

# UNIVERSITÉ D'AIX-MARSEILLE

## ÉCOLE DOCTORALE COGNITION, LANGAGE, ÉDUCATION

## LABORATOIRE PAROLE ET LANGAGE (UMR 7309)

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Sciences du Langage

Leonardo Francisco BARÓN BIRCHENALL

# Influence of sentence-level rhythmic regularity and phonological phrasing on linguistic accommodation during conversational interactions: the case of Spanish-speaking dyads

Soutenue le 14/12/2018 devant le jury :

Madame le Professeur Corine ASTÉSANO, Université Toulouse-Jean-Jaurès, Rapporteure
Madame le Professeur Lorraine BAQUÉ, Universitat Autònoma de Barcelona, Rapporteure
Madame Amandine MICHELAS, Chargée de recherche CNRS, LPL & Aix-Marseille Université, Examinatrice
Monsieur le Professeur Noël NGUYEN, LPL & Aix-Marseille Université, Directeur de thèse

# Table of contents

Table of contents
List of tables, figures, and graphics – iii
Abstract – 1
Résumé (français) – 3
Acknowledgements – 5
1. Prologue – 6
    1.1. Terminology – 6
    1.2. Introduction – 8
2. Accommodation – 10
    2.1. General – 10
    2.2. Types of accommodation – 12
    2.3. Development of accommodation – 14
        2.3.1. Phylogenetic development of accommodation – 14
        2.3.2. Ontogenetic development of accommodation – 17
    2.4. Modalities of accommodation – 20
        2.4.1. Phonetic accommodation – 20
        General – Accent and dialect – Speaking rate – Pitch (F0) – Vocal intensity / Amplitude – Voice onset time (VOT) and vowel spectra – Further types of phonetic accommodation – Metastudies
        2.4.2. Syntactic accommodation – 35
        2.4.3. Lexical accommodation – 37
        2.4.4. Rhythmic (and turn-taking) accommodation – 39
        2.4.5. Linguistic and speaking style accommodation – 44
        2.4.6. Gestural and postural accommodation – 46
    2.5. Characteristics of accommodation – 47
        2.5.1. Functions – 47
        2.5.2. Automaticity and degree of awareness – 51
        2.5.3. Role, gender, and social biases – 53
        2.5.4. Task difficulty and timing – 59
    2.6. Theoretical frameworks of accommodation – 61
        2.6.1. Communication accommodation theory (CAT) – 61
        2.6.2. Interactive alignment model (IAM) – 64
3. Speech rhythm – 69
    3.1. General – 69
        3.1.1. The isochrony hypothesis – 70

3.1.2. Alternative views of speech rhythm – **73**
3.2. The rhythm of Spanish – **75**
3.2.1. General facts and stress patterns – **75**
3.2.2. Phonological phrasing – **80**
Accentual feet – Accentual groups
3.2.3. Rhythmic classification of Spanish – **84**
3.2.4. Resyllabification and sirrema – **86**
Resyllabification – Sirrema
3.2.5. (Rhythmic) Secondary stress – **88**
3.2.6. Rhythmicity and rhythmic alternation principle – **92**
Rhythmicity – Rhythmic alternation principle
4. Experiments – **94**
4.1. Hypotheses – **94**
Expected results – Hypothesized results – Possible results
4.2. Materials and methods – **100**
4.2.1. Experiment 1: Acoustic evaluation – **100**
Participants – Stimuli – Procedure
4.2.2. Experiment 2: Perceptual evaluation – **106**
Participants – Stimuli – Procedure
4.3. Data analysis – **108**
Preparation of participants' recordings – Determination of dependent variables – Statistical approach
4.4. Results – **116**
4.4.1. Experiment 1: Acoustic evaluation – **118**
Rhythmic distance (RD) – Interval time (IT) – F0 range (F0R) – F0 range distance (F0RD) – F0 mean (F0M) – F0 mean distance (F0MD) – Speech rate (SR) – Speech rate distance (SRD) – Lexical repetitions (LR) – Survey
4.4.2. Experiment 2: Perceptual evaluation – **131**
Perceptual rating (PR) – Survey
5. Discussion and conclusions – **134**
6. References – **140**
Appendix 1: Complete list of sentences used in Experiment 1 – **165**

# List of tables, figures, and graphics

Table 1: Terms used interchangeably or in a very similar way in the literature on behavioral coordination – **7**
Table 2: Number of words and letters in the four blocks of stimuli – **102**
Table 3: Schematic description of the presentation of stimuli in Experiment 1 – **105**
Table 4: Schematic description of the presentation of stimuli in Experiment 2 – **107**
Table 5: Groups' and feet's lengths, quantity, and number of syllables – **116**
Table 6: General results – **116**
Figure 1: Levels of linguistic representation in the interactive alignment model – **66**
Figure 2: Relation feet-syllables according to the isochrony hypothesis – **71**
Graphic 1a: Example of a screen showed to participants in Experiment 1 – **104**
Graphic 1b: Example of a screen showed to participants in Experiment 1 – **104**
Graphic 2: Rhythmic distance score by dyad – **119**
Graphic 3: Rhythmic distance score by regularity of the sentences and type of phrasing – **120**
Graphic 4: F0 range by half of the test and type of dyad – **123**
Graphic 5: F0 range distance by regularity of the sentences and type of dyad – **124**
Graphic 6: F0 mean by mode of rendition and type of phrasing – **126**
Graphic 7: Speech rate by half of the test and type of dyad – **129**
Graphic 8: Perceptual rating by type of phrasing and type of dyad – **132**

# Abstract

This thesis has two principal aims. In the first place, we would like to offer an overview of the current academic knowledge, both theoretical and empirical, of the processes of linguistic accommodation between interlocutors, in a general sense, and of the rhythmic characteristics of the Spanish language, in particular. In the second place, we present two empirical studies designed to analyze the influence of sentence-level rhythmic regularity and phonological phrasing on the processes of linguistic accommodation.

The first study consists in the acoustic analysis of 12 dyadic interactions, in which a modified shadowing task was applied under laboratory conditions. This experiment was designed to assess the influence of using regular versus irregular rhythmic sentences, arranged in both accentual feet and accentual groups, on accommodation processes during dyadic conversational interactions between adult, unacquainted, Spanish speakers. Several acoustic-prosodic features were established as dependent variables in this experiment, including fundamental frequency (F0) average and range, speech rate, and response times. Furthermore, based on the rhythmic distance score employed by Späth et al. (2016) (which was also determined), and following the logic of the Euclidean distances, measures of similarity between speakers are proposed for speech rate, F0 average, and F0 range. Lexical accommodation within the dyads was also considered. Additionally, differences between same-gender dyads (both female / female and male / male) and mixed-gender dyads, as well as differences between modes of rendition (reading / repeating), were taken into account during this experiment.

The second study consists in a perceptual task, in which 24 participants were asked to assess the degree of rhythmic resemblance between the 12 dyads of the first study by means of a five-point Likert scale. Both studies ended with a brief survey in which participants were asked to report any particular perceived difference between the blocks of stimuli.

A series of linear mixed effects models applied to the data revealed a greater resemblance between speakers, in terms of rhythm and F0 range, during interactions involving regular rhythmic sentences compared to interactions involving irregular

rhythmic sentences, and during interactions involving sentences arranged in accentual groups compared to interactions involving sentences arranged in accentual feet. It was also observed a greater amount of lexical repetitions between participants of same sex dyads with respect to participants of mixed sex dyads. Moreover, a lower value of F0 mean and a narrower F0 range were observed during the use of both regular rhythmic sentences and sentences arranged in accentual groups compared to the opposite conditions. In addition, some known facts related to women having a higher F0 mean, a wider F0 range, and speaking slower regarding men were also found during the first experiment. As for the perceptual task, sentences of mixed dyads were rated more similar to each other with respect to sentences of female only and male only dyads (among other patterns found).

Taken together, the data gathered in this thesis indicate that regular rhythmic sentences, arranged in accentual groups, generate a greater amount of resemblance between Spanish speakers in terms of rhythm and F0 range, with respect to irregular rhythmic sentences and sentences arranged in accentual feet. Likewise, the results indicate that during conversational interactions, both rhythmic regularity and phonological phrasing have influence, not only on the degree of resemblance between speakers, but also on the average and range of the interlocutors' fundamental frequency. Finally, the complete set of results is reported and discussed in the light of the current theoretical and empirical framework presented in the first part of the thesis.

**Keywords:** linguistic accommodation, speech rhythm, temporal regularity, accentual foot, accentual group, Spanish-speaking dyads.

# Résumé (français)

Cette thèse a deux objectifs principaux. En premier lieu, on voudrait offrir un aperçu des connaissances académiques actuelles, tant théoriques qu'empiriques, des processus d'accommodation linguistique entre interlocuteurs, au sens général, et des caractéristiques rythmiques de la langue espagnole, en particulier. En second lieu, on présente deux études empiriques conçues pour analyser l'influence de la régularité rythmique au niveau des phrases et de l'arrangement phonologique sur les processus d'accommodation linguistique.

La première étude s'agit d'une analyse acoustique de 12 interactions dyadiques au cours desquelles une tâche consistant en redonner immédiatement les paroles d'un locuteur (*shadowing task*) a été modifiée et appliquée dans des conditions de laboratoire. Cette expérience visait à évaluer l'influence de l'utilisation de phrases avec un rythme régulier ou irrégulier, disposées bien en pieds accentuels ou bien en groupes accentuels, sur les processus d'accommodation lors des conversations dyadiques entre adultes hispanophones inconnus. Plusieurs caractéristiques acoustiques-prosodiques ont été établies en tant que variables dépendantes dans cette expérience, y compris la moyenne et l'étendue de la fréquence fondamentale (F0), le débit de parole, et les temps de réponse. De même, des mesures de similarité entre locuteurs (basées sur des distances Euclidiennes) sont proposées pour la moyenne et l'étendue de la F0, ainsi que pour le débit de parole. Ces mesures sont aussi basées sur le score de distance rythmique utilisé par Späth et al. (2016), qui est également employé dans cette étude. L'accommodation lexicale entre dyades, les différences entre dyades du même sexe (femme / femme et homme / homme) et dyades mixtes, et les différences entre les modes de rendu (lecture / répétition), ont aussi été prises en compte au cours de cette expérience.

La deuxième étude consiste en une tâche de perception dans laquelle 24 participants ont été invités à évaluer le degré de ressemblance rythmique entre les 12 dyades de la première étude en utilisant une échelle de Likert à cinq points. Les deux études se sont terminées par une brève enquête dans laquelle il était demandé aux participants de signaler toute différence particulière perçue entre les blocs de stimuli.

Une série de modèles linéaires à effets mixtes appliqués aux données a révélé une plus grande ressemblance entre locuteurs, en matière du rythme et de l'étendue de la F0, lors des interactions impliquant des phrases avec un rythme régulier par rapport à des interactions impliquant des phrases avec un rythme irrégulier, et lors des interactions impliquant des phrases disposées en groupes accentuels comparés aux interactions impliquant des phrases disposées en pieds accentuels. Il a également été observé une plus grande quantité de répétitions lexicales entre les participants des dyades du même sexe par rapport aux participants des dyades mixtes. De plus, une valeur inférieure de la moyenne de F0 et une étendue de F0 plus étroite ont été observées lors de l'utilisation de phrases avec un rythme régulier et de phrases disposées en groupes accentuels, par rapport aux conditions expérimentales opposées. En outre, certains faits connus concernant les femmes ayant une moyenne de F0 supérieure, une étendue de F0 plus large, et un débit de parole plus lent quant aux hommes ont également été observés au cours de la première expérience. En ce qui concerne la tâche de perception, les phrases des dyades mixtes ont été notées de manière plus similaire les unes aux autres par rapport aux phrases des dyades de femmes et des dyades d'hommes (parmi d'autres résultats trouvés). Dans l'ensemble, les données rassemblées dans cette thèse indiquent que les phrases avec un rythme régulier, disposées en groupes accentuels, produisent une plus grande ressemblance entre les hispanophones en matière de rythme et de l'étendue de la F0, par rapport aux phrases avec un rythme irrégulier et aux phrases disposées en pieds accentuels. De plus, les données indiquent que lors des interactions conversationnelles, la régularité rythmique et l'arrangement phonologique ont une influence non seulement sur le degré de ressemblance entre interlocuteurs, mais également sur la moyenne et l'étendue des fréquences fondamentales des interlocuteurs. Enfin, l'ensemble complet des résultats est rapporté et discuté à la lumière du cadre théorique et empirique actuel présenté dans la première partie de la thèse.

**Mots-clés** : accommodation linguistique, rythme de la parole, régularité temporelle, pied accentuel, groupe accentuel, dyades hispanophones.

# Acknowledgements

First of all, I would like to thank my family for all their support during these years as a PhD student far away from home. Additionally, this work would not have been possible without the financial support of Colciencias, from Bogotá, Colombia, and the academic support of my adviser, professor Noël Nguyen.

I am grateful as well to Aix-Marseille University (AMU), the Laboratoire Parole et Langage (LPL), and the CNRS, in France, and to Fundación Universitaria Los Libertadores (FULL), and the Behavioral Sciences Research Group (EC), especially to professor Oliver Müller, in Colombia.

I would like to express my very great appreciation to professors Guillermo Toledo and José Ignacio Hualde for their valuable and constructive suggestions and commentaries during the planning of this research work.

Finally, I would like to extend my thanks to Hernán Jurado, Óscar Galindo, Edwin Oliveros, Andrea Obando, Carlos Toro, the coordinators of the laboratory of the FULL, and the participants in the experiments.

# 1. Prologue

## 1.1. Terminology

In the literature on behavioral coordination diverse terms are used interchangeably or in a very similar way, hindering thus the compilation and comparison of data. For instance, several authors argue that interlocutors *align* between each other in a mainly automatic and unconscious manner (e.g. Garrod & Pickering, 2004). Confusion may arise when the term *alignment* is also referred to as *convergence* in terms of a *coordination* process (Pardo, 2006), whereas *convergence* is also referred to as *imitation* (Babel, 2011). Now, according to Louwerse, Dale, Bard and Jeuniaux (2012), "*coordination* appears to be a largely conscious mechanism, an intentional attempt to participate in a joint activity.... *Imitation*, on the other hand, is often unconscious and automatic" (p. 3; italics added). Table 1 illustrates the case.

