (in revision).

Finally, eleven items were selected for pre-training ($n = 3$) and training ($n = 8$). The list of familiar nouns and their pictures was composed of the same 9 items used in our final study, plus two additional familiar objects to complete a set of 8 training items (*tête* "head" and *frites* "fries"). As with the test items, eleven pictures of unfamiliar objects were chosen to match each familiar item. Each novel object in the pre-training list was assigned a nonce label that differed from the original word in varying degrees, starting with a very distant pair (**moutte* [mut] for *balle* [bal] "ball"), and following with two closer pairs (**kime* [kim] for *coeur* [kœʁ] "heart", and **pouke* [puk] for *poule* [pul] "hen"). The novel items in the training list were each assigned a label differing only in voicing ($n = 4$) or place of articulation ($n = 4$) of the final consonant (e.g., *glace* [glas] "ice cream" - **glaze* [glaz]). Each familiar word and each novel word were recorded twice in final position, once in an introduction

phrase (similarly to the test items), and once in a touching request phrase of the form *Touche le/la # !* "Touch the #".

All sentences were recorded by a female native French speaker, who was given the orthographic transcription of words and nonwords to ensure that she produced the correct form (i.e., completely assimilated or unassimilated) in each case. She was instructed to read them in child-directed speech and without pauses. Minor editing and intensity normalization (70 dB) were done using the software *Praat* (Boersma & Weenink, 2014). To verify that the target consonants in the test sentences were always produced in their fully assimilated form, we asked 10 adult monolingual French speakers to identify the consonants. Participants listened to the final V(C)C segments of each assimilated word (e.g., [az] from *Touche la [taz] maintenant*) in both viable and unviable conditions, and were asked to categorize the final consonant. In each trial, they were given two options to choose from: either the assimilated form (in this example, [z]) or the unassimilated form (here, [s]). To avoid a bias for the assimilated form, control samples extracted from the unassimilated words produced in sentence medial position were included as distractors (e.g., [as] from *Touche la tasse maintenant*). The order of presentation of the items, rules (place, voicing) and contexts (viable, unviable, control) were fully randomized. All items were recognized by at least 9 out of 10 participants in all conditions. Paired-samples t-tests revealed no difference in the identification scores of consonants in viable and unviable contexts, neither for the voicing assimilation rule ($t(11) = 0.43$, $p > 0.1$), nor for the place assimilation rule ($t(11) = -1.48$, $p > 0.1$).

Procedure

The setup in this first pilot was similar to that used in Carbajal et al. (in revision). However, following Skoruppa's design, children were presented with both a familiar and a novel object before hearing the test phrases. The exact sequence of events in each trial was as follows: first, the picture of the familiar object (for example, a cup) appeared on one side of the screen, while an introduction phrase labelling the object was played (e.g., *Voici une tasse* "Here's a cup"). Immediately after, the picture of the novel object appeared on the opposite side of the screen, while its introduction phrase was played (e.g., *Et voilà une taze* "And there's a [taz]"). Following a 3 second pause, the girl appeared in the middle of the screen and waited until the child clicked on her. After clicking, the touch response became blocked while the character produced the request phrase. The touch response reactivated immediately afterwards, allowing the child to give an answer by clicking either on the picture of the familiar or the

novel object. No time limit was given to produce an answer, however, if the child took more than 6 seconds to respond, the character would make a sound to remind the child to make a choice.

The experiment was composed of the same 3 stages as described in our final study, namely 3 *pre-training* trials, 8 *training* trials[2] and 24 *test* trials. However, due to the sequential introduction of the two objects prior to the touching request, and to a difference in the reward system, in which a pause was made to give the child a sticker after every star earned (instead of waiting until the end of the experiment as in the final study), the total duration of the experiment was approximately 20 minutes, that is, twice the duration of the final version.

One final difference between this first pilot and the final study was the order of trials in the counterbalancing lists. While in the final version all children saw 2 voicing viable and 1 voicing unviable trials in the first block, here, half of the children began with that ratio of trials, while the other half saw the opposite proportion. As in the final version, the ratio of viable to unviable trials for each rule alternated from one block to the other.

Exclusion criteria

As in Carbajal et al. (in revision), a limit on the number of errors accepted in the training phase was imposed. Specifically, children were excluded from analysis if they *a*) made mistakes in more than two items, or *b*) made more than two mistakes on the same item.

4.2.1.3 Results and discussion

Children's responses were automatically collected by the app and coded as a binary dependent variable representing whether the child chose the familiar or the novel object in each trial. Figure 4.2 shows the proportion of trials where the familiar object was chosen, split by condition. Responses were analyzed with the same generalized linear mixed model described in our final study, modeling the likelihood of choosing the familiar object. The model included fixed effects of assimilation rule (*voicing*, *place*) and context (*viable*, *unviable*), as well as the interaction between them. A maximal random effects structure was used (Barr, Levy, Scheepers & Tily, 2013), including an intercept as well as slopes for rule, context and their interaction by participant, and an intercept and slope for context by item. The

[2]As described in the Materials section, in this pilot there were 2 more training items than in the final study.

variables *rule* and *context* were treatment-coded, with *rule* = *voicing* and *context* = *viable* as baseline levels. P-values were obtained for all fixed effects using the *car* package (Fox & Weisberg, 2011).

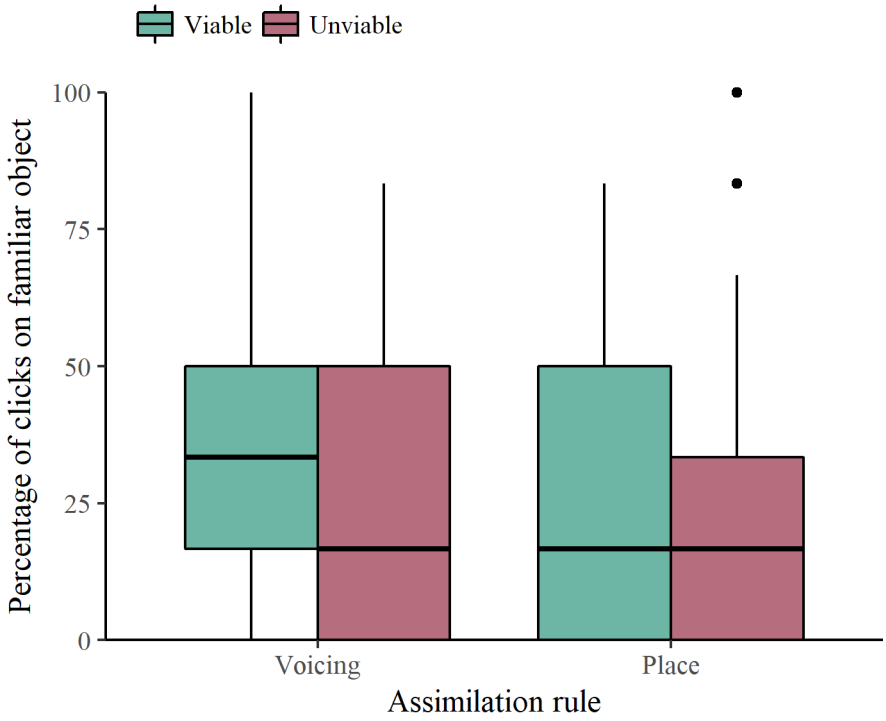


Figure 4.2: Boxplots showing the percentage of clicks on the familiar object per rule and context in Pilot 1.

The estimated intercept ($\beta_0 = -0.72$, $SE = 0.38$, $p = 0.055$)[3] shows that the percentage of trials where the familiar object was chosen in the *voicing viable* condition was marginally below chance. However, a significant effect of context ($\beta = -1.00$, $SE = 0.43$, $p = 0.02$) indicates that children clicked on the familiar object more often upon hearing voicing assimilation in a viable context than in an unviable context. No effect of rule ($\beta = -0.75$, $SE = 0.49$, $p > 0.1$) nor an interaction of rule by context ($\beta = 0.60$, $SE = 0.66$, $p > 0.1$) were found. A releveling of the rule variable setting *place* as baseline revealed no effect of context in place assimilation trials ($\beta = -0.40$, $SE = 0.46$, $p > 0.1$).

These results indicate that children compensated for voicing assimilation, as suggested by a significant difference between *voicing viable* and *voicing unviable* trials. However, although no effect of context was found for place assimilation trials, the difference between the native voicing and the non-native place assimilation rules remained inconclusive, as neither an effect of rule nor an interaction between rule and context were found. Given previous studies with toddlers (Skoruppa, Mani & Peperkamp, 2013; Skoruppa, Mani, Plunkett, et al., 2013) and adults (Darcy et al., 2009), and the large sample size

[3]All estimates are given in odds.

in this first pilot ($n = 32$), these null results were unexpected. As our goal was to use this paradigm to test bilingual children, from whom we may expect smaller effect sizes than from monolinguals, the current results would not allow any comparison between the two groups: not only is the lack of compensation for the non-native rule unclear, but also the effect of compensation for the native rule is too modest to ever observe a potential difference between monolinguals and bilinguals. Indeed, the probability of selecting the familiar object in *voicing viable* trials was below chance ($\beta_0 = -0.72$, representing a probability of only 33%), and the difference between the probability in *voicing viable* and *voicing unviable* trials was a mere 16%, the equivalent of only 1 more click on the familiar object for *voicing viable* trials. We therefore conducted a series of post-hoc analyses, presented below, in order to investigate the potential reasons for these quasi-null results, and finally implemented some changes in a new pilot study.

Post-hoc investigation of Pilot 1

First, we examined whether the game was age-appropriate. While the drop-out rate was very low -with only one child abandoning the game before it was finished - we found that 54% of the participants in the lower half of the age range (i.e., aged 4.0 to 4.5-years-old) failed the training stage. A generalized linear model using *passed* (yes/no) as binary dependent variable and *age* as predictor showed a significant effect of age ($\beta = 0.21$, $SE = 0.09$, $p = 0.026$). This suggests that while highly engaging, the task was found to be generally difficult for the younger half of the participants. This was not surprising given the high rejection rates reported in Skoruppa, Mani and Peperkamp (2013), where 3-year-olds were tested on a similar pointing task. Furthermore, we found a small but significant negative correlation between the age of the participants who passed the training and the general percentage of responses towards the familiar object ($r = -0.35$, $p = 0.026$). However, no correlation with age was found for the voicing difference score ($r = -0.03$, $p = 0.87$) nor for the place difference score ($r = -0.16$, $p = 0.36$). This may simply indicate that younger children have a strong familiarity bias. Thus, increasing the age of the participants by half a year may help reduce both the number of rejected children, and the variance due to bias.

Second, we investigated differences in response patterns for individual items. Figure 4.3 shows the percentage of clicks on the familiar object for each *voicing* item in both viable (AV) and unviable (AU) conditions. We found that, while some *voicing* items produced the expected responses (that is, more clicks on the familiar object in viable than in unviable contexts), others produced high familiarity (e.g.,

singe "monkey" - *sinche*) or novelty preferences (e.g., *crêpe* "pancake" - *crebe*) regardless of context. In one case (*douche* "shower" – *douge*) children clicked more on the familiar object in unviable than in viable contexts. Upon acoustic examination of these items, we had the impression that in most cases the length of the final VC segment was very long, which may have disrupted the prosodic cues that allow the assimilated word to be interpreted as the underlying form.