For the purposes of this thesis, we will use the following definitions of terms (based for the most part on the distinction proposed by Louwerse et al., 2012): **(a) *accommodation*:** phenomenon in which talkers alter diverse linguistic and paralinguistic features in response to specific characteristics of received stimuli; **(b) *imitation*:** a mainly unconscious and automatic form matching process, which does not have to be time aligned (in general terms, of course, imitation can be performed consciously) [1]; **(c) *convergence***, ***synchronization***, or simply ***sync*:** a symmetric or asymmetric increase of similarity of diverse linguistic and paralinguistic features of two or more individuals during an interaction, occurring in an involuntary and immediate manner, rather than intentionally[2]; **(d) *entrainment*:** the process of yoking together two oscillatory systems so that their periods of oscillation become related (Cummins, 2009a). We will try, as far as possible, to unify the use of terms when referring them from the different sources. When it cannot be

[1] Moreover, interlocutors can coordinate their behavior *without* necessarily imitating one another, for example, when collaborating to solve problems with a mutually understood structure (Louwerse et al., 2012).

[2] In fact, we consider the term *convergence* more suitable than *synchronization* for the phenomenon of increasing similarity. However, we will also use *synchronization,* or *sync*, as a synonymous term for stylistic purposes only.

done, or when a term is used in its common sense, a clarification will be provided. Terms such as *coordination*, *adaptation, similarity,* and *matching* will be used in their normal, broader, sense.

**Table 1:** Terms used interchangeably or in a very similar way in the literature on behavioral coordination (with some references as examples).

| |
|---|
| **Imitation:** mimicry, contagion (Louwerse et al., 2012), convergence, accommodation (Babel, 2011), and entrainment (Shockley, Santana & Fowler, 2003). |
| **Synchronization:** entrainment (Cummins, 2009a; Louwerse et al., 2012; Peelle & Davis, 2012), alignment, temporal convergence, and behavior matching (Wagner, Malisz & Kopp, 2014). |
| **Convergence:** alignment, accumulating common ground (Pardo, 2006), imitation, accommodation (Babel, 2011), similarity (Guardiola & Bertrand, 2013), and entrainment (Levitan et al., 2012). |
| **Temporal convergence**: synchronization, entrainment, and behavior matching (Wagner et al., 2014). |
| **Entrainment:** absolute coordination (Pardo, 2006), synchronization (Cummins, 2009a; Louwerse et al., 2012; Peelle & Davis, 2012), interpersonal imitation (Shockley et al., 2003), alignment, temporal convergence, behavior matching (Wagner et al., 2014), and similarity (Edlund, Heldner & Hirschberg, 2009). |
| **Alignment:** convergence, accumulating common ground (Pardo, 2006), synchronization, entrainment, temporal convergence, behavior matching (Wagner et al., 2014), similarity (Edlund et al., 2009), and adaptation (Fine, Jaeger, Farmer & Qian, 2013). |
| **Similarity**: coordination, priming, accommodation, inter-speaker influence, and interactional synchrony (Edlund et al., 2009). |

Note also that the degree of similarity during an interaction does not necessarily have to increase or decrease linearly. In contrast, similarity can be attained early in the interaction

and remain stable over time (Edlund et al., 2009), or maybe simply be there from the beginning and remain stable over time. Several authors refer to the maintenance of a certain degree of similarity during a conversational interaction as *synchrony* (e.g. Borrie, Lubold & Pon-Barry, 2015; Edlund et al., 2009). In this thesis, we will refer to this phenomenon as **proximity**.

Moreover, when speaking to infants or children, accommodation has been termed *child-directed speech* or *motherese* (Fernald et al., 1989), when interacting with non-native interlocutors, *foreign talk* or *foreignese* (Smith, 2007), and when accommodating to a noisy environment, *Lombard effect* (Van Summers, Pisoni, Bernacki, Pedlow & Stokes, 1988) (Further details are provided by De Looze, Scherer, Vaughan & Campbell, 2014).

Finally, sometimes the definition of the terms *stress* and *accent*, and related expressions, can be misleading. They have even been confused in some research (see Bolinger & Hodapp, 1961). In this thesis we follow the following definitions of *stress, nuclear stress, accent* and *pitch accent* provided by Hualde:

> Following Metrical Theory (...), I define '**stress**' as metrical prominence at the word, phrasal or utterance level. **Nuclear** or **main stress** is the most prominent stress in an utterance. '**Accent**' is defined as 'a clearly observable prominent F0 movement, which... consistently occurs in or near the stressed syllables of the key words' (...). '**Accent**' and '**pitch-accent**' are thus synonyms in our terminology. (Hualde, 2007, p. 60; bold added)

# 1.2. Introduction

It has been proposed that the main purpose of language is to communicate (Searle, 1972), and that conversations are the fundamental context of language use, and the primary one for children acquiring language (Clark & Wilkes, 1986)[3]. From this point of view,

[3] For an alternative point of view, see Chomsky (2010), according to whom, language main purpose is not to communicate but to create and express thoughts. From a Chomskyan theoretical framework, the need for communication is not an evolutionary force driving the phylogenetic development of language. Rather, language as we know it today would have evolved due to a "slight rewiring of the brain" in a particular individual in East Africa, some 75,000 years ago (Berwick & Chomsky, 2011). In this hypothetical scenario, the "lone mutant problem" must be

language use occurs prototypically in the context of cooperative activities (also called joint actions), that is to say, activities in which two or more people engage to achieve mutual goals. During these joint actions, conversations impulse the activities forward toward goal achievement (Shockley et al., 2003). In order to establish how interlocutors behave during these interactions, interpersonal coordination during cooperative conversations has been studied for at least thirty years (Shockley, 2005).

In this thesis, we are interested in the influence of sentence-level rhythmic regularity (regular versus irregular sentences) and phonological phrasing (rhythms arranged in feet versus rhythms arranged in accentual groups) on the processes of linguistic accommodation during conversational interactions (particularly, interactions between Spanish-speaking dyads).

For attaining our aims, we have divided this document in two main parts. In the first part, we offer an overview of the current academic knowledge, both theoretical and empirical, of the processes of linguistic accommodation between interlocutors, in a general sense, and of the rhythmic characteristics of the Spanish language, in particular. In the second chapter, the subject of accommodation is treated. The types and modalities, characteristics, theoretical frameworks, and patterns of development of accommodation can be found in this chapter. In the third chapter we discuss some generalities of the speech rhythm, the isochrony hypothesis, further stands on the rhythm of speech, and what we believe are the key factors for understanding the rhythm of Spanish, including phonological phrasing, resyllabification, sirremas, secondary stresses, and rhythmicity.

In the second part of this document we present two empirical studies designed to analyze our subject of interest. In the fourth chapter we explain our main hypotheses, the methodological and statistical approaches, and the results of both experiments. In the fifth chapter, we summarize the results and we discuss them in the light of the current theoretical and empirical framework presented in the first part of the document.

---

considered: "whom would the first language user talk to?" The answer would be: no one (Fitch, 2005).

# 2. Accommodation

## 2.1. General

Accommodation is the phenomenon in which talkers alter diverse linguistic and paralinguistic features in response to specific characteristics of received stimuli (Heath, 2015). The process of accommodation may be influenced by social, cultural, and personal aspects, such as perceived social status, social biases, language background, perception of attractiveness, and gender (e.g. Babel, 2011; Louwerse et al., 2012).

Accommodation can occur during natural conversations (Pardo, 2006), or in response to both natural and manipulated recorded stimuli (Goldinger, 1998; Nielsen, 2011). Furthermore, speech accommodation may take place even if the source of the stimuli is not directed to the listener. For instance, in Delvaux and Soquet's (2007) study, an imitative effect was obtained through the exposition of participants to a different regional dialect via loudspeakers (alternately performed live and by recordings), without specifically asking them to imitate the speech or even to listen to it.

Additionally, speech accommodation can result in modifications of the already existent phonetic repertoire of two or more interlocutors (Heath, 2014). A new variation may be constituted when a speaker realizes differently a subphonemic cue associated with a phonological contrast, with respect to the signal to which he is accommodating and to his own previous speech. In this way, accommodation processes are potential sources of permanent change in the speech patterns of a whole community (Heath, 2014).

To establish the relative importance of the information carried in the speech signal underlying the process of accommodation, Cummins (2009b) recorded and analyzed speakers trying to synchronize with different types of altered recordings. Speakers' productions were then aligned with the original, unaltered recordings, in order to assess the degree of synchronization. Cummins (2009b) concludes that the amplitude envelope, the pitch contour, and the spectral qualities of the signal, each contribute to the process of accommodation between speakers. Moreover, (referring to the F0) Babel and Bulatov

(2011) suggest that there is not one single acoustic feature in speech that serves as the only, or primary, imitable feature.

Accommodation processes between interlocutors have been studied with respect to several linguistic features, such as the ones presented in the following non-exhaustive list:

- **Speech rate** (Kousidis, Dorran, McDonnell & Coyle, 2009; Kousidis et al., 2008; Levitan et al., 2012; Levitan & Hirschberg, 2011; Lubold & Pon-Barry, 2014; Pardo, 2013a; Pardo, Jay & Krauss, 2010).
- **Speech rhythm** (Lelong & Bailly, 2011; McGarva & Warner, 2003; Späth et al., 2016).
- **Pitch level** and **range** (Babel & Bulatov, 2011; Collins, 1998; De Looze, Oertel, Rauzy & Campbell, 2011; Kousidis et al., 2008, 2009; Levitan et al., 2012; Levitan & Hirschberg, 2011; Lubold & Pon-Barry, 2014; Vaughan, 2011; Ward & Litman, 2007).
- **Pitch accents** (Gessinger et al., 2018).
- **Voice intensity** (De Looze et al., 2011; Gregory & Hoyt, 1982; Levitan et al., 2012; Levitan & Hirschberg, 2011; Lubold & Pon-Barry, 2014; Natale, 1975; Vaughan, 2011; Ward & Litman, 2007).
- **Voice quality** (Levitan et al., 2012; Levitan & Hirschberg, 2011).
- **Vowels duration** (Pardo, Jordan, Mallari, Scanlon & Lewandowski, 2013).
- **Pauses duration** (De Looze et al., 2011; Edlund et al., 2009; Gregory & Hoyt, 1982).
- **Sentences duration** (Lee et al., 2018).
- **Phonetic repertoire** (Pardo, 2006).
- **Regional accent** (Delvaux & Soquet, 2007; Evans & Iverson, 2007).
- **Speaking style** (understood as variations in the realization of linguistic units such as syllables) (Kappes, Baumgaertner, Peschke & Ziegler, 2009).
- **Syntactic complexity** (Xu & Reitter, 2016).
- **Syntactic constructions** (Branigan, Pickering & Cleland, 2000).
- **Lexical choices** (Brennan, 1996; Brennan & Clark, 1996; Garrod & Anderson, 1987; Nenkova, Gravano & Hirschberg, 2008; Wang, Yen & Reitter, 2015; Ward & Litman, 2007).

- **Linguistic style** (understood as variations in linguistic properties such as number of words, number of letters in words, types of words, and grammatical tenses) (Manson, Bryant, Gervais & Kline, 2013; Niederhoffer & Pennebaker, 2002; Thomson, Murachver & Green, 2001).
- **Voice onset time (VOT)** (Nielsen, 2011; Sanchez, Miller & Rosenblum, 2010; Sancier & Fowler, 1997; Shockley, Sabadini & Fowler, 2004).
- **Vowels spectra** (formants) (Babel, 2010, 2012; Pardo et al., 2010, 2013).
- **Pre-voicing** (of voiced stops) (Mitterer & Ernestus, 2008).
- **Turns duration** (Putman & Street, 1984; Street, 1984).
- **Turn-taking** (Bosch, Oostdijk & Boves, 2005; Himberg, Hirvenkari, Mandel & Hari, 2015).
- **Lips movements** (during phonemes production) (Gentilucci & Bernardis, 2007).

The most important results of the aforementioned studies are discussed in this thesis. Moreover, several authors present an overview of the accommodation processes between speakers, including: Bonin et al. (2013), De Looze et al. (2014), Louwerse et al. (2012), and Pardo (2006). Delaherche et al. (2012), in turn, present an overview of interpersonal synchronization, its functions, and the methods for studying it (understanding *synchronization* as the dynamic and reciprocal adaptation of temporal behavioral structures between interactive partners).

Further types of behaviors have also been studied with respect to accommodation between speakers, including:

- **Facial expressions** (Chartrand & Bargh, 1999; Hess & Blairy, 2001).
- **Mannerism**s (foot shaking and face rubbing) (Chartrand & Bargh, 1999).
- **Postural sway** (Shockley et al., 2003).
- **Moods** (Hess & Blairy, 2001; Neumann & Strack, 2000).
- **Social support type** (understanding *social support* as an interlocutor's reaction to another person's emotional and informational needs) (Wang et al., 2015).

# 2.2. Types of accommodation

Heath (2014, 2015) proposes that at least six different types of accommodation can occur when considering multiple phonetic features interacting simultaneously: **(a) convergence:** increase of similarity regarding the speech of the interlocutor[4]; **(b) divergence:** increase of dissimilarity regarding the speech of the interlocutor; **(c) orthogonal accommodation:** accommodation that does not increase the similarity or the dissimilarity regarding the model. That is to say, "changing one's own speech in response to an interlocutor's speech, but in a manner not reflected in that interlocutor's speech. An example would be speaking in a whisper in response to a statement about a sleeping baby" (Heath, 2014, p. 122); **(d) antagonistic accommodation:** increase of similarity regarding one specific factor of the interlocutor's speech, while increasing the dissimilarity with respect to another factor; **(e) hyperconvergence**: increase of similarity until matching the speech of the interlocutor, and then increase of dissimilarity in the opposite direction; and **(f) null accommodation:** lack of accommodation.

De Looze and Rauzy (2011, p. 1393), in turn, propose the existence of **(a) anti-similarity**, which would be divided in **(a1) anti-proximity** (anti-synchrony in the authors' terms), "the tendency for speakers to differentiate their speech from the other's, resulting in mirror or anti-correlated patterns," and **(a2) divergence**, "[speakers'] tendency to move apart towards different directions." In addition, **(b) no-similarity** would be a situation in which speakers do not exhibit proximity, anti-proximity, convergence, or divergence. As suggested by De Looze and Rauzy (2011), the combination of these phenomena would lead to seven different states (indicated here using our own terminology): three states of similarity (**proximity**, **convergence**, or both of them), three states of anti-similarity (**anti-proximity**, **divergence**, or both of them), and one state of no similarity (**no proximity and no convergence**).

It is worth noting that synchronization / convergence has been treated as the default form of accommodation during conversations. Nonetheless, the increase of similarity between linguistic and paralinguistic behaviors is not the only possible outcome of an

[4] Convergence between two speakers can be both unidirectional (A → B), or (B → A), as well as bidirectional (A ↔ B) (Kousidis et al., 2009).

interaction. As mentioned above, an increasing difference between speakers' speech features and associated behaviors (*divergence*) may also occur (Heath, 2014, 2015).

Divergence between interlocutors may be due to different reasons, such as an infrequent behavior that might not provide enough exposure to allow synchronization (e.g. to make an "O" shape with the mouth) (Louwerse et al., 2012). Not converging with an interlocutor may also be understood as showing creativity in linguistic choices: an attractive quality that would lead to a positive impression of the speaker (Schoot et al., 2016).

The concept of *speech complementarity* is considered by Muir, Joinson, Cotterill and Dewdney (2017) as a possible explanation of the divergence in linguistic style between interlocutors observed in their study (which is commented in Section 2.4.5.). According to this concept, some divergent communicative behaviors have the function of conveying and reinforcing social roles. This would be especially true in contexts such as organizational hierarchies, which often present highly expectations about appropriate behavior at the different levels of grading. In this scenario, individuals would rely on speech complementarity to maintain and reinforce hierarchical roles.

With respect to phonological accommodation, as believed by Heath (2015), physiological or learned restrictions on articulatory movements imply that convergence along incompatible phonetic features (e.g. VOT and stop closure duration), measured under experimental conditions, is not to be expected. It should be expected, instead, divergence in at least one of the measured features (an experiment supporting this claim is discussed in Section 2.4.1.).

Finally, according to Heath (2014, 2015), speech modifications due to divergence between interlocutors do not tend to persist beyond the interaction in which they are realized, thus it is unlikely that divergent behaviors can generate stable language variations.

# 2.3. Development of accommodation

## 2.3.1. Phylogenetic development of accommodation

Human speech implies an inherent rhythmic coordination between the articulatory, respiratory, and phonatory systems. This type of vocal and bodily coordination with an

external steady beat (speech) is a rare behavior in other animals, including non-human primates. Moreover, the ability of two or more organisms to engage in group behavioral coordination with a repeating beat, as is the case of conversations, seems to be unique for humans. Nonetheless, such coordination ability can also be found at some extent in species that are not closely related to men, including frogs, crickets, and ants. These animals take advantage of this behavioral coordination for mating and defending purposes (Merker, Madison & Eckerdal, 2009).

In this respect, Merker et al. (2009) propose that the ability to coordinate movements or vocalizations, or both, with a shared, repeating interval of time, would have evolved from specific primate behaviors, such as the so-called *carnival display* (i.e. groups of chimpanzees engaged in a chaotic voice and movement exhibition; stomping, running, and slapping trees, without any explicit indication of inter-individual coordination). In this scenario, the human ability to coordinate in pairs, or in groups, with a steady beat source of sound, is seen as a refinement of an ancient connection between calls and movements already present among the human hominoid ancestors. This ability would have evolved for purposes of mate attraction, by enabling the voice coordination needed for enhancing the signal directed to distant females.

In terms of empirical research, although the existence of behavioral synchronization between non-human species has been hardly investigated (Duranton & Gaunet, 2016), there are some studies of accommodation processes between non-human animals that include bird flocks, monkeys, and bats.

Several species of non-human primates, for instance, modify the structure of their calls in response to environmental acoustic signals and conspecifics' vocalizations, which may be considered as a process of vocal accommodation with respect to spatial location and social context (Barón, 2016). Furthermore, long-term vocal accommodation in non-human primates has been reported during pair and group formation, apparently aimed to reinforce dyadic bonds and group identity (Ruch, Zürcher & Barkart, 2017).

Regarding bats, it has been found that the first vocalizations of the Egyptian fruit bat pups consist in isolation calls produced when the pup is left alone in the roost, or when he fears that he may be separated from his mother (Prat, Taub & Yovel, 2015). These calls are

innate, appear at the first day after birth, and gradually converge, in terms of resemblance of acoustic features, toward adult-like calls during the first months of life. During this period, the highly diverse repertoire of pups' vocalizations disappears and becomes normalized (adult-like) within a month. This process can be compared to the crystallization stage in birds' singing, and with babbling and phonemic reduction in production and comprehension (respectively) in infants.

Regarding birds' songs, a critical acquisition period defines the time in which melodies may be learned. This sensitive period is divided in two phases: *sensory learning* (listening) and *sensory-motor learning* (listening-repeating) (Doupe & Kuhl, 1999). During sensory-motor learning, a subsong is produced by the young birds. These subsongs are generic among individuals, yet comprising variations, in a similar way of that of infants' babbling and mice ultrasonic vocalizations (Arriaga, Zhou & Jarvis, 2012). Subsongs eventually become plastic songs, which vary greatly between implementations, while gradually converging toward the song of the bird's tutor (an adult bird) in terms of resemblance of acoustic features. Plastic songs remain until the bird crystallizes them into stable mature songs, and from then on new songs cannot be learned (Doupe & Kuhl, 1999).

In terms of functionality, behavioral coordination has an adaptive value for clusters of animals, including decreasing the pressure of predation on offspring and increasing the effectiveness of protection against predators (Duranton & Gaunet, 2016). Rapidly matching acoustic signals, for instance, allows the vocalizer to address individual conspecifics, in a context where a signal can be directed at a multitude of listeners[5].

Furthermore, based on functional parallels between humans and other species, Ruch et al. (2017) suggest that the communicative function of vocal accommodation to signal social closeness or distance to a partner or a group, along with some level of vocal control, evolved before the emergence of language rather than being the result of it. From this

[5] Interestingly, in the animal kingdom the timing of a response is a key factor in vocal matching. Whereas a prolonged interval between emissions may not be perceived as a response to the first signal, a hasty reply may be perceived as a sign of aggression. Overlapping of the signal between individuals, however, is not a common occurrence, and sometimes serves an affiliative purpose in birds' duet signing (King et al., 2014).

standpoint, vocal accommodation is seen as a pre-adaptation that would have paved the way for language evolution.

On the other hand, the evolution of an imitation ability represents a major precursor to the evolution of language and one of the main steps in the evolution beyond the great ape level (MacNeilage, 1998). However, in the words of Fitch (2010, p. 163): "the capacity of human infants and children ... to imitate motor actions (as well as vocalizations) remains unparalleled in its richness, despite clear homologs in ape behavior."

In any case, beyond homologs in great apes' imitation behaviors, it has been proposed recently that the capacity for interactional synchrony is shared between humans, chimpanzees, bonobos, and macaques (Yu, Hattori, Yamamoto & Tomonaga, 2018). In line with this statement, and after a detailed analysis of vocal accommodation in non-human primates, Ruch et al. (2017) conclude:

> There are surprisingly strong parallels in the function of vocal accommodation in humans and primates, which are consistent with the optimization of signal transmission and CAT [communication accommodation theory] and suggest an early phylogenetic origin of these functions. These results strongly suggest that this social function of accommodation already existed prior to the evolution of language. (Ruch et al., 2017, p. 13)

## 2.3.2. Ontogenetic development of accommodation

Behavioral coordination between infants and their caregivers allows them to create and maintain a strongly attached relationship that is essential for the development of the child (Duranton & Gaunet, 2016) (An overview of this topic, including infant-infant early interactions, can be found in the third chapter of Inui, 2018). It has been proposed that this process relies on brain mechanisms operating by means of coupling coordinated rhythmic oscillators (Trevarthen, 1998). In the words of Beebe, Knoblauch, Rustin and Sorter (2003, pp. 818-819): "the pacemakers of motor systems are already coupled at birth, and all movements are played out in one time frame, 'intersynchronized'.... This coupling provides a physiological basis for endogenous coordination of perception and action..." Moreover, "the biological basis for the infant's capacity to partake in synchronous social dialogue is

provided by the organization of physiological oscillators during the neonatal period, such as the biological clock and heart rhythms..." (Feldman, Mayes & Swain, 2005, p. 24).

Additionally, Feldman et al. (2005) argue that parent-infant coordination has an important role in the development of the child's brain and predicts her or his cognitive skills and behavioral adaptation in later years. Interestingly, Feldman et al. (2005) also present evidence indicating that the degree of parent-infant coordination tends to decrease in cases of maternal depression, prematurity, or multiple births.

In the following lines we sketch a chronological non-exhaustive list of developmental stages of early infancy related to capacities that underpin the process of accommodation, including imitation, coordination, and rhythmicity:

- A seemingly universal proto-language is created between mother and infant during their early interactions. Prosodic modulations play a key role during these interactions to the point that even deaf mothers initially vocalize to their deaf infants, although neither of them can hear the sound (MacNeilage, 1998).
- As early as 42 minutes after birth, newborns exhibit a rudimentary form of deferred imitation of facial gestures (Beebe et al., 2003). According to Beebe et al. (2003), such imitation capacity implies the existence of multimodal pre-symbolic representations that are stored and compared with the own motor planning to match the gestures observed and produced. These multimodal representations include spatial, temporal, and visual aspects related to actions, which serve to remember and identify different persons. Progressively, the multimodal representations would become more stable, permitting thus longer periods of time between perceiving and imitating gestures.
- Already within the first hours of life, newborns can imitate tongue and lips protrusion, mouth opening, smiles, and an expression of surprise. Tongue protrusion can be imitated after two or three minutes after have seen the model (Beebe et al., 2003).
- Neonates (from 12 hours to 2 days old) are able to detect the rhythm of adult speech and synchronize their movements with it (Condon & Sander, 1974; for a criticism of the *visual scoring approach*, applied in this study, see Section 2.4.4.).

- Two- to five-day-old infants' cries exhibit tonal contours similar to those of their mother tongue (Mampe, Friederici, Christophe & Wermke, 2009).
- Since their first weeks of life, infants are able to perceive transmodal correspondences between what they see on the faces of other persons and the proprioceptive sensations of their own face (Beebe et al., 2003). Moreover, few weeks after being born, infants are already capable of maintain direct face-to-face interactions, coordinating vocal, oral, and gestural expressions (Beebe et al., 2003).
- Within the first eight weeks of life, infants are able to integrate speech characteristics highlighted by their caregiver into their own vocal production (Van Puyvelde, Loots, Gillisjans, Pattyn & Quintana, 2015).
- During the first months of life, infants exhibit diverse repetitive rhythmic movements, including kicking, rocking, and bouncing. In terms of articulatory and phonatory regularities, babbling is another one of these repetitive rhythmic behaviors (MacNeilage, 1998). Also during the first months of life, infants acquire diverse perceptual capacities that allow them to communicate with others, including binocular vision, selective attention, memory of contexts for object recognition, and discrimination of face patterns (Beebe et al., 2003 and references therein).
- The capacity to imitate sounds has been reported as early as two to six months of age, while other reports suggest that this particular ability does not develop entirely until the second year of life (Nguyen & Delvaux, 2015 and references therein).
- At about three months of age, infants are able to produce speech-like, as well as non-speech-like, vocal sounds (Van Puyvelde et al., 2015). Also, infants begin to open and close their mouths and move their tongues while paying attention to the adult's face and voice during episodes of interaction that involve eye-to-eye contact, and sometimes, infants' voicing (Bloom, 1998). It has also been reported that during mutual vocalizations, mothers and their three-month-old infants alter the pitch ratios and timing patterns of their utterances in such a way that the dyadic vocal exchanges become tonally synchronized (Van Puyvelde et al., 2015). Additionally, during the third month of life, infants begin to participate in synchronous social interactions, in which they learn to take turns in vocal exchanges and match their

partner's gaze directions and facial expressions. "These early face-to-face interactions between parents and infants are composed of microlevel behavioral units that follow dyad-specific rhythmic patterns, and infants at that stage can anticipate the partner's rhythms and coordinate their behavior accordingly" (Feldman et al., 2005, p. 24).

- At six months, infants can discriminate between different musical features, such as rhythm, melody, tempo, and key (Trehub, 1990).
- At roughly seven months of age, infants begin to babble, producing rhythmic mouth open-close alternations accompanied by phonation. From that point on, utterances will typically have a fixed rhythm (MacNeilage, 1998).
- Nearly at 14 months of age, infants exhibit more exploratory behaviors (including gaze direction), and smile more, toward adults who imitate them with respect to adults who perform non-imitative gestures (Beebe et al., 2003).
- The ability to synchronize with an external isochronous signal does not develop until late in infancy, and becomes steady just until puberty (Merker et al., 2009).

# 2.4. Modalities of accommodation

## 2.4.1. Phonetic accommodation

### General

Studies have focused on the occurrence of conversational accommodation for more than 50 years (Kousidis et al., 2008). A great amount of evidence has been accumulated, especially over the last decade, both in laboratory settings (in which speakers are exposed to another individual's speech productions) and in real conversational interactions (a review in Nguyen & Delvaux, 2015).

Specifically, the phenomenon of phonetic convergence is understood as an increase in the similarity of the acoustic-phonetic features of the speech of two or more persons. This type of convergence can occur both when someone listens passively to speech and when interlocutors are engaged in conversation (Pardo, 2013a). Moreover, phonetic convergence towards a model speaker may happen when speech is presented via audio, and also when it

is presented visually, during a lip-reading task (e.g. Gentilucci & Bernardis, 2007; Sanchez et al., 2010). Speech modifications due to phonetic convergence can be abstracted from interactions and generalized across the own speaker's linguistic system (Babel, 2011).

Nevertheless, in terms of acoustics, complete phonetic convergence between talkers is impossible to reach, because even for a single speaker two productions of the same phonetic segment are different in terms of articulatory and phonetic detail. Thus, phonetic convergence between individuals tends to be graded and inexact (Pardo, 2006). Moreover, during experiments, the effects of phonetic convergence are typically subtle (the size of the effect is usually small), which indicates that a complete, "perfect" imitation between speakers never occurs (Nguyen & Delvaux, 2015). Even experiments explicitly demanding impersonation do not attain a complete degree of phonetic synchronization (Wretling & Eriksson, 1998).

In the following, we present a non-exhaustive overview of studies on phonetic accommodation. A broader categorization is proposed. When the discussed work takes into account diverse acoustic-prosodic features at the same time, the feature presenting the most conclusive result is used to determine the category.

## Accent and dialect

Accommodation processes between speakers with different *accents* (understood as distinctive modes of pronunciation of a language) have been more studied than accommodation between different *dialects* (understood as particular forms of language, characteristic of particular regions or social groups). In both cases, as far as we are aware, the results tend to indicate the existence of synchronization between speakers (or towards a model speaker, or towards a typical style of pronunciation).