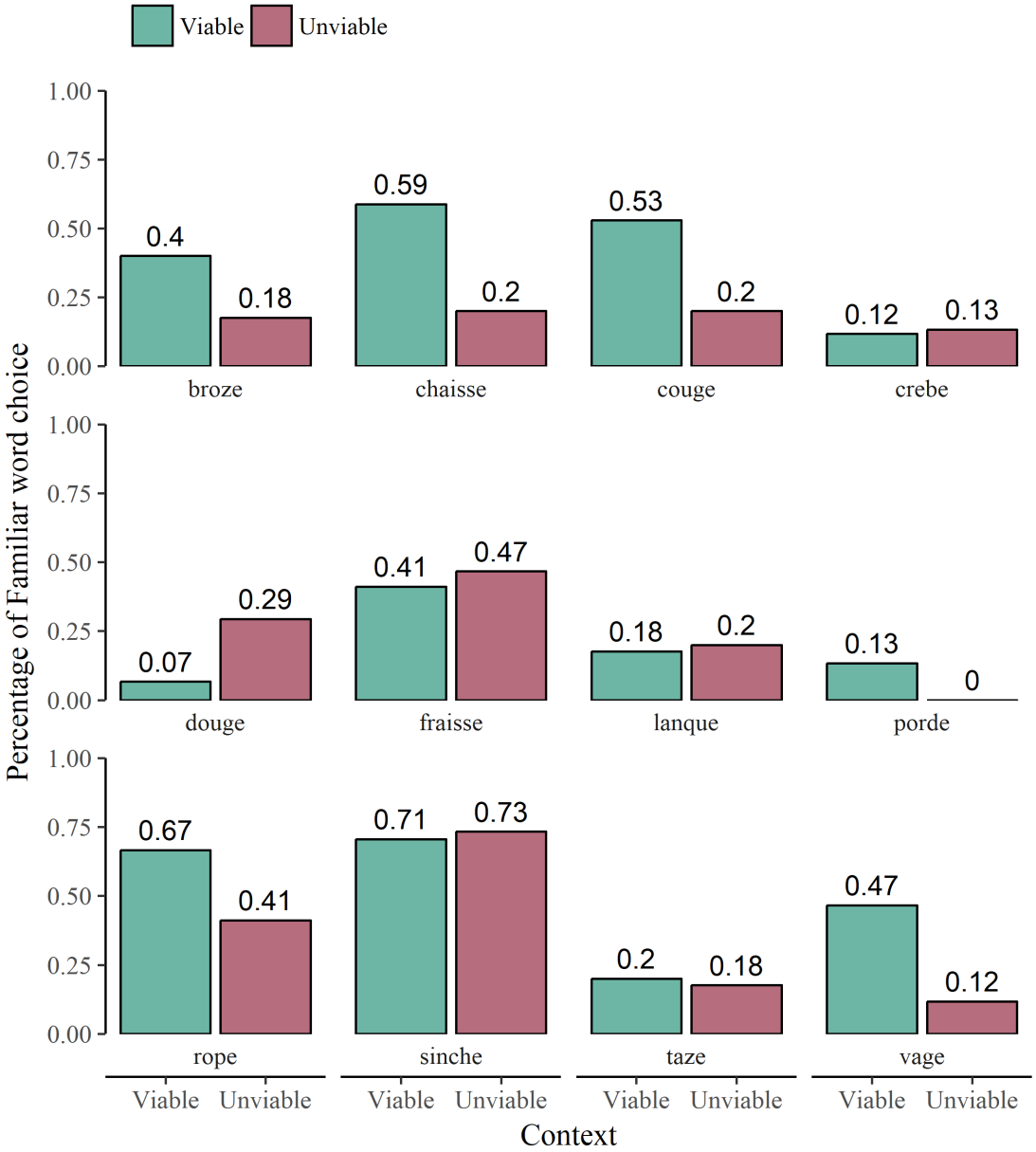


Figure 4.3: Barplots showing the percentage of clicks on the familiar object per item and context in voicing trials.

Furthermore, a comparison of the distribution of familiar object choices for the items in both counterbalancing lists showed a slight imbalance, by which one of the lists had overall higher familiarity preferences in *voicing viable* items than the other (mean % of familiarity preference, list A: 42%, list B:

32%). An imbalance was also present to a much smaller degree in the *place* items (overall, *place* items produced lower percentages of familiar object choices regardless of context, so the range of observed values is narrower and may not affect the overall results as much).

Finally, we noticed two additional small problems in this first pilot: First, while the pictures chosen to represent the unfamiliar items were not expected to be known by children, it often happened that they interpreted the image as being an object that they already knew, or sometimes they rejected the picture as a possible existing object. While the label was explicitly given to prevent them from incorrectly guessing the name of the novel object, some children made a statement about it, correcting the name: "no, that's a #!". Second, we noticed great difficulty during training in items where the contrast was the place of articulation change from /t/ to /p/ (for instance, *frites* [fʁit] "fries" to *fripes* [fʁip]), even though words were always produced in sentence-final position.

Based on these analyses, we decided to implement the following changes and run a second pilot:

- Increase the age-range by half a year, i.e., test children aged 4.5 to 5.5 years old.
- Swap one voicing item and one place item from each list to improve balance across lists.
- Make small edits to the length of final VC segments in items that showed poor performance to improve the ambiguity of the target word.
- Enhance the release of /t/ and /p/ in training items.

4.2.2 Pilot experiment 2

4.2.2.1 Methods

Participants

Twenty-one French monolingual 4.5 to 5.5-year-old children participated in the second pilot study (10 girls, mean age: 61.88 months, age range: 54.24 - 65.63 months). An additional 3 children participated but were excluded from analysis due to failing the training criteria (same criteria as used in Pilot 1).

Materials

The list of training and test items, the request phrases and the pictures used in this pilot were the same as described in Pilot 1. However, the counterbalancing lists differed from the ones used in Pilot 1 in two aspects: First, to improve the balance between the two lists based on our first pilot, the viable/unviable conditions of two voicing pairs (*brosse* "brush" - *broze*, & *fraise* "strawberry" - *fraisse*) and two place pairs (*crotte* "poop" - *crope*, & *pâtes* "pasta" - *papes*) were swapped. Second, after the first 9 children were tested, we decided to change the proportion of voicing viable items in the first block, such that both lists would have the same proportion (see *Results and discussion* section below for a discussion on why we made this decision). Thus, for the last 12 children in this pilot, the trial order was slightly different from that of Pilot 1 and from the first 9 children in Pilot 2.

While the recordings were the same as used in Pilot 1, small editing was done to the length of final VC segments of the following test items: *crebe, douge, fraisse, lanque, porde, taze* (in viable contexts) and *lanque, sinche* (in unviable contexts). Finally, the release of the final consonants /t/ and /p/ were enhanced for the following training items: *couette - coueppe, frites - fripes, tête - têpe.*

Procedure

The procedure was the same as in Pilot 1.

4.2.2.2 Results and discussion

We first looked separately at the two groups of children, before ($n = 9$) and after ($n = 12$) the change in the proportion of *voicing viable* items. The first group saw the same ratio of viable to unviable trials as in Pilot 1. From this first group we noticed that children who saw only one voicing viable item in the first block had overall less clicks on the familiar object throughout the whole experiment (15%) -with 4 out of 6 children having chosen the familiar object only twice out of 24 trials - compared to children who saw two voicing viable items in the first block (28%). One potential explanation for this is that children may form strong response biases early on in the game. If the first block contains only one trial that could potentially be interpreted as the familiar object, children may quickly create the belief that the correct answer is always the novel object. This belief may thus be sustained throughout the game. While the sample size is too small to reach a conclusion, we decided halfway through testing

to rearrange the proportions of the trials such that both lists would begin with 2 voicing viable and 1 voicing unviable trial (and the inverse proportion for place trials). As before, these ratios alternated from one block to the other. After the change, the children tested with the list that was rearranged clicked on the familiar object on 22% of the trials, on similar proportions on the first (24%) and second half (21%) of the experiment.

Pooled data from the two groups was analyzed with the same generalized linear mixed model as in Pilot 1. In addition to rule and context, we included group (*before* or *after* the change) in a triple interaction with rule and context. Furthermore, to control for a potential confound of trial order due to the rearrangement of the blocks mentioned in the *Materials* section, we included a simple effect of block. As no effect of group was found, neither as a simple effect nor in interaction with rule or context, we removed it from the model in subsequent analyses to improve the statistical power of the model. The random effects structure was the same as declared in Pilot 1. Figure 4.4 shows the proportion of clicks on the familiar object (both groups pooled together), split by condition.

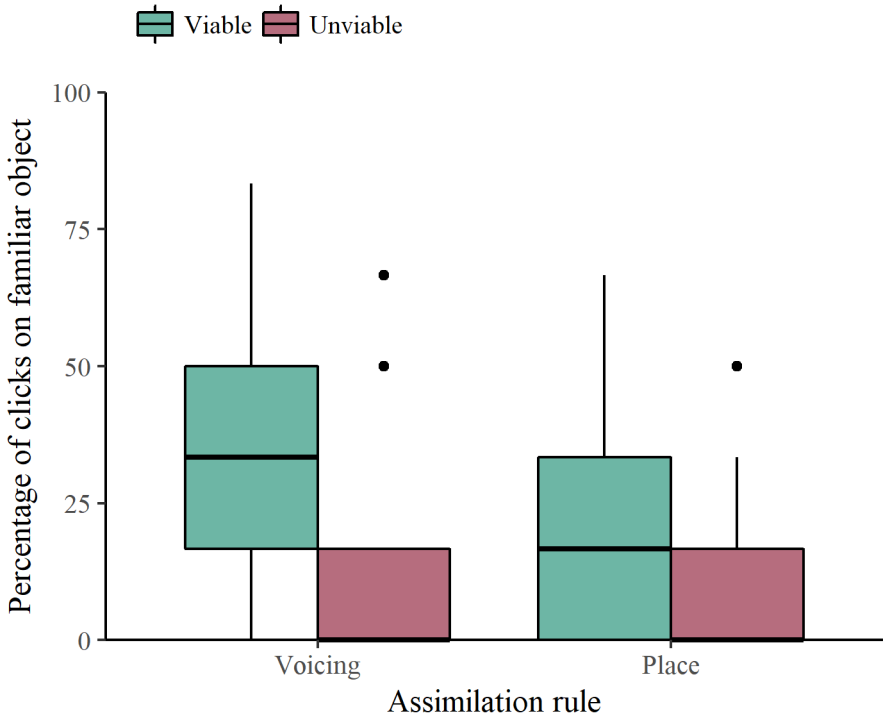


Figure 4.4: Boxplots showing the percentage of clicks on the familiar object per rule and context in Pilot 2.

The estimated intercept ($\beta_0 = -0.59$, $SE = 0.41$, $p > 0.1$) shows that the percentage of trials where the familiar object was chosen in the *voicing viable* condition was not significantly different from chance. A significant effect of context ($\beta = -2.58$, $SE = 0.86$, $p = 0.003$) indicates that children

clicked on the familiar object more often upon hearing voicing assimilation in a viable context than in an unviable context. An effect of rule ($\beta = -1.05$, $SE = 0.51$, $p = 0.04$) indicates that children chose the familiar object more often in *voicing viable* than in *place viable* trials. No effect of block ($\beta = -0.08$, $SE = 0.15$, $p = 0.63$) and no interaction of rule by context ($\beta = 1.03$, $SE = 1.02$, $p = 0.31$) were found. A releveling of the *rule* variable setting *place* as baseline revealed a marginal effect of context in place assimilation trials ($\beta = -1.55$, $SE = 0.82$, $p = 0.06$).

These results show a clearer compensation effect for voicing assimilation in comparison to the results in Pilot 1. However, the estimated difference between viable and unviable trials was approximately 31%, corresponding to a difference of only 2 more clicks on the familiar object for *voicing viable* trials. This difference might still not be large enough to eventually measure differences between monolinguals and bilinguals. These rather mild results -which we already observed in Pilot 1- might be a consequence of the design of the game: while children are not likely to choose the familiar object when hearing an unviable context, *voicing viable* contexts are in fact ambiguous, as either interpretation (that is, that the target is indeed the assimilated form of the familiar word, or alternatively that it is the name of the novel object) are equally plausible. Thus, children may be responding at chance in this condition, giving us a rather short range of possible outcomes. This small effect is in line with previous results with toddlers (Skoruppa, Mani & Peperkamp, 2013; Skoruppa, Mani, Plunkett, et al., 2013). However, as we need a stronger effect, a reconsideration of the paradigm might be necessary.

Furthermore, although an effect of rule was found (indicating that children chose the familiar object more often in voicing than in place trials), the interaction of rule and context was still not significant due to a marginal effect of context for place assimilation trials. While compensation for a non-native rule had never been reported in children, small non-native compensation effects have been observed in adults (Darcy et al., 2009). However, in Darcy et al. (2009)'s study, the interaction between rule and context was significant in spite of the presence of an effect of context for the non-native rule. Thus, to test whether these results were caused by a problem with the paradigm or otherwise due to children still being at an intermediate stage in their perceptual development, we conducted the following pilot with adult participants.

4.2.3 Pilot experiment 3

4.2.3.1 Methods

Participants

Thirteen French monolingual adults (age range: 19 - 30 years) participated in the third pilot study. They received a small compensation (2€) for their participation.

Materials

All materials were the same as used with the second group of children in Pilot 2.

Procedure

All procedures were the same as in Pilot 2, except that no pauses were made after completing each block.

4.2.3.2 Results and discussion

Responses were analyzed using the same generalized linear mixed model as in Pilot 1. However, the model with the maximal random effects structure failed to converge. Random effects were pruned until the model converged. The final random effects structure contained only intercepts for subject and item, and a slope for rule by subject. Figure 4.5 shows the percentage of clicks on the familiar object, split by condition.

The estimated intercept was significantly below chance ($\beta_0 = -2.11$, $SE = 0.56$, $p = 0.0002$), corresponding to a 10% chance of clicking on the familiar object in *voicing viable* trials. No effect of context ($\beta = -20.80$, $SE = 268.04$, $p = 0.94$) or rule ($\beta = -4.97$, $SE = 3.83$, $p = 0.19$), nor an interaction of rule by context ($\beta = 3.81$, $SE = 251.78$, $p = 0.99$) were found.

While 7 out of 13 participants did choose the familiar object at least once in the voicing viable condition, overall adults chose almost exclusively the novel object (with 6 of them never clicking on the familiar object at all during the test). This means that when given an alternative label, adults

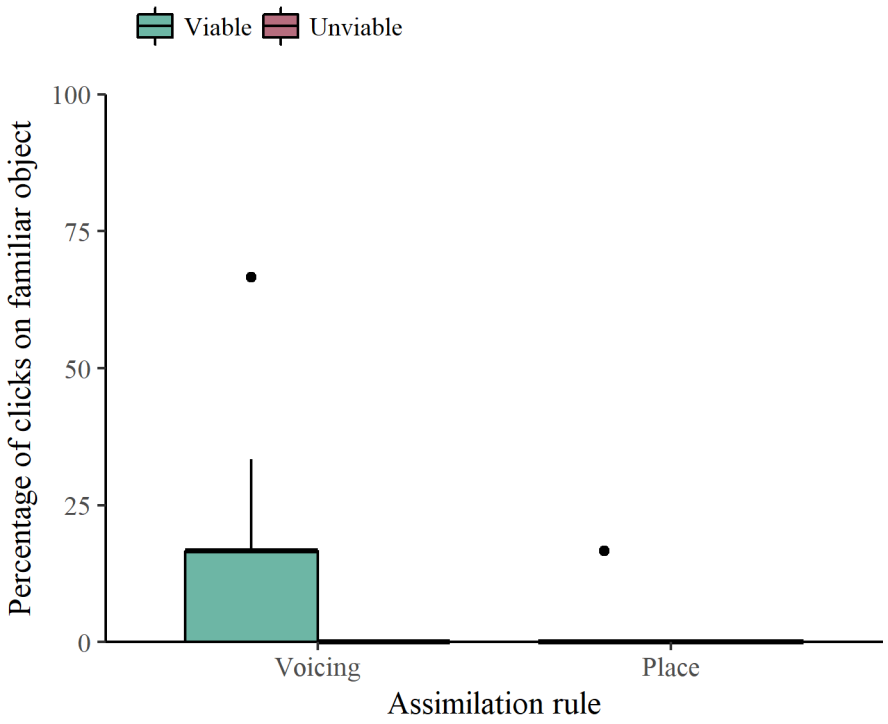


Figure 4.5: Boxplots showing the percentage of clicks on the familiar object per rule and context in Pilot 3.

unsurprisingly succeeded in listening for the specific contrast, and thus, as the target words were always pronounced with a voicing or a place change, they consistently chose the novel object. In Darcy et al. (2009), this problem did not emerge, as participants only heard the label of the familiar object and not the alternative form.

Overall, these results show a clear problem with the current paradigm. While having explicitly named objects may have worked with toddlers, older children may be better at detecting these contrasts, as adults do, thus showing a novelty preference and weak effects. These mild results are not a problem when testing for simple effects in separate groups, as was done in Skoruppa, Mani and Peperkamp (2013) and Skoruppa, Mani, Plunkett, et al. (2013), but it is not sufficient to test a 2 x 2 design. We therefore reconsidered the experimental design, and decided to remove the introduction phrases in a following pilot. As the items used in the current experiment are all well known to both children and adults, we expected the pictures of the familiar objects to be sufficient to activate the familiar word forms.

4.2.4 Pilot experiment 4

4.2.4.1 Methods

Participants

Fifteen French monolingual adults (age range: 18 - 26 years) participated in the third pilot study. They received a small compensation (2€) for their participation.

Materials

All materials were the same as used in Pilot 3. However, introduction phrases were removed.

Procedure

Procedures were globally the same as in Pilot 3, except for the removal of the introduction phrases. Note that, without the introductions, the game becomes in essence a mutual exclusivity task. The new sequence of events in each trial was as follows: first, the picture of the familiar object and the picture of the novel object appeared simultaneously, each on one side of the screen. Following a 3 second pause, the girl appeared in the middle of the screen and waited until the participant clicked on her. After clicking, the touch response became blocked while the character produced the request phrase. The touch response reactivated immediately afterwards, allowing the participant to give an answer by clicking either on the picture of the familiar or the novel object.

Before the experiment, participants were told that they would see pictures of familiar and novel items, and that their task was to listen to a phrase asking them to click on one of the two pictures, and finally decide which of the two pictures was asked for. As the images of the familiar objects would prime the familiar forms, and the alternative forms (which were not introduced) differed from the familiar word only on a subtle one-feature change in the final consonant, a strong familiarity bias was expected. To help reduce this bias, verbal feedback was given during training if the participant made a mistake, ensuring that the task was clearly understood before starting the test.

4.2.4.2 Results and discussion

Responses were analyzed using the same generalized linear mixed model as in Pilot 1, with the maximal random effects structure. Figure 4.6 shows the percentage of clicks on the familiar object, split per condition.

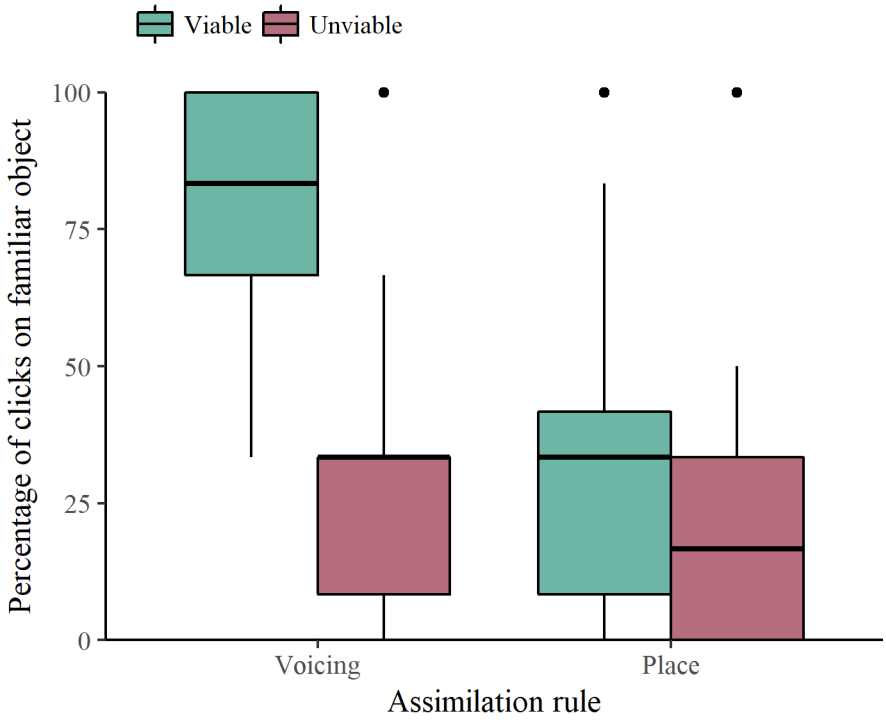


Figure 4.6: Boxplots showing the percentage of clicks on the familiar object per rule and context in Pilot 4.

The estimated intercept ($\beta_0 = 2.42$, $SE = 0.85$, $p = 0.004$) shows that the percentage of trials where the familiar object was chosen in the *voicing viable* condition was significantly above chance. We found a large effect of context ($\beta = -3.96$, $SE = 0.83$, $p < 0.0001$) for voicing assimilation trials. Furthermore, an effect of rule ($\beta = -3.85$, $SE = 0.91$, $p < 0.0001$) indicated a large significant difference between *voicing viable* and *place viable* trials. Finally, a positive interaction of rule by context ($\beta = 3.11$, $SE = 1.01$, $p = 0.002$) was found, suggesting a difference in the effect of context for voicing and place assimilation. A releveling of the *rule* variable setting *place* as baseline revealed no effect of context in place assimilation trials ($\beta = -0.85$, $SE = 0.72$, $p = 0.24$).

These results show a very clear compensation effect for voicing assimilation, with an estimated probability of 92% to choose the familiar object in voicing viable contexts against 18% in unviable contexts. The effect of rule and the interaction of rule by context replicate what was previously observed in

Darcy et al. (2009), showing a clear distinction in the way adults treat native and non-native rules.

Overall, these findings confirm that the problem with the previous task was the introduction of the two labels, which produced an ambiguity effect in the voicing viable trials, and also caused participants to become aware of the target contrast, thus often doing the task at an acoustic level and producing a strong novelty preference. While this new paradigm works well with adults, it may be difficult for young children to perform a mutual exclusivity task with a one-feature change in sentence-medial position. We thus conducted a small pilot with slightly older children to evaluate the feasibility of this paradigm.

4.2.5 Pilot experiment 5

4.2.5.1 Methods

Participants

Nine French monolingual children (4 girls, mean age: 69.60 months / 5.8 years, age range: 63.59 - 74.76 months) participated in the fifth pilot study. An additional 3 children were tested but excluded from analysis due to failure to pass the training criterion (same *Exclusion criteria* as in Pilots 1 and 2).

Materials

All materials were the same as used in Pilot 4. However, two training pairs were removed (*tête* "head" - *têpe*, and *frites* "fries" - *fripes*) due to the great difficulty observed in Pilots 1 and 2 in perceiving their final assimilated /p/, which caused children to misunderstand the task after being corrected several times without having perceived the contrast. Two other pairs containing the same contrast (*couette* "duvet" - *coueppe*, and *plante* "plant" - *plampe*) were kept in the training to ensure that children would have seen this contrast before moving on to the test. The final training list thus contained only 6 items.

Procedure

All procedures were the same as in Pilot 4, including the absence of a pause after each block. Unlike in Pilots 1 and 2, here we told the children they would win a sticker at the end of the game if they earned all 4 stars.

4.2.5.2 Results and discussion

Responses were analyzed with the same generalized linear mixed model as used in Pilot 1. Figure 4.7 shows the percentage of clicks on the familiar object, split per condition.