In a study by Evans and Iverson (2007), the speech of young adults from Northern England was evaluated before and after moving to Southern England. Acoustic analyses showed that the majority of participants altered their typically northern pronunciation of vowels towards a more southern accent. Moreover, the speech of the participants was rated by trained listeners, revealing an increasing resemblance towards the southern accent over time. Interestingly, this is one of the few studies in which perceptual and acoustic evaluations yielded equivalent results.

Borrie et al., (2015), in turn, examined the effects of pathological rhythmic production (dysarthric speech) and unfamiliar (foreign) accent on phonetic accommodation between interlocutors during face-to-face interactions. The results revealed that both pathological rhythmic production and unfamiliar accent influence phonetic convergence and interfere with the success of the interaction[6]. However, unfamiliar production parameters of accented speech hindered synchronization in a lesser amount that the variable production patterns present in dysarthric speech. That is to say, even when interlocutors have different accents, a certain degree of synchronization arises during the interaction (when comparing it to pathological speech).

According to Borrie et al., (2015) the results of their study indicate that acoustic-prosodic convergence becomes increasingly difficult to achieve as the production parameters of one interlocutor deviate from the usual speech characteristics known by the other. Nonetheless, given that dysarthria affects overall prosody, including speech rhythm and pitch regulation, it cannot be established the primary cause of the observed effect on accommodation.

In a recent study conducted by Lewandowski and Nygaard (2018), American English native speakers converged with both native English-speaking and Spanish-accented English-speaking model talkers. In this study, acoustic and perceptual assessments of convergence (AXB tasks[7]) showed that talkers synchronize with both native and non-native speakers, regarding F0, duration, and vowel spectra. Moreover, acoustic measures and perceptual assessments of convergence were related, but differed regarding the English-speaking and the Spanish-accented models. For instance, although participants did not exhibit differential convergence patterns regarding acoustic measures, perceptual assessments showed that native English speakers converged more toward the non-native

[6] Although Borrie et al. (2015) use the term *entrainment* for designing what we here call *synchronization* or *convergence*. They reserve the term *synchronization* to refer to the maintenance of a certain degree of similarity between interlocutors over time, despite the changes that occur during an interaction.

[7] An *AXB task*, as applied by Pardo (2013b), is a perceptual task in which listeners must select which item, A or B, sounds more similar in pronunciation to an item X. The item X, presented just in the middle of A and B, is a speech sound produced by one talker, and the items A and B are a repetition and a pre-task sample of the same sound, respectively, produced by the other talker.

English models. These results suggest that synchronization patterns may differ across accents (see also Davis & Kim's 2018 study in Section 2.4.4., which relates the strength of a foreign accent with prosodic characteristics of English speech).

In contrast to the mentioned experiments, Aubanel and Nguyen (2010) conducted a study of automatic recognition of regional accents during dyadic conversations. The authors did not find compelling evidence of phonetic convergence between interlocutors. In this experiment, a probabilistic classifier was used to determine if some pre-identified phonetic forms were uttered by a person with a Northern French accent or a person with a Southern French accent. The authors assumed that the classification performance would decrease across the interactions if phonetic convergence should occur between the two speakers. The results, however, showed that performance remained stable throughout the interactions.

Regarding dialect accommodation, Delvaux and Soquet (2007) found that, when exposed to a different regional dialect via loudspeakers, female talkers produced vowels that were significantly different from their typical realizations and significantly closer to the model speaker's realizations. Moreover, a substantial part of the imitation effect was reported to remain up to ten minutes after the end of the exposure. The authors of the study conclude that regional dialect imitation takes place automatically and unintentionally, leaving a memory trace in the speakers (at least in this kind of non-interactive situation).

Furthermore, conducting a study of phonetic accommodation between dyads, Kim, Horton and Bradlow (2011) found that interlocutors who speak the same language and the same dialect converge more than interlocutors who speak different languages or different dialects. According to the authors, these results indicate that synchronization is facilitated when the behavior to imitate already exists within the linguistic background of the person who imitates.

## Speaking rate

Mixed evidence has been found in studies of speaking rate accommodation. For instance, an early report presented by Street (1984) indicates that persons being interviewed during

20 to 30 minutes converged with their interviewers in terms of speech rate[8]. However, these results contrast with another series of studies conducted by the same author (Putman & Street, 1984) in which no consistent patterns of convergence in speech rate were found.

Levitan and Hirschberg (2011), for their part, analyzed the degree of acoustic-prosodic convergence (although they use the term *entrainment*) in dialogues extracted from the *Columbia games corpus* (which comprises spontaneous dyadic conversations between Standard American English native speakers, engaged in a cooperative game). The authors analyzed several features, including intensity, pitch, speaking rate, and voice quality (shimmer, jitter, and noise-to-harmonic ratio). The results of the study revealed a moderate degree of convergence between interlocutors in all features at turn level (i.e. the beginning of a speaker's turn with respect to the ending of her or his interlocutor's previous turn). Other patterns of accommodation at different levels were also found (e.g. from the first to the second half of the game, at session level, and for the entire multi-game session)[9]. Additional evidence of speech rate convergence in English conversations can be found in Kousidis et al. (2008), and Manson et al. (2013) (Both of them discussed below).

Using the methodology put forth by Levitan and Hirschberg (2011), another study with mixed results related to speaking rate accommodation was presented by Lubold and Pon-Barry (2014). In this work, the authors measured the degree of acoustic-prosodic synchronization between interlocutors, relative to pitch, intensity, and speaking rate. For this purpose, pairs of students working together in a mathematical problem were analyzed and two conditions were established: *remote interaction* (by means of tablets with shared audio, video, and workspaces), and *face-to-face interaction*. A higher degree of convergence in the face-to-face interaction was found at group level. However, at dyad level, whereas all remote partners exhibited some degree of convergence, only half of the face-to-face partners converged between each other. Taken together, these results do not support the

[8] Note that the effect of convergence reported by Street (1984) is based on *pooled data* (i.e. multiple observations of different units over a period of time). Conversely, individual analysis of the dyads did not reveal a significant overall effect of convergence.

[9] In the opinion of Pardo (2013a), Levitan and Hirschberg's (2011) methodological approach "yields interesting but chaotic data, and it is not known which acoustic attributes are perceptible to listeners, and which play a relatively minor role" (Pardo (2013a, p. 3).

hypothesis of a higher degree of synchronization during direct interactions regarding remote interactions.

In another study, Levitan et al. (2012) measured the degree of convergence (*entrainment*, in their terms) of several acoustic-prosodic features between two persons playing a series of cooperative computer games. The measured features were: intensity mean and maximum, pitch mean and maximum, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate (syllables per second). The greater amount of convergence was found between mixed-gender pairs, which became more similar during their interactions in every feature measured. Female / female pairs occupied the second place, failing to converge in pitch mean, pitch max, and NHR, but converging in the rest of measures. At the third place, male / male pairs only exhibited convergence in pitch mean, pitch max, and speech rate.

With respect to the perceptual assessment of speaking rate accommodation, Pardo and colleagues (Pardo 2013a; Pardo et al., 2010) examined the degree of articulation rate synchronization between speakers using acoustic measures and AXB tasks (note that for establishing the *articulation rate* of speech, the rate of syllable onsets within an utterance is measured while eliminating pauses and silences, which, in turn, removes important temporal information that contributes to speakers' speech rate and prosody; Schultz et al., 2015). In both studies (Pardo 2013a; Pardo et al., 2010), convergence between speakers was detected by listeners during the perceptual AXB task, but the acoustic analysis did not show talkers to converge.

As we have seen until this point, there is not conclusive evidence of a tendency of talkers to synchronize their speech rate. In fact, there is even evidence of divergence effects in articulation rate in spontaneous dialogues (between unacquainted German speakers; Schweitzer & Lewandowski, 2013). In agreement with these observations, and after a carefully review of the literature, Schultz et al. (2015) conclude that, at the time, it was impossible to draw solid conclusions about convergence of speaking rate between interlocutors, due to the existence of mixed evidence.

Nonetheless, it has been recently argued that speech rate synchronization does not necessarily happen over contiguous sequences, but can rather have an extended influence

over time (Duran & Fusaroli, 2017). For instance, an interlocutor can echo an increase of speech rate made by the other speaker earlier during a conversation. Additionally, variations in speaking rate may be due to variations in the number of pauses in speech and their mean duration rather than to variations in the actual articulation rate (De Looze et al., 2014). According to Bonin et al. (2013, p. 542), "while speakers' articulation rate is rather constant in nature, one may rather accommodate their speech in terms of pause duration." In consequence, the method used to establish the existence of speaking rate synchronization, as well as the conceptualization of the phenomenon, may determine the actual degree of convergence that can be recognized.

In this regard, Edlund et al. (2009) analyzed the length of pauses (within-speaker silences) and gaps (between-speaker silences) during Swedish spontaneous dialogues. The results of this study were inconclusive regarding convergence of pause duration. Likewise, De Looze et al. (2011) did no found compelling evidence of synchronization of pause duration between speakers. These results, however, contrast with the ones presented by Gregory and Hoyt (1982), indicating that participants in dyadic interviews tend to converge with respect to duration of pauses in speech.

Recent studies have also found mixed evidence in terms of speech rate accommodation. Wynn, Borrie and Sellers (2018), for example, found that the speech rate of typically developed adults converged towards the manipulated speech rate (slow / fast) of a female model speaker during a laboratory task. However, no evidence of speech rate convergence was observed in adults affected by autism, children affected by autism, and typically developed children, with respect to the model speaker. From these results, the authors suggest that speech rate synchronization is a developmentally acquired skill, which may be impaired in individuals with autism spectrum disorder.

In another recent study, adult speakers of Hebrew significantly reduced their speech rate in response to a confederate's speech rate reduction during five minutes conversations (Freud, Ezrati & Amir, 2018). Participants' speaking rate reduction, however, was significantly smaller than the reduction of the confederate. Additionally, when the same experience was recreated with the confederate intentionally speaking in a slower than normal speaking rate, instead of reducing their speech rate in order to converge with the

confederate, the participants increased their rate of speech. With respect to this *divergent* behavior, Freud et al. (2018) suggest that the participants may have felt the need to accelerate their speech rate in order to compensate for the "time loss" due to the confederate's slow speaking rate[10].

On the other hand, some works have found that, under certain circumstances, speakers indeed tend to synchronize their speech rate. For example, it has been found that healthy individuals synchronize their speech rate with recorded stimuli of speech with abnormal rhythmic production parameters (Borrie & Liss, 2014). Participants in this study automatically increased their speech rate in response to productions from individuals with hypokinetic dysarthria (characterized by fast speech rate), and decreased their speech rate in response to productions from individuals with ataxic dysarthria (characterized by slow speech rate).

In another study, Schultz et al. (2015) examined the process of speaking rate accommodation during the reading of scripted dialogues. With the help of a beat-tracking algorithm, the authors found that the participants' speech rates were faster when their interlocutor (a confederate) spoke at a faster rate (with reference to a slower rate). Moreover, although the reported effect of convergence was mostly bidirectional, the confederate's rate influenced the participant more than the other way around. An overall augmentation of the degree of speaking rate convergence over the course of each dialogue was also reported.

Cohen, Edelist and Gleason (2017), for their part, found that, during telephonic conversations between unacquainted participants, male interlocutors speak faster when talking with other male than when talking with a female. Moreover, the results of the study showed that participants interacting with a person who spoke slowly, or a person who spoke quickly, modified their speech rate in the same direction.

Regarding the interaction between humans and computers, it has been found that users of automatized tools of information adapt their speech rate to that of a spoken

[10] It is worth noting that intentional pausing and selective syllable prolongations, along with enhanced self-monitoring, were used in this study to produce the confederate's slower speech rate.

(computerized) dialogue system, maintaining a rate suitable for automatic speech recognition. Additionally, users show preference towards an interactive voice response system (computerized) that adapts its speech rate according to the user's speech rate (with reference to a non-adaptive system) (Kousidis et al., 2009 and references therein).

## Pitch (F0)

It has been proposed that pitch *proximity* (the maintenance of a certain degree of similarity during an interaction) is a critical acoustic-prosodic dimension for predicting a successful interaction (Borrie et al., 2015). Additionally, several studies demonstrate that pitch synchronization between speakers is positively correlated with the speakers' degree of affinity and involvement during interactions (De Looze et al., 2011; De Looze et al., 2014; De Looze & Rauzy, 2011).

It has also been found that filtering the F0 out of the speech signal hinders phonetic convergence between interlocutors (Babel & Bulatov, 2011). In this study, participants converged towards the F0 of a model talker in an unfiltered (natural) condition and tended to diverge when the F0 was absent of the signal. Moreover, using a shadowing paradigm with single words, participants' utterances in the unfiltered condition were judged by listeners to be more similar to the model talker's productions, with respect to the natural condition. Nevertheless, Babel and Bulatov (2011) report that the effect of convergence established using the acoustic measures was small, and that acoustic measurement and listener judgments were not significantly correlated.

Several other studies indicate that speakers tend to synchronize their pitch. For instance, synchronization between male dyads' F0 during American English and Egyptian Arabic interviews has been reported by Gregory et al. (1993). The results of this study also suggest that pitch convergence contributes to the perceived quality of a conversation. Collins (1998), in turn, reports convergence of the average fundamental frequencies of English speakers during unrestricted conversations. Moreover, it has been reported that, during spontaneous conversations, speakers synchronize the pitch of their backchannels with the pitch of their interlocutors' preceding utterance (i.e. backchannels' pitch is more similar to

the immediately preceding utterance with respect to non-backchannels[11]) (Heldner, Edlund & Hirschberg, 2010; Levitan, Gravano & Hirschberg, 2011).

In contrast to the studies just mentioned, Manson et al. (2013) report inconsistent patterns of accommodation of mean F0 and F0 variation (in terms of standard deviation) between English speaking dyads during ten minutes of unscripted conversations. Additionally, partial evidence of pitch synchronization between interlocutors has been presented by Levitan and Hirschberg (2011), Levitan et al. (2012), Lubold and Pon-Barry (2014; discussed above), Kousidis et al. (2008), and Vaughan (2011; discussed below). On the other hand, analyses of accommodation of the maximum, minimum, and average pitch between tutors and students during a conversational interaction did not reveal distinctive signs of convergence (Ward & Litman, 2007).

Taken together, these data suggest that the fundamental frequency is one of the acoustic-prosodic features more likely to be synchronized during conversational interactions. In this respect, and considering also speech rate convergence, Bonin et al. (2013) suggest the following (using the term *accommodation* to refer to the phenomenon we call *convergence*):

> It can be hypothesized that speakers show pitch accommodation much more than other types of prosodic adaptation as the human auditory system is very sensitive to changes in pitch. Speaker's states may be mainly expressed and recognized by changes in the amount of pitch accommodation. On the contrary, small changes in articulation rate may not be as well perceived, which would result in low levels of accommodation. (Bonin et al., 2013, p. 542)

## Vocal intensity / Amplitude

Several studies have found evidence of vocal intensity / amplitude synchronization between interlocutors. Natale (1975), for instance, reports mean vocal intensity synchronization in English non-directive interviews. According to the author, increases and decreases of the interviewer's vocal intensity level produced a corresponding change in the intensity level of the persons being interviewed. Natale (1975) also argues that the need for

[11] *Backchannels* are short vocalizations or gestures produced by a speaker to indicate that he or she is following and understanding the conversation, encouraging thus the other speaker to proceed and given continuity to the interaction (Heldner, Edlund & Hirschberg, 2010).

social approval predicts the degree of convergence of vocal intensity between speakers. Further evidence of vocal intensity synchronization during interviews is presented by Gregory and Hoyt (1982).

Kousidis et al. (2008), in turn, examined informal English conversations between an adult male speaker and three different interlocutors. The following acoustic-prosodic features were analyzed: total length (duration of speech intervals not including pauses), average pitch, pitch range, mean intensity, and speech rate (number of vowels per minute). The most significant degree of convergence was found for intensity, followed by speech rate. Average pitch was also found to converge, but in a lesser extent than intensity and speech rate. No signs of convergence were found for total length and pitch range.

In another study, Coulston, Oviatt and Darves (2002) found that seven- to ten-year-old children adapt their voice intensity level relatively to that of an animated character. In this study, children interacted with a virtual character that answered their questions about marine biology using text-to-speech conversion. Ward and Litman (2007), for their part, found convergence during human / human tutoring dialogs, with respect to the maximum, minimum, and mean values of the vocal amplitude. Finally, it has also been found that female dyads synchronize their voice intensity range during conversations in a cooperative task context (Vaughan, 2011) (However, in this study average intensity and average pitch synchronization was observed only in two of the four studied dyads, whereas pitch range was observed only in one).

On the other hand, additional investigations have found mixed evidence related to convergence of vocal intensity between interlocutors. For instance, Kousidis et al. (2009) studied acoustic-prosodic accommodation during English speakers' cooperative dialogues. The features under scrutiny were: pitch mean and range, mean intensity, and speech rate (vowels per minute). No consistent patterns of synchronization between the studied features were found. In the words of the authors (Kousidis et al., 2009, p. 5): "Theoretically, [the data found] indicates unidirectional convergence ... with a lag of 10 seconds.... However, such an interpretation would be naïve, especially in case a large coefficient at lag zero is also present."

In another study, De Looze and Rauzy (2011) analyzed conversational accommodation between English speakers. Measures of pitch level and range, voice intensity level, mean pause duration, and number of pauses were taken both for the whole conversation and at various points during the interaction. Participants in the study exhibited patterns of proximity between each other rather than convergence (i.e. a maintenance of a certain degree of similarity, rather than an augmentation, was observed between the speakers).

Supplementary partial evidence of vocal intensity synchronization between interlocutors has been presented by Levitan et al. (2012), and Levitan and Hirschberg (2011) (both of them discussed above).

## Voice onset time (VOT) and vowel spectra

Several studies indicate that the voice onset time (VOT) may be synchronized between interlocutors. Sancier and Fowler (1997), for instance, analyzed the VOT of voiceless stops in a bilingual speaker of English and Portuguese (English tends to have a longer VOT in most phonetic contexts, whereas Portuguese tends to have a shorter one; Ruch et al., 2017). After staying a couple of months in Brazil, participant's VOT was shorter in both languages. Correspondingly, it was longer after staying a comparable amount of time in USA.

Likewise, Shockley et al. (2004) found evidence of convergence between American English speakers and a model speaker regarding lengthened VOTs of word initial voiceless stops. In this study, the VOTs of words beginning with the consonants /p/, /t/, and /k/ were artificially lengthened. Participants were asked to shadow (*to repeat immediately after hearing*) the words, and then the VOTs of the shadowed words were measured.

In the study conducted by Sanchez et al. (2010), visual speech syllables uttered at different rates were dubbed onto auditory speech syllables with different VOTs. Participants were then asked to shadow visual speech tokens of a face articulating /pa/ syllables at two different rates. Then, the VOTs of the shadowed responses were measured to assess the degree of resemblance with respect to the audiovisual cues. The results showed convergence between the repeaters and the model speaker with respect to the syllables' VOTs. Additionally, both visible syllable rate and transformed VOT audio influenced the VOTs of the participants' syllables. The authors conclude that, as with

acoustic information, visual speech information can also induce convergence towards a phonetically relevant property within an utterance.

On the other hand, mixed evidence regarding VOT synchronization has also been presented. In the study conducted by Nielsen (2011), for example, participants were exposed to speech uttered by a model speaker with artificially altered VOT in words beginning with the phoneme /p/. Then, participants were recorded uttering words beginning with both /p/ and /k/. The results showed VOT imitation in both types of words when normal VOT was extended but not when it was shortened. With this seminal work, Nielsen (2011) demonstrated that perceived fine phonetic details within sub-lexical units are selectively imitable, and that such imitation can be generalized to other parts of the speech. However, according to Heath (2014), this type of generalization does not result in new phonetic material, since the resulting patterns were already present in English.

Furthermore, Yu, Abrego and Sonderegger (2013) reported no overall effect of convergence between female and male participants' and a male narrator's VOT. Yu et al. (2013) argue that the differences between their conclusions and the conclusions reported in Nielsen's (2011) study are due to a lack of contextualization of the exposure materials in the latter, differences in the presence of baseline imitation, and differences of the model talkers.

Heath (2014), for his part, exposed participants to the speech of a talker with artificially extended VOT and average stop closure duration (two subphonemic features that have an inverse correlation in English). At the end of the experiment, some participants converged toward the model talker's VOT, some toward the model's closure duration, and some others toward the ratio of the two features. However, half of the participants' stop closure duration diverged with respect to the model talker.

With respect to vowel spectra, Babel (2012) reported convergence of vowel formants between male and female talkers and two male speakers during a word-shadowing task. In this study, participants exhibited different degrees of synchronization regarding the different vowels tested and the different experimental conditions. These results contrast with other studies in which no convergence has been found with respect to vowel spectra during conversational interactions (e.g. Pardo et al., 2010), and with studies showing mixed

patterns of accommodation, like the one conducted by Pardo, Gibbons, Suppes and Krauss (2012). In this long-term study, linguistic accommodation between pairs of college roommates across the academic year was analyzed. The authors found inconsistent patterns of convergence in vowel spectra (formants), in both acoustic measures and perceptual similarity tests.

Moreover, in a series of experiments, Pardo et al. (2013) analyzed phonetic accommodation in shadowed monosyllabic words using a perceptual AXB task and acoustic measures of vowel duration, F0, and vowel formants. None of the three analyzed parameters yielded a significant degree of convergence in the acoustic measurement, whereas convergence was detected in all three of them in the perceptual task.

## Further types of phonetic accommodation

Phonetic accommodation has also been studied in other linguistic features such as sentences duration, backchannels, and prevoicing. Lee et al. (2018), for example, tested four English-speaking dyads before, during, and after a cooperative maze navigation task. Speakers in three out of four dyads synchronized their sentences duration, whereas one dyad showed significant divergence. Lee et al. (2018) propose an explanation of these mixed patterns of accommodation, relating them to some articulatory accommodation measures taken during the experiment (which are beyond the aim of this thesis).

Levitan et al., (2011), for their part, found that interlocutors tend to use similar sets of backchannel-preceding cues (BCPs) increasingly over time during spontaneous dyadic conversations. According to the authors of the study, BCPs indicate to an interlocutor that a backchannel is appropriate in a given moment. Additionally, as mentioned earlier, synchronization between the pitch of one speaker's backchannels and the pitch of her or his interlocutors' preceding utterance has been reported (Heldner, Edlund & Hirschberg, 2010; Levitan et al., 2011).

Finally, Mitterer and Ernestus (2008) found that native speakers of Dutch converged towards a female model speaker with respect to the presence of prevoicing in initial voiced stops of nonwords during a shadowing task. However, no effect of convergence was observed regarding the amount of prevoicing (according to the authors, this latter feature is phonologically irrelevant in Dutch).

## Metastudies

In the opinion of Lelong and Bailly (2011), empirical existent studies suggest that phonological and phonetic convergence between interlocutors is a very weak phenomenon. Besides, they add, the hypothesis that adaptation and synchronization are immediate and fast is questionable. As seen in this section, phonetic convergence does not occur in several cases, or only occur to a certain degree. According to Pardo (2013b, p. 2), "results from multiple studies examining phonetic convergence offer an array of often confusing and disparate findings."

In this respect, Pardo and colleagues (Pardo 2006; Pardo et al., 2010) have conducted a series of studies examining phonetic convergence in conversational interactions. In these studies, participants complete a conversational task designed to induce natural inter-talker repetitions of the same lexical items. These items are then extracted from the conversations and presented to another set of listeners via an AXB perceptual similarity test. Regarding the results of these experiments, Pardo (2013b, p. 2) concludes: "measures of individual acoustic attributes did not align with the perceptual measures of phonetic convergence. Neither item duration, speaking rate, nor vowel spectra in these studies produced consistent patterns of convergence, nor were the patterns consistent with the perceptual data." Ruch et al. (2017) suggest that these inconsistencies between acoustic and perceptual data in studies of linguistic accommodation are due to the fact that listeners rely on multiple cues, rather than just the measured acoustic attribute, when they are asked to judge the speech tokens.

Likewise, from an extensive analysis of the empirical literature, Nguyen and Delvaux (2015) conclude that:

> Phonetic convergence in conversational interactions is grounded on a low-level cognitive process involving a strong sensory-motor association (...) [However, it] does not result from a purely reflexive process, since it is selective (and as such, requires selective attention to the fine-grained phonetic details of the imitated sounds and some kind of higher-level matching process between production and perception [...]) and is typically modulated by a variety of psychological and social

factors which are partly under the control of the talker. (Nguyen & Delvaux, 2015, p, 48)

From the standpoint of Ruch et al. (2017), after a detailed analysis of the literature, phonetic convergence in humans can be explained by an automatic perception-production link, together with variations in exposure that explain the social effects. The authors add that individual differences in accommodation processes do not necessarily challenge the idea of an underlying automatic mechanism, and can be due to individual differences in perception or to different degrees of attention towards phonetic details of speech, which would lead, in both cases, to differences in the input of the perception-production mechanism.

Finally, in a recent study, Weise and Levitan (2018) did not find evidence of links between measurements of acoustic-prosodic convergence. In this study, pairwise linear correlations and principal component analysis, among other types of methods, were unsuccessfully tested to find a link between indices of pitch, voice intensity, speech rate, and lexical convergence. The authors of the study conclude that synchronization processes may consist in a set of loosely linked, perhaps independent, behaviors, rather than a structured collection of interrelated conducts.

## 2.4.2. Syntactic accommodation

Syntactic synchronization is often viewed as an instance of *structural priming*. This type of priming occurs when a syntactic structure (or a lexical item), used by one interlocutor, influences the other interlocutor to reuse the same structure (or item) at a later choice-point (i.e. influences the other interlocutor to make the same linguistic decision) (Reitter et al., 2006). "Structural priming would predict that a rule (*target*) occurs more often shortly after a potential *prime* of the same rule than long afterwards – any repetition at great distance is seen as coincidental" (Reitter et al., 2006, p. 122). A critical review of this subject can be found in Pickering and Ferreira (2008), and a discussion from the evolutionary point of view in Hurford (2012, p. 185 ff.).

There are several studies that support the existence of syntactic convergence between speakers. According to Bock's (1986) findings, for instance, participants in a conversation

are likely to use syntactic constructions that they have already heard from their interlocutor rather than other possible and sometimes more common constructions. For instance, uttering "The man gave the cake to the woman" instead of "The man gave the woman the cake," after hearing "The boy gave the toy to the teacher" (Babel, 2011, p. 16).

In another study (Branigan et al., 2000), dyads of interlocutors alternately described cards depicting actions to each other. One member of the dyad, a confederate in the experiment, produced scripted descriptions with varied syntactic structures. Results showed that speakers tended to use syntactic forms that they had just heard in the confederate's description. According to Louwerse et al. (2012), the repetition of syntactic structures in Branigan et al.'s (2000) experiment occurred quickly enough to be considered as a case of convergence between interlocutors.

Wizard of Oz-studies have also shown that users of dialogue systems synchronize their syntactic structures with those emitted by a simulated computer[12] (Reitter et al., 2006). For instance, Branigan, Pickering, Pearson, McLean and Nass (2003) reported an experiment in which naive participants played a dialogue game under the impression that they were interacting with either a person or a computer (although in both cases they were interacting with a computer program that produced pre-scripted utterances). Participants in this experiment exhibited a strong tendency to repeat the syntactic form of their interlocutor's immediately preceding utterance.

For their part, Reitter et al. (2006) examined priming of syntactic rules in annotated corpora of spoken dialogue. In this study, the authors analyzed both spontaneous telephone conversations between English speakers, and task-oriented (map tasks[13]) dyadic dialogues in English. Resulting data showed that structural repetition between speakers occurs in both types of interactions, but decays quickly and nonlinearly. Moreover, task-oriented dialogues exhibited significantly more priming than spontaneous conversations, indicating

[12] In the field of human-computer interaction, a *Wizard of Oz-study* is an experiment in which participants interact with a computer system under the impression that it is autonomous, but, actually, the system and interface are operated totally or partially by a concealed person.
[13] In a *map task*, speakers must collaborate verbally to reproduce on one participant's map a route printed on the other participant's map (Aubanel & Nguyen, 2010).

that interlocutors are more likely to accommodate their syntactic choice when a goal is clearly established in the conversation than during spontaneous interactions.

As for grammatical hierarchy, Xu and Reitter (2016) reported that interlocutors converge at different levels of syntactic complexity during natural conversations (understanding *syntactic complexity* as closely related to the amount of lexical information and degree of sophistication of a sentence). Using corpus data of spoken dialogue between dyads, including telephonic conversations, in a topic-following or -leading scenario, the authors found convergence in sentence length, tree depth, and branching factor (the three levels of syntactic complexity proposed). According to Xu and Reitter (2016), these findings support the interactive alignment model (IAM) of linguistic accommodation (discussed in Section 2.6.2.).