The model revealed a significant intercept above chance ($\beta_0 = 2.81$, $SE = 0.88$, $p = 0.001$), but no effect of context ($\beta = 0.89$, $SE = 2.17$, $p > 0.1$) or rule ($\beta = -1.17$, $SE = 0.84$, $p > 0.1$), nor an interaction of rule by context ($\beta = -0.84$, $SE = 1.91$, $p > 0.1$). Indeed, by looking at the pattern of responses in Figure 4.7, it is evident that children were unable to correctly play this mutual exclusivity game, as these word-final one-feature changes embedded in difficult contexts were too subtle to compensate for the strong familiarity bias. As adults can perform this task with no major difficulties, it is evident that increasing the age of the participants would probably solve this problem. However, as we would like to have a paradigm that works with preschoolers and that could eventually be used with younger children, this was not an option.

We therefore decided to revise the paradigm one last time. As children were able to detect the contrast when given the label of the novel object, but showed a strong familiarity bias in the mutual exclusivity task, a possible explanation is that they found it highly unlikely that a new object would be named almost exactly as the familiar object, and thus disregarded these one-feature changes as small deviations from the canonical pronunciation. Therefore, if we removed the novel object, children would not need to make assumptions about the likelihood of the label belonging to a new item. We finally decided to replace the image of the novel object with a cross, thus eliminating any assumptions about the labels, and transforming the task into a word detection task, similar to the one used by Darcy et al. (2009) with adults. We implemented this change in our final study (Carbajal et al., in revision), and obtained for the first time a significant interaction showing a difference in the perception of native and non-native rules in children.

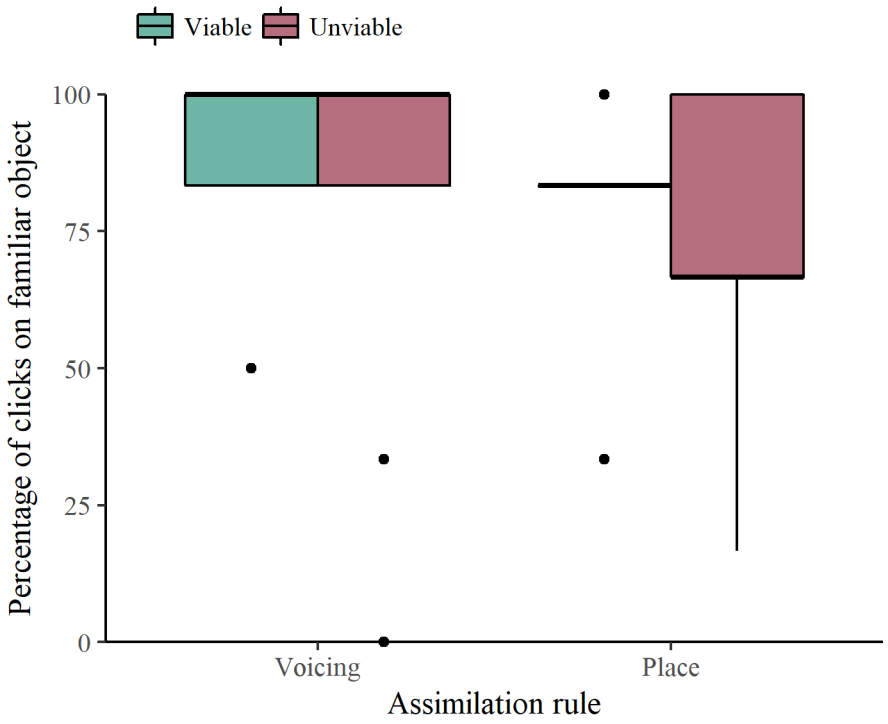


Figure 4.7: Boxplots showing the percentage of clicks on the familiar object per rule and context in Pilot 5.

4.2.6 General discussion

In a series of five pilot studies, we examined French monolingual children and adults' compensation for both a native voicing assimilation rule and a non-native place assimilation rule using an iPad videogame. We began with an experimental design inspired by Skoruppa, Mani and Peperkamp (2013), in which participants were first introduced to familiar and novel objects, and then heard a touching request phrase in which the assimilated form (that is, with a place or a voicing change in its final consonant) was followed by either a viable or an unviable context for assimilation. Using this paradigm, we found that overall children (aged between 4- and 5.5-year-old) and adults showed small or null effects of compensation for voicing assimilation (see summary in Table 4.1) and strong novelty preferences. Furthermore, we consistently failed to find a significant interaction of *rule* (voicing, place) and *context* (viable, unviable), indicating that our experiment was not suitable to test differences in the way listeners treat a native and a non-native rule.

After the first three pilots using Skoruppa, Mani and Peperkamp's (2013) design failed to show a significant interaction, we removed the introduction phrases - thus turning the experiment into a mutual exclusivity task - and tested adults again in Pilot 4. The removal of the introductions revealed

Table 4.1: Summary of findings in Pilots 1, 2 & 3 using Skoruppa et al.'s design.

	Compensation for voicing	Compensation for place	Interaction (rule x context)
Pilot 1 (4.0 - 5.0 y.o.)	✓	×	×
Pilot 2 (4.5 - 5.5 y.o.)	✓	(✓)	×
Pilot 3 (adults)	×	×	×

the expected compensation pattern, as adults compensated for voicing but not for place assimilation, with a significant interaction of rule by context indicating a difference in the way the two rules were perceived. Finally, in Pilot 5 we tested 5.5 to 6.5-year-olds on the same task and found that unfortunately by 6-years-old, children are still unable to deal with a difficult mutual exclusivity task, showing a strong familiarity bias and null effects for both voicing and place.

We finally removed the novel item and replaced it with a red cross, as described in Carbajal et al. (in revision), which yielded the expected interaction of rule by context in 6-year-old children. This experimental design is similar to the word detection task used with adults in Darcy et al. (2009), and allowed us to test both rules simultaneously with a clear distinction of the compensation effects for native and non-native rules, which was necessary to test bilingual children. In Figure 4.8 we show a comparison of the results from toddlers in Skoruppa, Mani and Peperkamp (2013) and from our five pilots as well as our final study[4].

The interesting patterns of results observed in this series of pilots suggest that while compensation for assimilation has been regarded as an automatic process (Sun et al. 2015, Fort et al. 2017), it is still dependent on the perceptual task, as listeners may treat their speech input differently depending on the dimension they are focusing on. Thus, if explicitly given the target contrast, both children and adults can inhibit their compensation - at least to a certain extent - in order to categorize the consonants and recognize the novel labels. Without the explicit contrast, children seem to be sensitive to assimilation (as we have shown in Carbajal et al., in revision), yet may disregard subtle one-feature changes -even in unviable contexts - as small variability in the signal if presented with a familiar object and a new object that is unlikely to have such a similar name, as happened in the mutual exclusivity task. These task-dependent aspects of assimilation should thus be taken into account when planning perceptual experiments, both with adults and children.

[4]Since we did not have access to the full dataset presented in Skoruppa, Mani & Peperkamp (2013) but we had the means and standard errors per condition as reported in the paper, we show barplots instead of boxplots for all 7 studies.

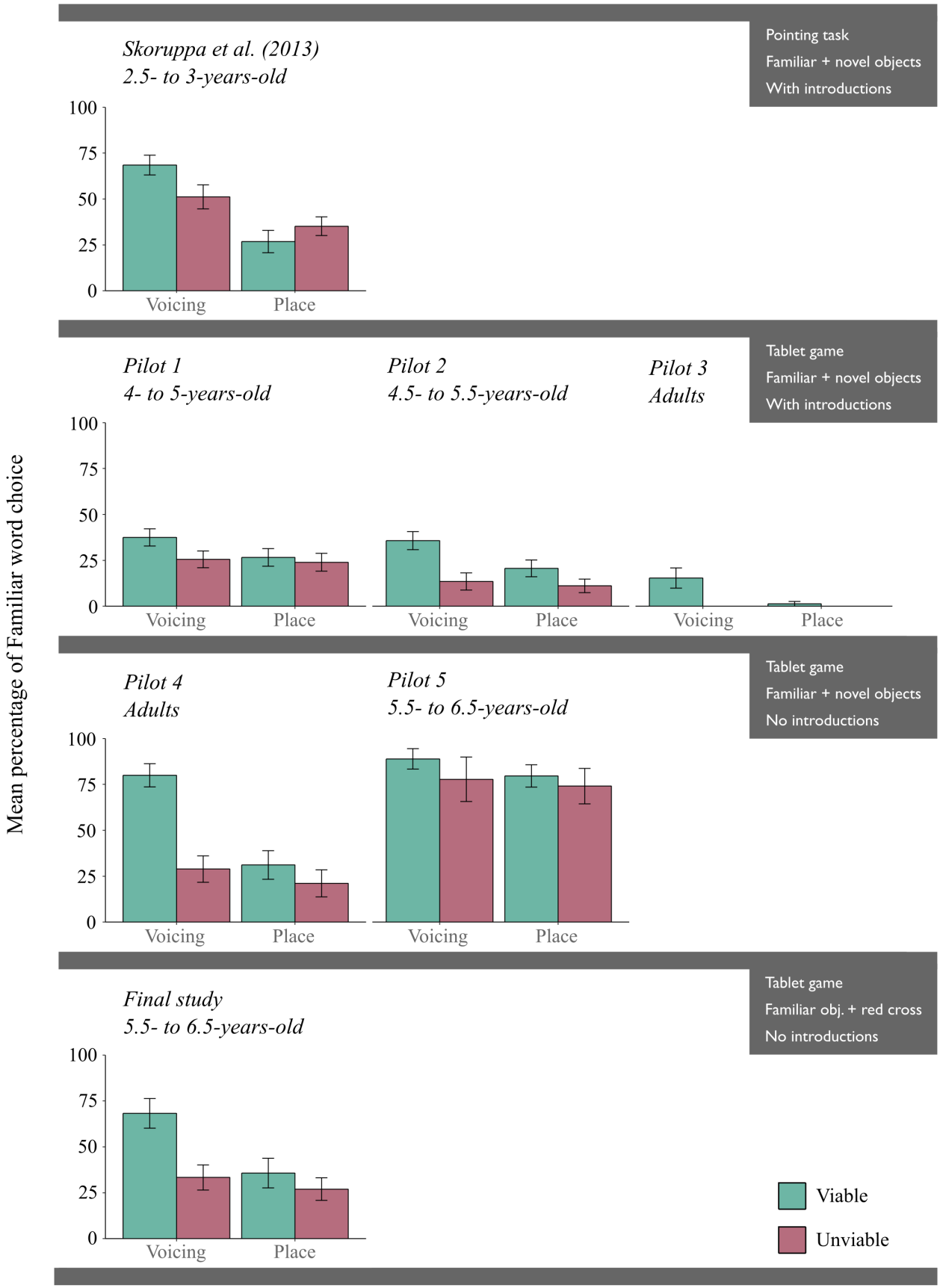


Figure 4.8: A comparison of results obtained in Skoruppa, Mani and Peperkamp (2013), Pilots 1-5 and the final study presented in Carbajal et al. (in revision). Barplots represent means and standard errors per condition.

Chapter 5

General discussion

In this thesis, we explored three different aspects of early bilingual development related to the problem of language separation: the discovery of two languages in the input, the potential role of environmental separation of the input on early lexical development, and the perceptual separation of phonological rules. In this chapter we will summarise our findings from these three research projects, discussing future directions of research. We will separate the discussion in two sections; the first one, regarding our findings from Chapter 2 and 3, concerns issues on language separation in the first year of life; the second one, based on our findings from Chapter 4, concerns later separation abilities.

5.1 Language separation in the first year of life

Chapters 2 and 3 in perspective

The goal of our first study (Chapter 2) was to revive the discussion on how young bilinguals come to discover the presence of two languages in the speech signal. Based on a large body of research on language discrimination abilities in newborns and infants, and inspired by previous attempts at modelling this perceptual skill, we aimed to provide a computational account that could explain infant perception of multilingual speech. While previous models of language discrimination had focused on rhythmic properties of speech (e.g., Ramus et al., 1999, Dominey & Ramus, 2000), we argued that this type of representation is insufficient to describe the full spectrum of results that has been observed in infants. As an alternative to previous models, we borrowed a state-of-the-art system from speech technologies, the i-vector model, which allowed us to incorporate in a single representation both

prosodic and phonetic properties of speech.

In a series of six computational experiments, we explored the behavior of the i-vector model when confronted with multilingual audio recordings. Our results showed overall good agreement with previous empirical evidence, confirming a) that distant language pairs (such as French-English) are easier to discriminate than close language pairs (such as Spanish-Catalan and English-Dutch), b) that these discrimination patterns remain generally unchanged when presented with filtered speech, and c) that different (monolingual) language backgrounds result in similar discriminability skills. An interesting finding from these experiments is that, in general, between-speaker distance in our model's representational space is equal to or larger than between-language distance. This type of variability may pose a big challenge for the dual language learner. An empirical illustration of this is the fact that newborns fail at discriminating a distant language pair (Dutch-Japanese) when presented with natural speech from multiple speakers, while they succeed with reduced speaker variability (Ramus, 2002). Going further than previous models, our i-vector system reproduced for the first time these two different responses to the same language pair, thus emphasizing the importance of considering models that can integrate multiple sources of variability in one representation.

In our last computational experiment, we began to explore the potential consequences of blindly accumulating speech samples from two languages, something that bilingual infants with close language pairs (such as Spanish and Catalan) might accidentally do at the beginning of life. Our results suggested that this lack of separation could lead to some initial stage of confusion, but that attention to additional information (in this case, speaker identity perfectly correlating with language) could help to overcome this problem. The idea that young bilinguals may need other linguistic and extralinguistic cues to help in the task of language separation is not new. For instance, it has been shown that bilingual infants can discriminate languages through facial gestures (Weikum et al., 2007, Sebastián-Gallés et al., 2012), and some studies have suggested a negative impact of parental language mixing on lexical acquisition (e.g., Byers-Heinlein, 2013). However, information on the effects of these factors in language separation remains scarce. It is thus worth further investigating if and how young children integrate these additional cues with their acoustic representations to achieve separation, and whether this has any impact on later language development.

In Chapter 3, we investigated several environmental aspects of bilingual exposure in 11-month-olds learning French and an additional language. Previous research on quantitative and qualitative properties of bilingual input had shown strong influences of each language's relative amount of exposure

on infants' lexical development, but diverging results were reported regarding the impact of the separation of the two languages in their environment (Byers-Heinlein, 2013; Place & Hoff 2011, 2016; De Houwer, 2007). We discussed that these contradicting observations could be caused by age differences, as extra-linguistic cues to language separation might be relevant only during the first two years of life. Following this line of research, and in the light of our computational findings regarding the role of by-speaker language separation, we used a Language Diary method to investigate the co-existence of two languages in young bilinguals' input.

Based on diary records, we computed several measures quantifying the amount of exposure to each language, as well as the degree of overlap of both languages in time and by speaker. First, we observed that despite most families generally adopting a "one-parent one-language" approach, infants varied in the frequency of co-occurrence of the two languages in their environment, both in time (regardless of speaker) and within speaker. We then investigated potential effects of these and other environmental factors on infants' lexical development, by conducting exploratory correlational analyses on their vocabulary scores. An interesting observation stemming from these analyses was that by-speaker language separation had a positive correlation with comprehension vocabulary scores in both languages, in line with previous results on 1.5-year-olds by Byers-Heinlein (2013). As this was an exploratory study, these results would need a confirmatory replication. However, if confirmed, they would suggest that additional cues to language separation are indeed relevant during the first one or two years of life. It is in fact not unreasonable to think that after a certain age, all bilinguals will have figured out the presence of two languages in their environment, thus making the task of tagging their input and learning from it less challenging.

Finally, we made an unexpected observation in our heterogeneous sample: infants who were learning two languages from the same family (that is, French plus a Romance language) had overall higher vocabulary scores than those learning two distant languages, particularly French plus a Germanic language. While we had not planned for this analysis, this finding seems to align with very recent data from a large bilingual cohort studied by Floccia and colleagues (2018), in which infants learning language pairs with close phonological, morphological or word typology properties had larger vocabularies than those learning distant languages. This might suggest that, while distant languages may be easier to discriminate, this might not always translate into an advantage in language learning. Indeed, close language pairs may share properties that are crucial to vocabulary acquisition, such as prosody, which is known to be used in word segmentation (Curtin et al., 2005). Furthermore, they may share a

larger number of cognates, thus reducing the number of approximate word forms that the child needs to memorize. Thus, unlike phonological development, where close language pairs seem to pose more problems than distant ones in the acquisition of phonetic categories (Bosch & Sebastián-Gallés, 2003b; Sundara & Scutellaro, 2011), lexical acquisition might show the inverse pattern.

Another interesting difference between the language pairs that we studied is that, when comparing the effect of by-speaker language separation across groups, we observed an effect in the French + Romance group but not in the French + Germanic group. This might reflect that infants learning distant languages do not need any additional cues to notice the presence of two languages, and thus might not be affected by this factor. Due to the small sample size in each language group, it is not possible to conclude whether this difference is actually meaningful until more data are gathered. However, it hints at a possible interaction between the various linguistic challenges an infant faces during the first year of life, which are likely dependent on the specific language pair. This emphasizes the importance in early bilingualism research of comparing an effect across several language pairs, as what is true for one group may not be true for the other. Future research on similar topics should thus aim at systematically assessing children from different linguistic backgrounds on the same tasks or measures, to begin forming a broader picture of what is general to all language pairs, and what is specific to individual language pairs.

A natural question that begins to emerge is whether early language discrimination is actually necessary to bilingual language acquisition. Given the difficulty of the task of language sorting (as evidenced from infants' failure in some language discrimination tasks, as well as from our computational work), it could be expected that infants produce frequent errors in their tagging. Would a failure to correctly separate the speech samples into pure language clusters affect their general learning mechanisms? This kind of question is difficult to answer empirically, as it is perhaps impossible to know how infants have clustered their input in real life, and behavioral methods, such as artificial language learning, may or may not reflect what infants do with their actual input. We consider that in this situation, computational models could offer a useful insight into what is possible to learn from mixed input. One way to do this is to feed the output of the i-vector classification (with its imperfect classification) to models of phonological development or spoken word discovery. It may be the case that a failure to find the correct language clusters does not necessarily result in a completely merged phonology or a nonsense vocabulary. We thus hope that in the near future it will be possible to combine these kinds of computational cognitive algorithms (which currently exist mainly as separate modules) into one big

pipeline to examine the whole process of language learning in bilingual infants, from their very first speech representations to the acquisition of words.

5.2 Separation of phonological rules

Summary of Chapter 4

In the last chapter of this thesis we investigated bilingual preschoolers' perception of language-specific phonological rules. Unlike other properties of young bilinguals' phonological systems, their acquisition and separation of phonological rules had barely been explored, with the only prior evidence coming from production studies. Here, we investigated the perception of phonological assimilation, a rule by which a consonant at a word edge adopts a phonological feature of a neighboring consonant. We argued that this type of rule may pose an interesting challenge to the bilingual learner for two reasons: first, bilingual children may be exposed to speech with more phonetic variability than what a typical monolingual encounters. Due to this experience, bilinguals may be more flexible regarding mispronunciations, and thus learning a subtle, context-specific rule might be a difficult task. The other problem that bilinguals face is the language-specificity of these rules, as phonological assimilation affects different consonants and features depending on the language. Keeping track of the alterations that occur in one or the other language may be challenging. To begin investigating these questions, we designed a touchpad videogame to test French-English bilingual children's compensation for voicing and place assimilation rules, the first of which exists in French but not in English, and vice-versa for the other one.

In this first study we assessed both rules in French sentences. Our results showed that, like French monolinguals of the same age, bilinguals compensated for voicing but not for place assimilation when hearing French sentences, and they did so in a context-specific manner. These results show that bilingual preschoolers have good perceptual knowledge of a subtle phonological rule, which they correctly interpret in its corresponding language. As we so far only tested them on French sentences, we cannot conclude from these results that their knowledge is indeed language-specific. These bilinguals, who were being raised in a French-speaking country, might have better mastery of their French phonology than their English phonology. In a future study which we have already begun to prepare, we will test bilinguals on these two same rules, applied on English sentences. By having the full square design, crossing rules and languages, we will be able to get a full picture of the specificity of their perception

of these phonological rules.

An important point to raise regarding these results, is the fact that French and English are two distant languages, which differ indeed in many linguistic properties (such as their phonologies, their rhythmic patterns, and their syntactic structures). Thus, in the context of our previous discussion on language distance and its impact on phonological and lexical development, it might be expected that French-English bilinguals have no trouble tracking their phonological systems with little cross-linguistic interference. In future work, it would be interesting to test similar language-specific phonological rules in a more closely related language pair, such as Spanish and Catalan, which might reveal a more complex scenario.

In conclusion, this experiment, a first of its kind in this population, opens the door to further work on phonological processing in bilinguals beyond the second year of life, an area of research that has been so far largely unexplored. To finalise this discussion, we would like to add a few comments on the methodological effort that was involved in designing this experiment. As any experimentalist working with bilingual children will know, designing a behavioral experiment for this kind of population is a long and arduous task. Beyond the well-known difficulties in finding good, sensitive methods to test the perception of young children in general, bilinguals come with an additional level of complexity, as their prior linguistic knowledge (which is often needed to plan an experiment) may be hard to estimate. Furthermore, bilinguals' responses may differ from that of monolinguals in many predictable and unpredictable ways. As this population is difficult to recruit, the traditional trial and error process that one could use to fine tune an experiment is hardly feasible with bilinguals. Instead, in this study, we conducted a total of 5 pilot studies with monolingual children and adults, testing a total of 113 participants before we were convinced that our method was appropriate for the age, and sufficiently sensitive to test the interaction we were after. This was to us no wasted time, as in the process we learnt many valuable things about how children (and adults too) may interpret the very same acoustic signal when faced with different scenarios. Following this idea, we made a report of our pilot studies freely available on our OSF project page. Reporting pilot studies is rarely done, but we would like to encourage more researchers to make this information widely available, as we often learn more from our mistakes than from our success. This leads us to the last point, which is the importance of deciding beforehand what we will consider a success. As bilingual infants and children may respond in many different ways to the experimental stimuli, it is all the more important to make a clear analysis plan before testing. After our immense effort to fine tune our experimental method, we pre-registered the

procedure, as well as the rejection criteria, the number of participants, and the full analysis. Thanks to this effort done prior to testing, we can rest assured that our results reflect an honest analysis of the data, something invaluable when testing a population that is so difficult to recruit. As times are quickly changing, we expect more and more developmental studies to go this way.

5.3 Conclusion

In this thesis we explored different aspects of language separation in young bilingual children, using a combination of three research methods: computational modelling, home diary records, and behavioral experiments. Each method allowed us to have a new perspective on one specific aspect of the problem, and combined, they provided several insight into the challenges that infants face when discovering and learning language from dual input. Our research questions covered different stages of bilingual development, from the very first steps in speech perception to the acquisition of complex phonological rules. While our work offered new perspectives on this problem and contributed additional evidence on how infants from different backgrounds develop their lexical and phonological systems, an immense amount of work remains to be done. The field of bilingual language acquisition is, much like the infants it studies, just beginning to discover a complex world.