In another study (Schoot, Heyselaar, Hagoort & Segaert, 2016), it was found that the use of passive transitive sentence structures during conversations results in a priming effect on syntactic choices in subsequent sentences. For this purpose, pairs of participants were asked to play a card game in which they had to describe photographs to each other. On the other hand, priming with an active transitive sentence structure did not influence subsequent syntactic choices. The authors explain that, in many structural alternatives, priming with the less preferred structure results in stronger syntactic priming effects (i.e. a higher degree of syntactic convergence).

Regarding the relation between structural priming and social indices, Schoot et al. (2016) analysed the influence of syntactic accommodation on the participants' degree of *likeability* (property that makes someone likeable). The results of the study are inconclusive in this regard, but according to the authors it is likely that syntactic convergence is not automatically affected by social goals (in this particular case: making your partner like you). In fact, after a detailed review, Schoot et al. (2016) conclude that: "on the one hand, studies suggest a positive influence of likeability of the partner on the strength of syntactic alignment [convergence] (...) while on the other hand, others have reported that the more speakers like or want to be liked by their partner, the *weaker* their syntactic alignment [convergence] magnitude" (p. 3).

## 2.4.3. Lexical accommodation

Lexical accommodation has been investigated in many academic works. The earlier studies measured it as primed and unprimed words frequency in halves of conversations (Bonin et al., 2013). Later, based on evidence from their own experiments, Brennan and Clark (1996) suggested that people establish temporary agreements during conversations about how to refer to, and conceptualize, objects related to the conversation topic. Once the conceptual pact has been established, either interlocutor can appeal to it to refer to the object in question. Consequently, the same or closely related terms would be used repeatedly during the conversation to refer to an object or situation, resulting in lexical convergence.

Several studies have found evidence of lexical synchronization between speakers. For example, in Goldinger's (1998) seminal study of imitation during speech shadowing, episodic effects in single words and nonwords shadowing were observed. For instance, in both immediate and delayed shadowing, imitation of low frequency words was greater than high frequency words (the effect was stronger in immediate shadowing). Furthermore, in immediate shadowing imitation increased with increasing repetitions prior to shadowing. According to Pardo et al. (2013), however, the methodology used in Goldinger's (1998) study increases the likelihood of finding convergence, limiting at the same time the generalizability of the findings. In a subsequent experiment (Goldinger & Azuma, 2004), effects of frequency and repetition were also observed. In this experiment, imitation of English words was more evident among lower frequency words than among high frequency words. Imitation also tended to increase with more training repetitions.

In another early study, Garrod and Anderson (1987) analyzed a series of dialogues between two speakers during a computer maze game task. The results revealed a general tendency of the interlocutors to convergence with respect to the spatial descriptors used during each course. During this experiment, both speakers were likely to use the same basic types of spatial descriptors in order to reach their goal. Additionally, Garrod and Anderson (1987) report an increase of the effect of lexical convergence during the dialogues.

Analyzing also a series of computer games that required verbal communication between dyads, as well as spontaneous telephonic conversations, Nenkova, Gravano and Hirschberg (2008) found that lexical synchronization of high frequency words correlates with task

success, coordinated turn-taking behavior, and perceived naturalness of the dialogue (although in the article *synchronization* is termed *entrainment*).

Regarding human-computer interactions, in a series of experiments on lexical convergence between humans and computers, Brennan (1996) found that participants tended to repeat the terms used by the computer in both text and speech interfaces. Moreover, participants were more likely to use the same term used by the computer after an explicit "correction" (e.g. *User: what college does Aida attend? / Computer: by college, do you mean school? / User: yes / Computer: the school Aida attends is Williams*), than after an "embedded" correction (e.g. *User: what college does Aida attend? / Computer: the school Aida attends is Williams*).

More recently, based on the method of Reitter, Keller and Moore (2006) to detect syntactic priming, Ward and Litman (2007) analyzed a corpus of human-human physics tutoring transcripts. The authors tested lexical convergence between speakers established by means of words priming (i.e. students repeating words uttered by the tutor). A priming approach was also used to establish the levels of convergence of the maximum, minimum, and average values of F0 and voice amplitude over each turn, for each tutor, and for each student. The results of the study revealed the existence of convergence during the tutoring dialogs, in both lexical choices and maximum and average amplitude. Analysis of F0 did not show distinctive signs of convergence.

## 2.4.4. Rhythmic (and turn-taking) accommodation

Several factors must be considered regarding rhythmic accommodation, including the definition of rhythm, the role of rhythm in turn-taking accommodation, and the level of analysis of the phenomenon (syllables, sentences, paragraphs, etc.). In the following (and in Section 3.) we present a summary of the academic work related to these topics.

In a seminal study consisting in the visual analysis of a social interaction film, Kendon (1970) reported that listeners synchronized their movements with the speech rhythm and movements of their interlocutor, even when they were not looking at each other[14]. The

[14] Here, Kendon (1970) is referring to a process known at the time as *interactional synchrony*, according to which: "the larger movement waves fit over larger segments of speech, such as

author concludes that the synchronization of the listener's movements with the speaker's general behavior depends on the listener's response to the stream of speech. For arriving at this conclusion, a *visual scoring approach* was applied by Kendon (1970). This approach consists in the assessment of the postural synchronization between individuals, scoring videotaped interactions by hand and quantifying the number of joint angle changes during such interactions. The recordings allow the evaluation of the timing of the listeners' movements with respect to the rhythmic properties of their interlocutors' speech (Shockley, 2005). In the opinion of Shockley (2005), a disadvantage of this approach is that the visual analysis of angle changes of 3-dimensional movements from 2-dimensional tapes will be distorted unless the movements are aligned with the viewing plane. In addition, this type of analysis can reveal that a joint angle has changed, but not the degree of change.

Years later, following the isochrony hypothesis tradition for the English language (the regular recurrence of stressed syllables in time; see Section 3.1.1.), Couper-Kuhlen (1993) conducted auditory and acoustic analysis of short fragments of everyday English conversations as well as a corpus of British and American radio programs and family conversations. The author concluded that speech rhythm allows the timing of turn transitions during conversations. Moreover, Couper-Kuhlen (1993) suggests that perceptually isochronous patterns may be considered as gestalt-like rhythmic structures.

In this regard, in accordance to the Gestalt's principle of good continuation, two nearby prominent syllables establish an interval in time that allows interlocutors to project how the rhythm should continue. In the words of Couper-Kuhlen (1993), referring to the speech rhythm function within verbal activity sequences:

> The establishment or maintenance of a rhythmic structure can be regarded as the preferred option. It produces an effect which [sic] might be labelled 'harmony', 'interaction proceeding smoothly' or 'take no notice', as the case may be. The destruction or breaking down of a rhythmic structure is on the whole a dispreferred option, producing by contrast effects such as 'disharmony', 'we have a problem', 'notice this', depending on situational factors ... Anticipated and early, delayed and

---

words or phrases, the smaller movement waves, contained within the larger, fit over the smaller segments, such as the syllables and the sub-syllabic changes" (Kendon, 1970, p. 103).

> late beats have in common that they draw attention to themselves by virtue of pre-empting, postponing or destroying the completion of a rhythmic pattern. (Couper-Kuhlen, 1993, p. 267)

In the opinion of Shockley et al. (2003), the data and analysis provided by Couper-Kuhlen (1993) imply that participants in a conversation converge in their speaking rhythms even to the extent that, in turn-taking, one speaker picks up on the beat of the other speaker's rhythm. In this regard, Kawasaki, Yamada, Ushiku, Miyauchi and Yamaguchi (2013) suggest that along with the content of the conversation, turn-taking requires interpersonal synchronization of speech rhythms, in terms of timing, speed, duration, and intervals.

Turn-taking, in turn, has been traditionally thought as being governed by a set of linguistic rules, but contemporary theories suggest that is rather driven by the entrainment of oscillatory processes, operating at the level of prosody and timing (Himberg et al., 2015).

> The oscillation-based theory of turn-taking ... assumes that conversation participants are entrained to a common rhythm that is established by shared syllable timing.... This shared rhythm governs the participants' 'readiness' to take turns, and it helps them to optimize turn-taking so that it does not comprise overlaps and long silences (Himberg et al., 2015, pp. 4-5).

Further empirical evidence regarding the role of speech rhythm in turn transitions can be found in Bosch et al. (2005). In this study, involving face-to-face and telephonic dyadic interactions, a significant correlation between the average duration of between-turn pauses produced by the two interlocutors in the telephonic condition was found. However, no correlation was observed for the duration of pauses between utterances within turns. In a subsequent study, Himberg et al (2015) found that when two persons were creating stories together, in turns, one word at a time, their word rhythms were strongly entrained (understanding *word-rhythm entrainment* as a phase-locking of the temporal sequences of the interlocutors' words onset times). In this study, pairs of participants were asked to create novel stories, each one contributing with one word at a time (turn-taking thus occurred after every word). Participants interacted between each other via audio-only or audio-video. In both cases, word-rhythm entrainment was observed. In the opinion of Himberg et al. (2015), taken together, their results and the results of Bosch et al. (2005)

indicate the existence of a synchronization process to a common rhythm during conversational interactions, even during telephonic conversations.

Additionally, Mooney and Sullivan (2015) analyzed twelve non-task-based dyadic English conversations. The authors found that rhythmically coordinated speech tends to occur during turn transitions between speakers, rather than during conversations. Mooney and Sullivan (2015) conclude that rhythmic coordination during turn transitions works as a contextualization cue, signaling a high degree of cohesion between speakers.

Rhythmic accommodation has also been approached from different levels of analysis. Lelong and Bailly (2011), for instance, examined the degree of resemblance between renditions of eight French peripheral oral vowels by two persons involved in a turn-by-turn game. A degree of similarity of speech rhythm was established analyzing the mean changes of relative difference between syllabic durations. The results revealed convergence of speech rhythm during the interactions, (according to the authors) probably due to the task focusing on rhyme matching (although it is not clear if it was a phenomenon of proximity rather than convergence).

Rao, Smiljanic and Diehl (2013), for their part, conducted a study involving female and male English-speaking dyads reading syllables and a short written paragraph before and after an interactive map task. Resulting data showed that male participants were more likely to converge between each other regarding speech rhythm whereas women were more likely to converge in vowels' production. In this experiment, convergence in vowels was established using measurements of the first and second formants, while the proposed measure to explore rhythmic accommodation took into account disfluencies, pauses, syllable prominence, stressed and unstressed syllables variation, and syllabic distribution (see Rao & Smiljanic, 2011 for details).

From a sentence-level rhythmic accommodation point of view, Späth et al. (2016) conducted an experiment with healthy persons and patients suffering from Parkinson's disease. The authors found that speech rhythm resemblance is greater in sentences with a metrically regular structure with respect to sentences with an irregular structure (especially in individuals with Parkinson's). For arriving at these conclusions, German iambs and trochees were used to create metrically regular sentences consisting of an

entirely uniform alternation of stressed and unstressed syllables, and metrically irregular sentences consisting of a less regular succession of stressed and unstressed syllables. A mixed reading-repetition paradigm was used to record participants' renditions of the regular and irregular sentences. Then, speech wave analyses were conducted to examine the data. Given that this study is fundamental for this thesis in terms of methodology, conceptualization, and inspiration, further details of its procedure are provided in the section *Data analysis*.

Other studies have approach the phenomenon of speech rhythmic accommodation from the standpoint of the speech rate. For instance, it has been found recently that a normal and constant speaking rate improves word recognition against informational speech masking (understanding *normal speaking rate* as a natural rate of roughly 5.4 syllables per second, and *informational speech masking* as a loop of combined continuous recordings of nonsense sentences spoken by different talkers) (Wang, Kong, Zhang, Wu & Li, 2018). Note that for Wang et al. (2018) a *normal speech rhythm* consists in a naturally and constant speech rate, whereas an "*abnormal" speech rhythm* is seen as a speech rate variation (slower / normal / faster) within a sentence rendition. During this study the authors examined the recognition of keywords at different positions within a target sentence uttered in a normal constant speaking rate, or in an artificially modulated concatenation of speaking rates: slower / normal / faster. The effect of word recognition improvement, however, was not observed when the target sentences were masked by a stream of speech-spectrum noise.

Additionally, based on studies of phonetics of impersonation, Wretling and Eriksson (1998) (also Eriksson & Wretling, 1997) suggest that whereas speakers can vary their speech rate to produce phrases or other mayor segments of speech within a given time span, articulatory timing patterns at a more local level are more rigid and very difficult to change. The authors propose that individual timing speech patterns may be mainly stable and specific: more or less "hard coded" in individual speakers. In this sense, imitation of someone's speech, or convergence with another interlocutor in terms of metrical patterns, may occur mostly with respect to the overall rate of speech, whereas the relative durations of the words or other minor segments would remain almost invariant (for a given speech rate).

From another point of view, individual speech rhythm is understood as the duration of voices and the intervals between them (Kawasaki et al., 2013). According to this premise, the authors measured the duration and intervals between letters of the alphabet alternately and sequentially pronounced by two individuals, and by one individual and one machine. Kawasaki et al. (2013) found a higher degree of synchronization during the human-human tasks compared to the human-machine tasks (in which the machine pronounced the letters at a fixed interval). Moreover, during the human-machine tasks, subjects' speech rhythms were more likely to synchronize when the voices used were familiar to the participants. These results indicate that convergence of speech rhythm is influenced by social aspects of the interaction between interlocutors, rather than being a simply automatic response to a given structural characteristic of the speech signal.

Yet another definition of vocal rhythm was proposed in a study by McGarva and Warner (2003). These authors analyzed vocal rhythm accommodation between conversational dyads (defining *vocal rhythm* as a periodic fluctuation observed in on-off vocal activity during conversations). Convergence between dyads was only found in some of the interactions. Moreover, when convergence was observed, it did not occur at the start of conversations, but rather gradually. In this respect, McGarva and Warner (2003) suggest that rhythms of dialogue may be resistant to change, due perhaps to an underlying (¿innate?) chronobiology of the speakers.

Finally, in a recent study, Davis and Kim (2018) examined the rhythmic characteristics of an English passage uttered by Korean and French L2 English talkers with strong and weak accents. Two characteristics of prosodic rhythm were measured: the degree of synchrony between stress and syllable amplitude modulation rates, and the nested clustering of peaks in the amplitude envelope over different timescales. The results of the Korean speakers showed that the two characteristics measured differed as a function of the strength of the (foreign) accent. Conversely, the data of the French speakers showed that neither measure differed as a function of the foreign accent strength.