Bibliography

Albareda-Castellot, B., Pons, F., & Sebastián-Gallés, N. (2011). The acquisition of phonetic categories in bilingual infants: New data from an anticipatory eye movement paradigm. *Developmental science*, *14*(2), 395–401.

Allen, G. D., & Hawkins, S. (1980). Phonological rhythm: Definition and development. In *Child phonology* (pp. 227–256). Elsevier.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351–373.

Bahrick, L. E., & Pickens, J. N. (1988). Classification of bimodal English and Spanish language passages by infants. *Infant Behavior and Development*, *11*(3), 277–296.

Barron-Hauwaert, S. (2004). *Language strategies for bilingual families: The one-parent-one-language approach*. Multilingual Matters.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID - Corpus of Interactional Data - Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, *49*(3), 1–30.

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, *13*(4), 525–531.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer [Computer program]. Version 5.3.86. Version 5.3.86. Retrieved from http://www.praat.org

Bosch, L., & Ramon-Casas, M. (2014). First translation equivalents in bilingual toddlers' expressive vocabulary: Does form similarity matter? *International Journal of Behavioral Development*, *38*(4), 317–322.

Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*(1), 33–69.

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy*, *2*(1), 29–49.

Bosch, L., & Sebastián-Gallés, N. (2003a). Language experience and the perception of a voicing contrast in fricatives: Infant and adult data. In *Proceedings of the 15th international conference of phonetic sciences* (pp. 1987–1990). UAB/Casual Prods.

Bosch, L., & Sebastián-Gallés, N. (2003b). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and speech*, *46*(2-3), 217–243.

Bosch, L., & Sebastián-Gallés, N. (2005). Developmental changes in the discrimination of vowel contrasts in bilingual infants. In *Proceedings of the 4th international symposium on bilingualism* (pp. 354–363). Cascadilla Press Somerville, MA.

Bridges, K., & Hoff, E. (2014). Older sibling influences on the language environment and language development of toddlers in bilingual homes. *Applied psycholinguistics*, *35*(2), 225–241.

Burns, T. C., Yoshida, K. A., Hill, K., & Werker, J. F. (2007). The development of phonetic representation in bilingual and monolingual infants. *Applied Psycholinguistics*, *28*(3), 455–474.

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition*, *16*(1), 32–48.

Byers-Heinlein, K., Burns, T. C., & Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychological science*, *21*(3), 343–348.

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: Infants' language experience influences the development of a word-learning heuristic. *Developmental science*, *12*(5), 815–823.

Campbell, W. M., Singer, E., Torres-Carrasquillo, P. A., & Reynolds, D. A. (2004). Language recognition with support vector machines. In *Odyssey 2004 - The Speaker and Language Recognition Workshop*.

Carbajal, M. J., Fér, R., & Dupoux, E. (2016). Modeling language discrimination in infants using i-vector representations. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 889–894).

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., & Vair, C. (2007). Acoustic language identification using fast discriminative training. In *Eighth Annual Conference of the International Speech Communication Association*.

Cattani, A., Abbot-Smith, K., Farag, R., Krott, A., Arreckx, F., Dennis, I., & Floccia, C. (2014). How much exposure to english is necessary for a bilingual toddler to perform like a monolingual peer in language tests? *International Journal of Language & Communication Disorders*, *49*(6), 649–671.

Chevrot, J.-P., Dugua, C., & Fayol, M. (2009). Liaison acquisition, word segmentation and construction in French: a usage-based account. *Journal of child language*, *36*(3), 557–596.

Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2-month-olds. *Developmental Science*, *1*(2), 215–219.

Conboy, B. T., & Thal, D. J. (2006). Ties between the lexicon and grammar: Cross-sectional and longitudinal studies of bilingual toddlers. *Child development*, *77*(3), 712–735.

Curtin, S., Byers-Heinlein, K., & Werker, J. F. (2011). Bilingual beginnings as a lens for theory development: PRIMIR in focus. *Journal of Phonetics*, *39*(4), 492–504.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition*, *96*(3), 233–262.

David, A., & Wei, L. (2008). Individual differences in the lexical development of French–English bilingual children. *International Journal of Bilingual Education and Bilingualism*, *11*(5), 598–618.

Davidson, D., & Tell, D. (2005). Monolingual and bilingual children's use of mutual exclusivity in the naming of whole objects. *Journal of experimental child psychology*, *92*(1), 25–45.

Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *Readings in speech recognition* (pp. 65–74). Elsevier.

De Houwer, A. (2007). Parental language input patterns and children's bilingual use. *Applied psycholinguistics*, *28*(3), 411–424.

De Houwer, A. (2009). *Bilingual first language acquisition*. Multilingual Matters.

De Houwer, A. (2011). Language input environments and language development in bilingual acquisition. *Applied Linguistics Review*, *2*, 221–240.

De Houwer, A. (2014). The absolute frequency of maternal input to bilingual and monolingual children. *Input and experience in bilingual development*, *13*, 37–58.

De Houwer, A. (2018). The role of language input environments for language outcomes and language acquisition in young bilingual children. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language: The state of the science across its subfields*. John Benjamins Publishing Company.

De Houwer, A., & Bornstein, M. (2003). Balancing on the tightrope: Language use patterns in bilingual families with young children. In *4th International Symposium on Bilingualism, Tempe, AZ*.

De Houwer, A., & Bornstein, M. H. (2016). Bilingual mothers' language choice in child-directed speech: Continuity and change. *Journal of multilingual and multicultural development*, *37*(7), 680–693.

De Houwer, A., Bornstein, M. H., & De Coster, S. (2006). Early understanding of two words for the same thing: A CDI study of lexical comprehension in infant bilinguals. *International Journal of Bilingualism*, *10*(3), 331–347.

De Houwer, A., Bornstein, M. H., & Putnick, D. L. (2014). A bilingual–monolingual comparison of young children's vocabulary size: Evidence from comprehension and production. *Applied Psycholinguistics*, *35*(6), 1189–1211.

DeAnda, S., Bosch, L., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2016). The language exposure assessment tool: Quantifying language exposure in infants and children. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1346–1356.

Dehaene-Lambertz, G., & Houston, D. (1998). Faster orientation latencies toward native language in two-month-old infants. *Language and Speech*, *41*(1), 21–43.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association*.

Deuchar, M., & Clark, A. (1996). Early bilingual acquisition of the voicing contrast in English and Spanish. *Journal of Phonetics*, *24*(3), 351–365.

Dominey, P. F., & Ramus, F. (2000). Neural network processing of natural language: I. sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, *15*(1), 87–127.

Döpke, S. (2000). Generation of and retraction from cross-linguistically motivated structures in bilingual first language acquisition. *Bilingualism: Language and Cognition*, *3*(3), 209–226.

Duanmu, S. (1994). Syllabic weight and syllabic duration: A correlation between phonology and phonetics. *Phonology*, *11*(1), 1–24.

Duez, D. (2006). Syllable structure, syllable duration and final lengthening in Parkinsonian French speech. *Journal of Multilingual Communication Disorders*, *4*(1), 45–57.

Dunn, L. M., Dunn, D. M., Lenhard, A., Lenhard, W., & Suggate, S. (2015). *PPVT-4: Peabody picture vocabulary test;[manual]*. Pearson.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Eilers, R. E., Oller, D. K., Levine, S., Basinger, D., Lynch, M. P., & Urbano, R. (1993). The role of prematurity and socioeconomic status in the onset of canonical babbling in infants. *Infant Behavior and Development*, *16*, 297–315.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Fabiano-Smith, L., & Barlow, J. A. (2010). Interaction in bilingual phonological acquisition: Evidence from phonetic inventories. *International journal of bilingual education and bilingualism*, *13*(1), 81–97.

Fabiano-Smith, L., & Goldstein, B. A. (2010). Phonological acquisition in bilingual Spanish-English speaking children. *Journal of Speech, Language, and Hearing Research*, *53*(1), 160–178.

Fabiano-Smith, L., Oglivie, T., Maiefski, O., & Schertz, J. (2015). Acquisition of the stop-spirant alternation in bilingual Mexican Spanish-English speaking children: Theoretical and clinical implications. *Clinical linguistics & phonetics*, *29*(1), 1–26.

Farinas, J., Pellegrino, F., Rouas, J.-L., & André-Obrecht, R. (2002). Merging segmental and rhythmic features for automatic language identification. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Vol. 1, pp. I–753). IEEE.

Fennell, C. T. (2005). Balanced and language-dominant bilinguals' perception of a consonant contrast in infancy. In *International Congress for the Study of Child Language, Berlin*.

Fennell, C. T., & Byers-Heinlein, K. (2014). You sound like mommy: Bilingual and monolingual infants learn words best from speakers typical of their language environments. *International Journal of Behavioral Development*, *38*(4), 309–316.

Fennell, C. T., Byers-Heinlein, K., & Werker, J. F. (2007). Using speech sounds to guide word learning: The case of bilingual infants. *Child development*, *78*(5), 1510–1525.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, i–185.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur communicative development inventories. *Applied Psycholinguistics*, *21*(1), 95–116.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science*, *16*(2), 234–248.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Bell Labs Technical Journal*, *12*(4), 377–430.

Fletcher, J., & McVeigh, A. (1993). Segment and syllable duration in australian english. *Speech Communication*, *13*(3-4), 355–365.

Floccia, C., Sambrook, T., Delle Luche, C., Kwok, R., Goslin, J., White, L., ... Krott, A., et al. (2018). Vocabulary of 2-year-olds learning English and an additional language: norms and effects of linguistic distance. *Monographs of the Society for Research in Child Development*, *83*(1).

Friend, M., & Keplinger, M. (2003). An infant-based assessment of early lexicon acquisition. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 302–309.

Frota, S., Butler, J., Correia, S., Severino, C., Vicente, S., & Vigário, M. (2015). Questionários MacArthur-Bates (CDI) para o Português Europeu: Formas reduzidas (8 aos 30 meses). *II Jornadas–Comunicação e Desenvolvimento da Linguagem*.

Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In *Speech Prosody 2002, Aix-en-Provence*.

Garcia-Sierra, A., Rivera-Gaxiola, M., Percaccio, C. R., Conboy, B. T., Romo, H., Klarman, L., ... Kuhl, P. K. (2011). Bilingual language learning: An ERP study relating early brain responses to speech, language input, and later word production. *Journal of Phonetics*, *39*(4), 546–557.

Gathercole, V. C. M. (2002a). Command of the mass/count distinction in bilingual and monolingual children: An English morphosyntactic distinction. In D. K. Oller & R. E. Eilers (Eds.), *Language and literacy in bilingual children*. Clevedon, UK: Multilingual Matters.

Gathercole, V. C. M. (2002b). Grammatical gender in bilingual and monolingual children: A Spanish morphosyntactic distinction. In D. K. Oller & R. E. Eilers (Eds.), *Language and literacy in bilingual children*. Clevedon, UK: Multilingual Matters.

Gathercole, V. C. M. (2002c). Monolingual and bilingual acquisition: Learning different treatments of that-trace phenomena in English and Spanish. In D. K. Oller & R. E. Eilers (Eds.), *Language and literacy in bilingual children*. Clevedon, UK: Multilingual Matters.

Gathercole, V. C. M. (2014). Bilingualism matters: One size does not fit all. *International Journal of Behavioral Development*, *38*(4), 359–366.

Gathercole, V. C. M., & Thomas, E. M. (2009). Bilingual first-language development: Dominant language takeover, threatened minority language take-up. *Bilingualism: Language and Cognition*, *12*(2), 213–237.

Genesee, F. (1989). Early bilingual development: One language or two? *Journal of child language*, *16*(1), 161–179.

Genesee, F., & Nicoladis, E. (2006). Bilingual acquisition. In E. H. M̃. Shatz (Ed.). Oxford, Eng.: Blackwell.

Gervain, J., & Werker, J. F. (2008). How infant speech perception contributes to language acquisition. *Language and Linguistics Compass*, *2*(6), 1149–1170.

Glembek, O., Matějka, P., Burget, L., & Mikolov, T. (2008). Advances in phonotactic language recognition. In *Ninth Annual Conference of the International Speech Communication Association*.

Goldstein, B. A., Fabiano, L., & Washington, P. S. (2005). Phonological skills in predominantly English-speaking, predominantly Spanish-speaking, and Spanish-English bilingual children. *Language, Speech, and Hearing Services in Schools*, *36*(3), 201–218.

Goldstein, B. A., & Washington, P. S. (2001). An initial investigation of phonological patterns in typically developing 4-year-old Spanish-English bilingual children. *Language, Speech, and Hearing Services in Schools*, *32*(3), 153–164.

Gollan, T. H., Starr, J., & Ferreira, V. S. (2015). More than use it or lose it: The number-of-speakers effect on heritage language proficiency. *Psychonomic bulletin & review*, *22*(1), 147–155.

Grosjean, F. (2010). *Bilingual*. Harvard University Press.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34.

Hallé, P. A., & de Boysson-Bardies, B. (1994). Emergence of an early receptive lexicon: Infants' recognition of words. *Infant Behavior and Development*, *17*(2), 119–129.

Hammer, C. S., Hoff, E., Uchikoshi, Y., Gillanders, C., Castro, D. C., & Sandilos, L. E. (2014). The language and literacy development of young dual language learners: A critical review. *Early Childhood Research Quarterly*, *29*(4), 715–733.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Havy, M., Bouchon, C., & Nazzi, T. (2016). Phonetic processing when learning words: The case of bilingual infants. *International Journal of Behavioral Development*, *40*(1), 41–52.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental review*, *26*(1), 55–88.

Hoff, E., & Core, C. (2013). Input and language development in bilingually developing children. In *Seminars in speech and language* (Vol. 34, *4*, p. 215). NIH Public Access.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of child language*, *39*(1), 1–27.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child development*, *73*(2), 418–433.

Hoff, E., Welsh, S., Place, S., & Ribot, K. (2014). Properties of dual language input that shape bilingual development and properties of environments that shape dual language input. *Input and experience in bilingual development*, *13*, 119–140.

Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy*, *15*(2), 125–150.

Hulk, A., & Müller, N. (2000). Bilingual first language acquisition at the interface between syntax and pragmatics. *Bilingualism: language and cognition*, *3*(3), 227–244.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental psychology*, *27*(2), 236.

Jackson-Maldonado, D., Marchman, V. A., & Fernald, L. C. (2013). Short-form versions of the Spanish MacArthur–Bates Communicative Development Inventories. *Applied Psycholinguistics*, *34*(4), 837–868.

Johnson, C. E., & Lancaster, P. (1998). The development of more than one phonology: A case study of a Norwegian-English bilingual child. *International Journal of Bilingualism*, *2*(3), 265–300.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of memory and language*, *44*(4), 548–567.

Juan-Garau, M., & Perez-Vidal, C. (2001). Mixing and pragmatic parental strategies in early bilingual acquisition. *Journal of child language*, *28*(1), 59–86.

Junker, D. A., & Stockman, I. J. (2002). Expressive vocabulary of German-English bilingual toddlers. *American Journal of Speech-Language Pathology*, *11*(4), 381–394.

Jusczyk, P. W. (1985). The high-amplitude sucking technique as a methodological tool in speech perception research.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630.

Kandhadai, P., Hall, D. G., & Werker, J. F. (2017). Second label learning in bilingual and monolingual infants. *Developmental science*, *20*(1), e12429.

Kehoe, M. (2002). Developing vowel systems as a window to bilingual phonology. *International Journal of Bilingualism*, *6*(3), 315–334.

Kehoe, M., Lleó, C., & Rakow, M. (2004). Voice onset time in bilingual German-Spanish children. *Bilingualism: Language and Cognition*, *7*(1), 71–88.

Kenny, P., Boulianne, G., & Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on speech and audio processing*, *13*(3), 345–354.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., & Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(5), 980–988.

Kern, S., Langue, J., Zesiger, P., & Bovet, F. (2010). Adaptations françaises des versions courtes des inventaires du développement communicatif de MacArthur-Bates. *Approche Neuropsychologique des Apprentissages chez l'Enfant*, *107*(108), 217–228.

Kreiman, J., Gerratt, B. R., & Antoñanzas-Barroso, N. (2007). Measures of the glottal source spectrum. *Journal of speech, language, and hearing research*, *50*(3), 595–610.

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

Kuwabara, H. (1996). Acoustic properties of phonemes in continuous speech for different speaking rate. In *Proceedings of the fourth international conference on spoken language, 1996. ICSLP 96.* (Vol. 4, pp. 2435–2438). IEEE.

Ladefoged, P. (1975). A course in phonetics. *University of California.*

Legacy, J., Reider, J., Crivello, C., Kuzyk, O., Friend, M., Zesiger, P., & Poulin-Dubois, D. (2017). Dog or chien? Translation equivalents in the receptive and expressive vocabularies of young French-English bilinguals. *Journal of child language*, *44*(4), 881–904.

Leopold, W. F. (1970). *Speech development of a bilingual child: Sound-learning in the first two years*. Ams Press.

Li, H., Ma, B., & Lee, K. A. (2013). Spoken language recognition: From fundamentals to practice. *Proceedings of the IEEE*, *101*(5), 1136–1159.

Liu, L., & Kager, R. (2015). Bilingual exposure influences infant VOT perception. *Infant Behavior and Development*, *38*, 27–36.

Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Gonzalez-Rodriguez, J., & Moreno, P. (2014). Automatic language identification using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5337–5341). IEEE.

Lyon, J. (1996). *Becoming bilingual: Language acquisition in a bilingual community*. Multilingual matters.

MacLeod, A. A., & Fabiano-Smith, L. (2015). The acquisition of allophones among bilingual Spanish–English and French–English 3-year-old children. *Clinical linguistics & phonetics*, *29*(3), 167–184.

Maneva, B., & Genesee, F. (2002). Bilingual babbling: Evidence for language differentiation in dual language acquisition. In *Proceedings of the Annual Boston University Conference on Language Development* (Vol. 26, *1*, pp. 383–392).

Marchman, V. A., Fernald, A., & Hurtado, N. (2010). How vocabulary size in two languages relates to efficiency in spoken word recognition by young Spanish-English bilinguals. *Journal of Child Language*, *37*(4), 817–840.

Marchman, V. A., Martínez, L. Z., Hurtado, N., Grüter, T., & Fernald, A. (2017). Caregiver talk to young Spanish-English bilinguals: comparing direct observation and parent-report measures of dual-language exposure. *Developmental science*, *20*(1).

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive psychology*, *47*(3), 241–275.

Martinez, D., Burget, L., Ferrer, L., & Scheffer, N. (2012). Ivector-based prosodic system for language identification. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4861–4864).

Matejka, P., Schwarz, P., Cernocky, J., & Chytil, P. (2005). Phonotactic language identification using high quality phoneme recognition. In *Ninth European Conference on Speech Communication and Technology*.

Mather, E., & Plunkett, K. (2011). Mutual exclusivity and phonological novelty constrain word learning at 16 months. *Journal of Child Language*, *38*(5), 933–950.

Mattock, K., Polka, L., Rvachew, S., & Krehm, M. (2010). The first steps in word learning are easier when the shoes fit: Comparing monolingual and bilingual infants. *Developmental Science*, *13*(1), 229–243.

Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental Psychobiology*, *56*(2), 154–178.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental science*, *11*(1), 122–134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Mehler, J., & Christophe, A. (1995). Maturation and learning of language in the first year of life. In M. Gazzaniga (Ed.), *The cognitive neurosciences*. Bradford Books, The MIT Press.

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.

Merriman, W. E., Bowman, L. L., & MacWhinney, B. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the society for research in child development*, i–129.

Molnar, M., Gervain, J., & Carreiras, M. (2014). Within-rhythm class native language discrimination abilities of Basque-Spanish monolingual and bilingual infants at 3.5 months of age. *Infancy*, *19*(3), 326–337.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant behavior and development*, *16*(4), 495–500.

Müller, N., & Hulk, A. (2001). Crosslinguistic influence in bilingual language acquisition: Italian and French as recipient languages. *Bilingualism: Language and cognition*, *4*(1), 1–21.

Munro, S., Ball, M. J., Müller, N., Duckworth, M., & Lyddy, F. (2005). Phonological acquisition in Welsh-English bilingual children. *Journal of Multilingual Communication Disorders*, *3*(1), 24–49.

Nakamura, J. (2016). Hidden bilingualism: ideological influences on the language practices of multilingual migrant mothers in Japan. *International Multilingual Research Journal*, *10*(4), 308–323.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, *24*(3), 756.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*(1), 1–19.

Ng, R. W., Leung, C.-C., Lee, T., Ma, B., & Li, H. (2010). Prosodic attribute model for spoken language identification. In *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (pp. 5022–5025).

Nicoladis, E., & Paradis, J. (2011). Learning to liaise and elide comme il faut: Evidence from bilingual children. *Journal of Child Language*, *38*(4), 701–730.

O'Toole, C., Gatt, D., Hickey, T. M., Miekisz, A., Haman, E., Armon-Lotem, S., ... Kern, S. (2017). Parent report of early lexical production in bilingual children: A cross-linguistic CDI comparison. *International Journal of Bilingual Education and Bilingualism*, *20*(2), 124–145.

Oller, D. K., Eilers, R. E., Urbano, R., & Cobo-Lewis, A. B. (1997). Development of precursors to speech in infants exposed to two languages. *Journal of child language*, *24*(2), 407–425.

Paquette-Smith, M., & Johnson, E. K. (2015). Spanish-accented English is Spanish to English-learning 5-month-olds. *Scottish Consortium for ICPhS*.

Paradis, J. (2001). Do bilingual two-year-olds have separate phonological systems? *International journal of bilingualism*, *5*(1), 19–38.

Patterson, J. L., & Pearson, B. Z. (2004). Bilingual lexical development: Influences, contexts, and processes.

Pearson, B. Z., & Fernandez, S. C. (1994). Patterns of interaction in the lexical growth in two languages of bilingual infants and toddlers. *Language learning*, *44*(4), 617–653.

Pearson, B. Z., Fernandez, S. C., Lewedeg, V., & Oller, D. K. (1997). The relation of input factors to lexical learning by bilingual infants. *Applied Psycholinguistics*, *18*(1), 41–58.

Pearson, B. Z., Fernández, S. C., & Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language learning*, *43*(1), 93–120.

Pearson, B. Z., Fernández, S., & Oller, D. K. (1995). Cross-language synonyms in the lexicons of bilingual infants: One language or two? *Journal of child language*, *22*(2), 345–368.

Pellegrino, F., Chauchat, J.-H., Rakotomalala, R., & Farinas, J. (2002). Can automatically extracted rhythmic units discriminate among languages? In *International Conference on Speech Prosody 2002*.

Petitto, L. A., Katerelos, M., Levy, B. G., Gauna, K., Tétreault, K., & Ferraro, V. (2001). Bilingual signed and spoken language acquisition from birth: Implications for the mechanisms underlying early bilingual language acquisition. *Journal of child language*, *28*(2), 453–496.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release)[www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology. *Ohio State University (Distributor)*.

Place, S., & Hoff, E. (2011). Properties of dual language exposure that influence 2-year-olds' bilingual proficiency. *Child development*, *82*(6), 1834–1849.

Place, S., & Hoff, E. (2016). Effects and noneffects of input in bilingual environments on dual language skills in 2 1/2-year-olds. *Bilingualism: Language and Cognition*, *19*(5), 1023–1041.