## 2.4.5. Linguistic and speaking style accommodation

Analyzing dyadic interactions in laboratory-based Internet chat rooms, and recorded dyadic conversations between President Nixon and his assistants, Niederhoffer and

Pennebaker (2002) found evidence of *linguistic style matching* (LSM) between interlocutors, both at the conversation level and at the turn level (i.e. the beginning of a speaker's turn with respect to the ending of his or her interlocutor's previous turn). In this study, *LSM* was defined as a correlation between different psychometric properties of language, including number of words, number of letters in the words, types of words, and grammatical tenses. Additionally, LSM did not correlate with ratings of the quality of the interaction made by the participants themselves and by judges.

Manson et al., (2013) also found evidence of dyadic linguistic style matching related to function words, but not to conjunctions or quantifiers. In this study, triads of same-sexed strangers conversed freely during ten minutes and then each participant played a one-shot prisoner's dilemma game toward each co-participant[15]. Furthermore, the authors of the study found inter-individual convergence of speech rate (mean syllable duration) over the course of the whole interactions. Additionally, speech rate convergence was positively correlated with cooperation in the prisoner's dilemma game.

In another study, Muir et al. (2017) analyzed accommodation processes of linguistic style in contexts of low and high power interactions (*linguistic style* is understood here as the way in which an individual conveys a message, using function words that are processed and produced unconsciously). In this study, a group of participants had multiple short conversations with each other using an online chat system. During these conversations, participants played either a *worker role* (low power) or *a judge role* (high power), in a situation in which workers had to make a bid to the judges, who in turn had to decide about it. As reported in the paper, judges and workers exhibited divergence in their linguistic style within conversations, and not additional increase or decrease with each additional conversation was observed. Moreover, participants playing the low power role were reported as more likely to change their linguistic style to be similar to their higher power interlocutor, rather than the other way around.

[15] In the *one-shot prisoner's dilemma* participants choose whether to cooperate or defect toward a recipient. Participants gain the largest payoff when they defect while the recipient cooperates, the second largest when both cooperate, the third largest when both defect, and the lowest when the participant cooperates while the recipient defects (Manson et al, 2013).

Kappes et al. (2009), for their part, present evidence of *speaking style* convergence (understanding *speaking style* as variations in the realization of linguistic units such as syllables). The results of this study revealed a significant degree of convergence of healthy German individuals towards a model speaker, who was asked to pronounce nonwords in two different styles (with a fully expressed schwa-syllable, and with a reduction of the schwa-vowel). No effect of converge was reported for aphasic patients. The authors of the experiments conclude that "phonologically irrelevant acoustic or phonetic information contained in a spoken model may survive the translation from perception into action, even though the instruction to repeat a word or nonword is only meant to induce a reproduction of its linguistic information" (Kappes et al., 2009, p. 148).

## 2.4.6. Gestural and postural accommodation

A *gesture* is a visible movement of any part of the body, especially a hand or the head, used as an utterance or as a part of an utterance (excluding self-adaptors, such as scratching or neck massaging). Gestures produced while speaking are called *co-speech gestures* (Wagner et al., 2014). *Postures* are particular positions of the body, or the characteristic way in which someone holds its body when standing or sitting.

There is some empirical evidence of *postural sway* and gestural synchronization[16]. For instance, Shockley et al. (2003) compared the shared activity of two persons engaged in conversation with each other with the shared activity of the same two persons engaged in conversation with others. Sensors were placed on the persons' waist and forehead, to track their movements in the anterior-posterior direction. Greater shared waist activity was found when participants were conversing with each other than when they were conversing with a confederate. The shared waist activity was observed when participants were facing each other, interestingly, even in absence of visual interaction. No significant evidence of synchronization was observed when calculations were based on the displacement of the

[16] The term *postural sway* refers to the instability of the position of the center of mass over time. It occurs when supra-postural tasks, as reaching or talking, are done during upright stance, and even when standing without doing any other activities. Verbal and visual interaction may influence postural sway (Shockley, 2005).

head (the authors of the study ascribed this to the disruption of vocal gesturing involved in speaking).

Note that visibility has an important effect on the type of gestures produced during conversations. Under visibility conditions (i.e. when the speaker and the listener see each other) gestures are larger and have clear interactive functions, such as turn-taking regulation. In non-visibility conditions, the rate of gestures between speakers decreases. Additionally, *non-obligatory iconic gestures* (i.e. gestures that are not necessary to understand the accompanying speech) have been found under both visibility and non-visibility conditions (Wagner et al., 2014 and references therein).

As far as we know, to this date the mechanism responsible for the shared postural activity found by Shockley et al. (2003) is unknown. Different properties of the interaction could have been the main factor defining the synchronization (e.g., speaking rhythm, conversational turn-taking, or word similarity).

Louwerse et al. (2012), in turn, found that participants in a conversation synchronize their head movements and facial behaviors within a margin of 1.5 seconds. This temporal lag is short enough to permit imitation between interlocutors from one conversational turn to the next one. Conversely, about half of the behaviors measured by Louwerse et al. (2012) did not exhibit signs of convergence (e.g. to pout and to make an "O" shape with the mouth). The authors of the study argue that this lack of convergence was due to the lower overall frequency of occurrence of the behaviors, so no sufficient opportunity for the synchronization to take place was available.

# 2.5. Characteristics of accommodation

## 2.5.1. Functions

In general terms, imitative behaviors may improve social interactions by increasing affiliation and empathy between interactional partners and supporting *vicarious learning* (which occurs when observing, processing, and replicating other people's actions) (Adank, Hagoort & Bekkering, 2010). Additionally, it has been proposed that imitation of posture and movements increases liking and smooth interactions between interlocutors, and that

people who are naturally empathetic tend to imitate more than non-empathetic people (Chartrand & Bargh, 1999).

Regarding linguistic interactions, it has been found that imitating the pronunciation of sentences being listened improves unfamiliar accent comprehension (Adank et al., 2010). In this study, Dutch participants listened to sentences spoken in an altered accent of Dutch, then, some of them were asked to repeat the sentences while imitating the strange accent while others were asked to imitate the sentences with the standard Dutch pronunciation. Comprehension improved only to participants that imitated the specific pronunciation of the sentences. According to the authors of the study, replicating specific aspects of the execution of an action may update representations associated with it, allowing thus a better anticipation and understanding of the action to imitate. In this sense, imitating other people's actions may facilitate the prediction of such actions, particularly when their conveyed meaning is not clear.

With respect to the phenomenon of synchronizations between interlocutors, convergence processes have been associated with several nonlinguistic functions of the social discourse, as well as with linguistic functions such as accent change and dialects formation (Pardo, 2006). For instance, Borrie et al. (2015) argue that synchronization between speakers contributes to communication success by facilitating sense making and exchange of information, and establishing affiliation, *rapport* (i.e. a harmonious relationship), and intimacy. Moreover, synchronization between interlocutors would help to decrease misunderstandings, build rapport and affiliation, and attain goals faster (Bonin et al., 2013 and references therein).

More specifically, empirical evidence suggests that prosodic convergence correlates with the communicative success of an interaction by preventing misunderstandings and facilitating goal achievement (De Looze et al., 2014). Prosodic convergence would also increase the social success of interactions creating affiliation and rapport. Moreover, it has been found that the degree of lexical and syntactic repetition and the similarity between interlocutors' backchannel-preceding cues are both associated with the level of task success (Levitan et al., 2011; Reitter & Moore, 2007). Reitter and Moore (2007) conclude that the

more successfully interlocutors collaborate between each other, the more they exhibit linguistic synchronization.

Convergence between speakers also correlates with a positive evaluation towards the conversational partner, and it is positively evaluated by interlocutors (Kousidis et al., 2009). For instance, after conducting a study of problem-solving collaboration, Lee et al. (2010) found that higher values of pitch convergence between married couples were correlated with positive attitudes towards the interaction. In this study, the more the couple's pitch converged, the better became their judgment of a more satisfying interaction. Besides, in addition to correlate with a positive evaluation towards the interlocutor, acoustic-prosodic convergence may enable communication efficiency and help establishing common ground during interactions (Kousidis et al., 2008 and references therein).

Synchronization between speakers has also been understood as a multifunctional pervasive phenomenon that, once established, serves to create common ground and reinforce social affiliation (Louwerse et al., 2012). According to this view, speakers' convergence implies different levels of interactional organization, from motor behavior to linguistic descriptions, which actively constrain each other while adapting to different contexts. Establishing and increasing common ground between interlocutors, however, may not only be reached by speaking, but also by gesturing, particularly by using head gestures (nods) to express ongoing attention and understanding (Wagner, et al., 2014).

Convergence between interlocutors may also serve to accomplish mutual goals, to help people defining their identity by categorizing others and themselves into groups, and to establish a mutual comprehension by decreasing social distance (Lelong & Bailly, 2011). More specifically, prosodic synchronization has been related to functional social aspects of interactions, such as perceived naturalness of the conversation flow, interlocutors' degree of involvement, and interlocutors' affinity between each other (De Looze et al., 2014). Interestingly, in the study by De Looze et al. (2014) the functional role of prosodic synchronization just mentioned was perceived similarly by native and non-native speakers.

Further studies have found positive correlations between the amount of involvement in an interaction and the degree of synchrony between interlocutors' pitch and voice intensity level (De Looze et al., 2011; De Looze & Rauzy, 2011), and between the degree of speaking

rate convergence and the amount of cooperation during conversations (Manson et al., 2013).

Additionally, several linguistic functions have been associated with convergence between speakers. For instance, it has been argued that phonetic convergence may play an important role in the acquisition of the phonology and phonetics of a second language (Sancier & Fowler, 1997). In this regard, L2 phonetics and phonology acquisition is understood as partly relying on the ability to reproduce foreign speech sounds, so individual different capacities to imitate speech may result in behavioral foreign accent differences in late L2 learners (Nguyen & Delvaux, 2015). Moreover, in the field of sociolinguistics, accommodation processes in general are considered as one of the mechanisms responsible for channeling linguistic variation towards dialect formation, and eventually into language change (Nguyen & Delvaux, 2015).

Synchronization processes may also function as a recovery device, by marking a point on the conversation to which one can return when communication fails (Louwerse et al., 2012). Furthermore, synchronizing processes may relieve the speaker's cognitive system of the burden of constantly computing the next behavior of her or his interlocutor, reducing thus detailed planning for each behavioral channel during interaction (Louwerse et al., 2012). Finally, it has been found that convergence of prosodic features correlates positively with learning gain during interactions between students and computer tutors (Thomason, Nguyen & Litman, 2013).

According to a different stand on the specific functions of behavioral accommodation, interpersonal coordination between interlocutors results from strategic adaptive conversational goals that depend on unique contextual demands. According to this point of view, "interpersonal coordination is not beholden to any single functional explanation, but can strategically adapt to diverse conversational demands" (Duran & Fusaroli, 2017, p. 1). Moreover, even if it has been suggested that synchronization plays an important role in social interactions, understanding the phenomenon as a linear causation, the other way around is also a possibility. Cooperation and closer relationships, for instance, could increase the attention paid to the other's conduct, improving the representation of their

motor behavior and thus facilitating synchronization processes (Koban, Ramamoorthy & Konvalinka, 2017).

### 2.5.2. Automaticity and degree of awareness

According to Louwerse et al. (2012), behavioral synchronization is immediate and involuntary, rather than strictly intentional. It comprises different features in different channels, such as postural sway, eyebrow movements, and speech rate, which would be very difficult to control intentionally by the participants in a conversation. On the other hand, relations between speech perception and production are constrained by situational aspects that influence the direction and magnitude of linguistic accommodation during a conversational interaction (Pardo, 2006, referring specifically to phonetic convergence). Consequently, an ongoing debate takes place over whether or not speech perception produces linguistically significant parameters (gestural, lexical, syntactic, semantic and/or phonological), which lead automatically to imitation (Garrod & Pickering, 2004; Pardo, 2006; Pickering & Garrod, 2004). In this scenario, the process of imitation takes place through an unconscious and automatic connection between perception and production (Chartrand & Bargh, 1999).

In this respect, Goldinger (1998) proposed that the automaticity of imitation might rely on structural and functional characteristics of the episodic memory systems. “In such systems, frequency and repetition effects are an expected outcome, and the data patterns from shadowing imitation closely followed the predicted impact of frequency and repetition (Pardo, 2013b, p. 2).

Furthermore, Koban et al., (2017) suggest that interpersonal spontaneous motor synchronization is a consequence of individual brains in interaction with each other, operating under a general optimization principle of neural computation (in this context, *optimization* refers to the selection of the best possible element from a set of alternatives). The authors argue that synchronization of motor behavior between individuals is computationally more efficient and energetically less costly (than the lack of sync): therefore it would arise automatically. Additionally, greater optimization would improve coordination and greater coordination would in turn promote optimization. In sum, “synchronized behavior results in synchronized neural representations for self- and other-

generated behavior, which are reinforced as they are in line with the brain's general tendency to compress information and to reduce prediction errors" (Koban et al., 2017, p. 13).

From this perspective, *the mirror-neuron system* is a fundamental mechanism for motor synchronization, because it allows the understanding of other individuals' actions and intentions by means of a common coding scheme for self-generated actions and perception of multiple agents' actions (Koban et al., 2017). The *mirror-neuron system*, in turn, implies the existence of a perception-action coupling from low-level action to higher-level goals (Louwerse et al., 2012). This coupling relies on a neural mechanism present in the human brain, involving Broca's area, among others, and consisting of neurons similar to the *mirror neurons* found within the ventral premotor area F5 in the monkey's brain. The mirror neurons discharge both when the monkey performs an action and when it observes another individual performing it, allowing the recognition of others' motor actions by matching them with an internal motor copy (Rizzolatti, 1998).

As for the empirical evidence, automaticity of linguistic imitation has been observed in several studies, such as the one of Delvaux and Soquet (2007; discussed in Section 2.4.1.), in which imitation of speech was demonstrated in a non-interactive situation. The simple exposure of the participants to a different regional dialect, without specifically asking them to imitate the speech or even to listen to it, was sufficient to trigger imitation. Consequently, the results of this study suggest that imitation occurs automatically and unintentionally.

From another point of view, also assuming automaticity of accommodation, "research has suggested that prosodic adaptation [accommodation] is a subconscious method of achieving social approval and acceptance and is utilized to identify with a particular social group" (De Looze et al., 2014, p. 13) (See also Giles, Coupland & Coupland, 1991; and Chartrand & Bargh, 1999). As believed by Lakin and Chartrand (quoted by De Looze et al., 2014, p. 14]:

> Accommodation would have become automatic over the course of human evolution, playing an important role as a necessary pre-requisite for communicating and for maintaining harmonious relationships within a group. It would have evolved to act as a social glue, creating, facilitating and enhancing social links between individuals.