Poulin-Dubois, D., Bialystok, E., Blaye, A., Polonia, A., & Yott, J. (2013). Lexical access and vocabulary development in very young bilinguals. *International Journal of Bilingualism*, *17*(1), 57–70.

Povey, D. [Daniel], Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Schwarz, P., et al. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017). The impact of early social interactions on later language development in Spanish–English bilingual infants. *Child development*, *88*(4), 1216–1234.

Ramon-Casas, M., & Bosch, L. (2010). Are non-cognate words phonologically better specified than cognates in the early lexicon of bilingual children. In *Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology* (pp. 31–36). Cascadilla Sommerville, MA.

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive psychology*, *59*(1), 96–121.

Ramus, F. (2002a). Acoustic correlates of linguistic rhythm: Perspectives. In *Speech Prosody 2002, Aix-en-Provence*.

Ramus, F. (2002b). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, *2*(1), 85–115.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292.

Redlinger, W. E., & Park, T.-Z. (1980). Language mixing in young bilinguals. *Journal of child language*, *7*(2), 337–352.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*(1-3), 19–41.

Richardson, F., Reynolds, D., & Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, *22*(10), 1671–1675.

Ronjat, J. (1913). *Le développement du langage observé chez un enfant bilingue*. H. Champion.

Rouas, J.-L., Farinas, J., Pellegrino, F., & André-Obrecht, R. (2005). Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication*, *47*(4), 436–456.

Sadjadi, S. O., Slaney, M., & Heck, L. (2013). MSR Identity Toolbox v 1.0: A MATLAB toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter*, *1*(4), 1–32.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Schatz, T. (2016). *ABX-discriminability measures and applications* (Doctoral dissertation, Université Paris 6 (UPMC)).

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association* (pp. 1–5).

Schnitzer, M. L., & Krasinski, E. (1994). The development of segmental phonological production in a bilingual child. *Journal of Child Language*, *21*(3), 585–622.

Schnitzer, M. L., & Krasinski, E. (1996). The development of segmental phonological production in a bilingual child: A contrasting second case. *Journal of child language*, *23*(3), 547–571.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*(47), 12860–12867.

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. F. (2012). A bilingual advantage in visual language discrimination in infancy. *Psychological Science*, *23*(9), 994–999.

Sebastián-Gallés, N., & Bosch, L. (2002). Building phonotactic knowledge in bilinguals: Role of early exposure. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(4), 974.

Silven, M., Voeten, M., Kouvo, A., & Lunden, M. (2014). Speech perception and vocabulary growth: a longitudinal study of Finnish-Russian bilinguals and Finnish monolinguals from infancy to three years. *International Journal of Behavioral Development*, *38*(4), 323–332.

Singer, E., Torres-Carrasquillo, P. A., Gleason, T. P., Campbell, W. M., & Reynolds, D. A. (2003). Acoustic, phonetic, and discriminative approaches to automatic language identification. In *Eighth European conference on speech communication and technology*.

Stevens, K. (2000). *Acoustic phonetics*. Current Studies in Linguistics Series. CogNet.

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*(3), 329–353.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190.

Sundara, M., Polka, L., & Molnar, M. (2008). Development of coronal stop perception: Bilingual infants keep pace with their monolingual peers. *Cognition*, *108*(1), 232–242.

Sundara, M., & Scutellaro, A. (2011). Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*, *39*(4), 505–513.

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, *116*(2), 918–933.

Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental science*, *8*(5), 432–443.

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child development*, *72*(3), 748–767.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental psychology*, *39*(4), 706.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language learning and development*, *3*(1), 73–100.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, *15*(4), 426–445.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175.

Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*(4), 432–444.

Tong, R., Ma, B., Li, H., & Chng, E. S. (2009). A target-oriented phonotactic front-end for spoken language recognition. *IEEE transactions on audio, speech, and language processing*, *17*(7), 1335–1347.

Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., & Deller Jr, J. R. (2002). Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In *Seventh international conference on spoken language processing*.

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989.

Verma, P., & Das, P. K. (2015). I-vectors in speech processing applications: A survey. *International Journal of Speech Technology*, *18*(4), 529–546.

Vihman, M. M., Nakai, S., DePaolis, R. A., & Hallé, P. (2004). The role of accentual pattern in early lexical representation. *Journal of Memory and Language*, *50*(3), 336–353.

Vihman, M. M., Thierry, G., Lum, J., Keren-Portnoy, T., & Martin, P. (2007). Onset of word form recognition in English, Welsh, and English-Welsh bilingual infants. *Applied Psycholinguistics*, *28*(3), 475–493.

Volterra, V., & Taeschner, T. (1978). The acquisition and development of language by bilingual children. *Journal of child language*, *5*(2), 311–326.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, *316*(5828), 1159.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, *24*(11), 2143–2152.

Werker, J. F. (2012). Perceptual foundations of bilingual acquisition in infancy. *Annals of the New York Academy of Sciences*, *1251*(1), 50–61.

Werker, J. F., & Byers-Heinlein, K. (2008). Bilingualism in infancy: First steps in perception and comprehension. *Trends in cognitive sciences*, *12*(4), 144–151.

Werker, J. F., Byers-Heinlein, K., & Fennell, C. T. (2009). Bilingual beginnings to learning words. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3649–3663.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language learning and development*, *1*(2), 197–234.

Wong Fillmore, L. (1991). When learning a second language means losing the first. *Early childhood research quarterly*, *6*(3), 323–346.

Yamamoto, M. (2001). *Language use in interlingual families: A Japanese-English sociolinguistic study*. Multilingual Matters.

Yang, L.-c. (1998). Contextual effects on syllable duration. In *The third ESCA/COCOSDA workshop (ETRW) on speech synthesis*.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... Povey, D., et al. (2006). The HTK book (v 3.4). *Cambridge University*.

Zhang, Y., Chen, L., & Ran, X. (2010). Online incremental EM training of GMM and its application to speech processing applications. In *2010 IEEE 10th International Conference on Signal Processing (ICSP)* (pp. 1309–1312). IEEE.

Zissman, M. A. (1995). Language identification using phoneme recognition and phonotactic language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on* (Vol. 5, pp. 3503–3506). IEEE.

Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on speech and audio processing*, *4*(1), 31.

Résumé

Durant les premières années de leur vie, les enfants apprennent rapidement à traiter la parole à partir d'un signal acoustique continue, et très vite, ils sont capables de comprendre et de produire les sons, les mots et la structure de leur langue maternelle. Les enfants qui grandissent dans un environnement bilingue rencontrent un défi supplémentaire : ils doivent simultanément découvrir et séparer les deux langues en deux systèmes individuels, avec des unités sonores, des vocabulaires et des grammaires indépendants, sans savoir a priori combien de langues sont parlées dans leur environnement. Malgré cela, l'acquisition du langage chez les jeunes bilingues suit, dans une large mesure, une chronologie similaire à celle des enfants monolingues. Comprendre comment les enfants arrivent à découvrir la présence de deux langues dans ce qu'ils entendent, et dans quelle mesure ils arrivent à les séparer, sont des questions cruciales pour le domaine de la recherche en bilinguisme chez les enfants. Dans cette thèse, nous nous concentrons sur ces deux questions en explorant comment les propriétés perceptuelles et environnementales de ce qu'ils entendent peuvent aider ou ralentir la découverte et le développement lexical des deux langages, et si les représentations phonologiques formées par les jeunes bilingues sont spécifiques à chaque langue. Nous adoptons une approche pluridisciplinaire, en utilisant à la fois des techniques empiriques et computationnelles, qui permettent d'apporter différents éclaircissements sur la tâche de la séparation des langues à un jeune âge.

Dans la première partie de cette thèse, nous nous intéressons au problème de la découverte de deux langues dans l'input d'un point de vue acoustique. Basés sur des recherches existantes concernant les capacités de discrimination des langues chez les nouveau-nés et les enfants, et inspirés par de précédents travaux de modélisation, nous cherchons à décrire d'un point de vue computationnel la perception de la parole dans plusieurs langues chez l'enfant. En empruntant un système de l'état de l'art dans les technologies de la parole, nous avons mené une série d'expériences qui peuvent servir à comprendre quel type de représentations les jeunes enfants créent quand ils entendent des langues différentes, et comment différents facteurs peuvent influencer leur perception de la distance entre les langues.

Dans la deuxième partie, nous étudions plusieurs aspects environnementaux de l'exposition à deux langues. Des recherches précédentes sur les propriétés quantitatives et qualitatives de l'input bilingue a montré des influences fortes de la quantité relative de l'exposition à chaque langue sur le développement lexical de l'enfant, mais des résultats divergents ont été rapportés quant à l'influence de la séparation des deux langues dans leur environnement. Nous avons utilisé une méthode de journal que les parents tenaient chez eux afin d'étudier la co-existence de deux langues dans ce qu'entendent les jeunes bilingues, et d'explorer comment cela, ainsi que d'autres facteurs environnementaux, peuvent influencer l'acquisition du vocabulaire.

Dans la dernière partie de cette thèse, nous examinons la perception de règles phonologiques spécifiques à chaque langue chez les enfants à l'âge de la maternelle. Contrairement à d'autre propriétés des systèmes phonologiques chez les jeunes bilingues, leur acquisition et séparation de différentes règles phonologiques ont été très peu explorées, et les seules données antérieures proviennent d'études sur la production. Nous avons mené une expérience comportementale en utilisant un jeu vidéo, afin de mesurer la perception inter-langue de l'assimilation phonologique chez les bilingues français-anglais.

Dans son ensemble, cette thèse contribue en apportant de nouveaux éclaircissements sur la question de la séparation et l'acquisition des langues dans le bilinguisme précoce, avec de nombreuses perspectives pour des recherches futures sur le sujet.

Abstract

During the first years of life, children rapidly learn to process speech from a continuous acoustic signal, and soon become able to understand and produce the sounds, words and structure of their native language. Children growing up in a bilingual environment face an additional challenge: they must simultaneously discover and separate their bilingual input into individual (yet potentially overlapping) systems, with independent sound units, vocabularies and grammars, without knowing a priori how many languages are spoken in their environment. In spite of this, language acquisition in young bilinguals follows, to an extent, a similar time-line as in monolinguals. Understanding how children come to discover the presence of two languages in their input, and to what extent they are able to keep them apart, are to this day crucial questions to the field of childhood bilingualism. In this thesis we focus on these two questions by exploring how perceptual and environmental properties of the input can help or hinder the discovery and lexical development of two languages, and whether the phonological representations formed by young bilinguals are language-specific. In order to investigate these questions, we take a multidisciplinary approach, using both empirical and computational techniques, which can provide different insights on the task of early language separation.

In the first part of this dissertation we examine the problem of discovering two languages in the input from an acoustic perspective. Based on a large body of research on language discrimination abilities in newborns and infants, and inspired by previous modelling work, we aim to provide a computational account of infant perception of multilingual speech. Borrowing a state-of-the-art system from speech technologies, we conducted a series of computational experiments that can help us understand what kind of representations young infants form when hearing different languages, and how different factors may shape their perception of language distance.

In the second part, we investigate several environmental aspects of bilingual exposure. Previous research on quantitative and qualitative properties of bilingual input had shown strong influences of each language's relative amount of exposure on infants' lexical development, but diverging results were reported regarding the impact of the separation of the two languages in their environment. We used a home diary method to investigate the co-existence of two languages in young bilinguals' input, and explore how this and other environmental factors may influence their vocabulary acquisition.

Finally, in the last part of this dissertation, we consider bilingual preschoolers' perception of language-specific phonological rules. Unlike other properties of young bilinguals' phonological systems, their acquisition and separation of phonological rules has barely been explored, with the only prior evidence coming from production studies. We conducted a behavioral experiment using a touchpad videogame to test French-English bilinguals' cross-linguistic perception of phonological assimilations.

Overall, this thesis contributes new insights to the question of language separation and acquisition in early bilingualism, with multiple perspectives for future research on this topic.