On the other hand, authors such as Heath (2014) and Koban et al. (2017) consider that behavioral synchronization occurs both consciously and unconsciously. In the opinion of Koban et al. (2017), normally, individuals do not need to be aware (or intend to be aware) for spontaneous synchronization to occur. Nevertheless, the consequences of the synchronization are consciously accessible, allowing a person, for example, to easily note that is clapping at the same time with the rest of the public during a concert.

Moreover, several contextual and intentional factors may play an important role in accommodation processes (some of them are discussed in the next section). For instance, Duran and Fusaroli (2017) reported a more stable and sustained degree of speech rate convergence during deceiving conversations than during truthful conversations. The authors of this study employed an experimental paradigm that selectively elicits deception during dyadic conversations while controlling whether interlocutors agree or disagree with each other. In this scenario, one of the interlocutors is secretly asked to argue an opinion opposite of his or hers actual beliefs, whereas the other interlocutor remains naive. In this way, a situation is created in which a significant goal for the deceiver is to maintain consistency in believability while delivering information known to be false. From the results of the study, Duran and Fusaroli (2017) conclude that low-level behavioral synchrony is sensitive to a wide-array of contextually relevant intentional goals.

### 2.5.3. Role, gender, and social biases

In the opinion of Louwerse et al. (2012), during instruction-giving tasks, asymmetry in roles tends to cause asymmetry in synchrony, because the instruction follower is more likely to do what the instruction giver has just done than vice versa. For instance, during a map task the role assignation may produce a social asymmetry, because instruction givers know what the next subgoal of the assignment is, and they are likely to initiate subtasks and determine strategies. Therefore, synchronization tends to be toward the instruction giver. This assumption is supported by Louwerse et al.'s (2012) own study (discussed in Section 2.4.6.), in which the instruction follower matched the head movements of the instruction giver more often than the other way around.

Another asymmetry in roles during dyadic tasks has been pointed out by Konvalinka et al. (2014). With the help of recordings of brain activity during dyadic interactions in a

finger-tapping task, Konvalinka et al. (2014) found that leaders (defined as the less adaptive member of the dyad) invest more brain resources in prospective planning and control of the task with respect to the other member of the dyad (the follower). In this respect, Koban et al. (2017) propose that leading a rhythm, instead of following it, implies inhibition of the representation of the other person's motor actions, or enhancement of the representation of one's own actions. This would be associated with an increase in effort and a stronger need for cognitive control.

Additionally, in the context of the communication accommodation theory (CAT) (See Section 2.6.1.), individuals in a low-power role are thought to be motivated to seek social approval from individuals in a high-power role (Muir et al., 2017). This would be obtained by the accommodation of their communications depending on specific situations (e.g. a job interview or a courtroom situation). Supporting this idea, Muir, Joinson, Cotterill and Dewdney (2016) found that, during face-to-face interactions, the linguistic style of interlocutors in a low position of power tends to converge towards the linguistic style of a partner with a higher degree of power. Similar results are presented by Muir et al. (2017) in their study about online conversations (discussed in Section 2.4.5.).

From an empirical standpoint, several studies (yielding mixed results) have focused on the influence of role, and also the influence of gender (sometimes both of them at the same time), on the process of linguistic accommodation. For instance, conducting an early study of professional interviews, Street (1984) found that male / male dyads tended to converge with respect to turn duration, whereas female / male dyads tended to diverge. Additionally, as mentioned earlier, Street (1984) found that persons being interviewed synchronized their speech rate with their interviewers. However, no influence of the participants' role during the interview was reported in this study.

Namy, Nygaard and Sauerteig (2002), for their part, conducted a shadowing task in which male and female participants repeated isolated words uttered by both male and female speakers. The results of the study showed that, overall, females converged with their interlocutor more than males. In addition, females converged more with male than with female interlocutors, whereas males exhibited a similar degree of convergence with both male and female interlocutors (convergence was established using an AXB forced-choice

perceptual procedure). Namy et al. (2002) suggest that the greater effect of convergence observed in females may be due to the fact that, perceptually speaking, women are more sensitive than men to indexical cues in their interlocutors' speech.

The results of Namy et al. (2002), just mentioned, contrast with the data provided by Pardo (2006). In Pardo's (2006) study, in which a map task was implemented, pairs of male talkers converged more than pairs of female talkers with respect to phonetic features. Additionally, in female pairs, instruction givers converged towards receivers, but receivers did not converge towards givers. On the contrary, in male pairs, instruction receivers converged towards givers more than vice versa. Further evidence of male talkers converging more than female talkers with respect to phonetic features was obtained in a similar study conducted by Pardo et al. (2010). However, in Pardo et al. (2010) one member of each dyad was instructed to imitate her or his partner. The results showed that "phonetic convergence was impacted by the role of the talker who was given the instruction. In this case, instructing Receivers to imitate led to greater convergence" (Pardo, 2013b, p. 2).

A greater degree of linguistic convergence between men (regarding women) has also been reported by Thomason et al. (2013). These authors analyzed the relation between the entrainment of students with a tutoring dialogue system and the degree of learning (*entrainment* here is understood as an unconsciously mimic of voices, diction, and other behaviors, between interlocutors)[17]. Using an audio feature extractor, Thomason et al. (2013) examined verbal interactions of students with either a pre-recorded or a synthesized tutor voice. Several prosodic features were analyzed, including mean, minimum, maximum, and standard deviation of F0 and loudness for every utterance. It was found that male mean entrainment was significantly higher than female mean entrainment on loudness min and max.

After a detailed analysis of the empirical evidence, Pardo (2006) concludes that, although imitation of socially dominant individuals is likely, the interpretation of dominance is not

[17] A *dialog system* (or *conversational agent*) is a computational agent capable of recognize and understand social behaviors. Frequently used in mobile communications, internet search engines, and assistive technologies for the elderly or communicatively impaired, these agents are intended to interact with humans, in a coherent manner, using text, speech, graphics, gestures, and other modes of communication (De Looze et al., 2014).

necessarily caused directly by the nominal role in a conjoint task. Furthermore, the roles assumed during asymmetrical tasks and conversations are not necessarily constant and not necessarily cause an asymmetrical pattern of convergence. For instance, Xu and Reitter (2016, discussed in Section 2.4.2.) found that, during dyadic conversations, the syntactic complexity of the *topic leader* decreased whereas the syntactic complexity of the *topic follower* increased. In this study, conversations were understood as consisting of several topic episodes. The unfolding of each new topic was mostly controlled by the *topic leader*, with the *topic follower* playing a more passive role in the shift. The results also showed that a leader was able to become a follower during a single conversation, and the other way around.

Another series of studies has provided further information of the role of gender on the process of linguistic accommodation. For example, in a study by Lelong and Bailly (2011, discussed in Section 2.4.4.) examining the degree of resemblance between two persons' renditions of French peripheral oral vowels, a stronger phonetic effect of proximity was observed in dyads of the same sex, particularly female / female dyads, with respect to mixed gender dyads[18]. Contrarily, Levitan et al. (2012, discussed in Section 2.4.1.) concluded that male / male pairs converge the least whereas female / male pairs converge the most (with respect to acoustic-prosodic features in the context of a cooperative computer game). The authors suggest that convergence "is more important to the perception of social behavior for mixed-gender pairs than it is for same-gender pairs, [while] it is more important to the smoothness and flow of dialogue for male-male pairs than it is for female-female or female-male pairs" (Levitan et al., 2012, p. 11).

More evidence of an advantage of mixed-gender dyads regarding linguistic convergence is provided by Quezada, Robledo, Román and Cornejo (2012). These authors propose that empathy between two interlocutors increases the degree of phonetic convergence between them, particularly with respect to the F0. For testing their hypothesis, Quezada et al. (2012) conducted a study with 27 dyads of Spanish-speaking participants. Although their overall

[18] With respect to the difference between *convergence* and *proximity*, please note the following statement of Lelong and Bailly (2011, p. 283), referring to the vocalic targets: "An evolution of convergence rates with time was expected. Convergence rates as a function of time as been plotted for each interlocutor but nothing relevant has been observed."

results did not reach a significant statistical level, collected data indicated a stronger effect of empathy on F0 convergence on mixed-gender dyads with respect to same-gender dyads.

On the other hand, several researchers have failed to find differences in convergence between men and women (e.g. Kawasaki et al., 2013; Thomson et al., 2001 and references therein; also, an overview in Pardo et al., 2010). Thomson et al. (2001), for example, analyzed how women and men accommodated to gender-preferential linguistic style during e-mail interactions. In this study, linguistic styles consisted in female- and male-preferential language features previously identified in e-mail messages (e.g. proportion of adjectives, opinions, apologies, insults, and personal information within each message). The authors of the study did not find differences between females and males with respect to accommodation to both female- and male-preferential linguistic styles. Additionally, in the opinion of De Looze and Rauzy (2011), certain types of prosodic convergence between female and male speakers may require too much vocal effort, and thus may not be exhibited during interactions.

As for the influence of social biases on the process of linguistic accommodation, researchers have mostly focused on characteristics such as desirability, attractiveness, biases, attitudes, and social status. Natale (1975), for example, found that persons who obtained high scores on a *social desirability* test were more likely to converge towards their interlocutors in terms of voice intensity level, compared to persons with lower scores on the same test[19].

Babel (2012), for her part, reports that the more attractive female participants rated a "White model talker" (sic; *White* referring to the racial identity of the talker), the more likely they were to imitate his vowels. On the contrary, the less attractive male participants rated the same model talker, the more likely they were to imitate his vowels. However, no significant relation between attractiveness and vowel imitation was found for the "Black model talker." Additionally, Babel (2012) found a higher amount of vowel convergence in the condition that involved a visual image of the model talker with respect to the condition

[19] The concept of *social desirability* "refers to the need for social approval and acceptance and the belief that this can be attained by means of culturally acceptable and appropriate behaviors" (Aubanel & Nguyen, 2010, p. 6).

with no image. For the author, this fact suggests that the social context facilitates the process of convergence.

In another study conducted by Babel (2010), it was found that the more positive the implicit social biases toward a person's place of origin are, the more that person is imitated. In this study, the convergence of New Zealand participants towards an Australian talker in terms of vowel formants frequencies was positively affected by the implicit bias of the New Zealanders towards Australia. According to Babel (2010), social biases concerning how a person feels about his or her interlocutor help to predict the extent of convergence between them.

Regarding attitudes and accommodation, in a study by Yu et al. (2013) a positive attitude towards a male narrator, and the personality trait "openness," both increased the degree in which participants imitated the narrator's extended VOT. Moreover, the authors of the study suggest that dynamics of VOT imitation are modulated by the attitudes of the person who imitates, but not by his or her gender, or by the perceived sexual orientation of the model speaker.

Additionally, within the field of sociophonetics, Giles et al. (1991) argue that the speech of individuals of lower social status tend to converge towards their interlocutor's speech, if such interlocutor is considered to be of a higher social status (see Section 2.6.1.). In line with this idea, Gregory and Webster (1996) found that, during dyadic interviews, a television host synchronized more with higher status guests with respect to lower status guests in terms of F0. Degrees of convergence during this experiment, however, were low and did not show an increase over time (so, in the terms of this thesis it was a case of *proximity*).

Further evidence of the influence of social factors on linguistic accommodation was provided by Hay, Jannedy and Mendoza (1999). Conducting a study of the speech of Oprah Winfrey (a well-known USA television personality), Hay et al. (1999) found that both lexical frequency (understanding *frequent* as occurring five or more times in Ms. Winfrey's speech, and *infrequent* as occurring fewer than five times) and ethnicity of the interlocutor (the guest in the television show) influenced the phonetic implementation of /ay/ in different words uttered by Ms. Winfrey.

## 2.5.4. Task difficulty and timing

It has been proposed that the degree of synchronization between participants in a task correlates positively with the constraints of such task: the more constraints, the more degree of synchronization (Louwerse et al., 2012). Given that the difficulty of a task correlates positively with the *cognitive workload* of the individual performing the task (understanding *cognitive workload* as "the information processing load placed on [a] human operator while performing a particular task"; Abel & Babel, 2016, p. 3), it is plausible to assume that the amount of cognitive load during a conjoint task correlates positively with the degree of synchronization between the participants involved in such task.

However, Abel and Babel (2016) reported that information processing load during a conjoint task reduces perceived linguistic convergence between interlocutors. In this study, unacquainted female / female dyads participated in a block-building task with varied levels of difficulty. Judgments of perceptual similarity performed by independent listeners and acoustic measures were employed to establish the amount of convergence between dyads. Interestingly, the results of the acoustic measures did not show the inverse correlation between the cognitive load and the linguistic convergence found with the perceptual assessments. The authors conclude that the observed data support the existence of automatic mechanisms underlying speech convergence. In this scenario, further cognitive workload may have caused a reallocation of attentional resources that would lead to "a global damping of the perception-behavior link/production-perception coupling or to dyad members working with less detailed phonetic representations than are necessary for phonetic convergence" (Abel & Babel, 2016, p. 18).

As for the temporal development of accommodation processes, it has been proposed that, during conversations, behavioral coordination may take a few seconds to occur, and its effects may persist after the end of the interaction, perhaps to be carried to the next interaction (Louwerse et al., 2012; Pardo, 2006). More specifically, it has been observed that phonetic accommodation may start during the first minutes of an interaction (Goldinger, 1998; Pardo et al., 2010), and its effects may persist even up to a week after the initial exposure (according to Goldinger & Azuma, 2004).

In Kousidis et al.'s (2008) study (discussed in Section 2.4.1.), for instance, speakers were found to converge early during the interaction regarding voice intensity and speech rate. Likewise, in the study by Delvaux and Soquet (2007, discussed in Section 2.4.1.) only a couple trials were necessary to obtain the imitation effect of a different regional dialect. Furthermore, such effect of imitation was observable on the speakers' realizations up to 10 minutes after the last exposure to the stimuli. From another point of view, due to the temporally reactive nature of conversational speech, interlocutors may not start to accommodate between each other immediately (Bonin et al., 2013; De Looze et al., 2014).

It has also been proposed that the degree of resemblance between interlocutors tends to increase over time (in our terms, to *converge* or *synchronize*) (Louwerse et al., 2012). This has been true for several studies commented in this thesis, for instance, the increasable resemblance of nodding behaviors between participants in the study by Louwerse et al. (2012, discussed in Section 2.4.6.). One way to assess this linear progression, with respect to prosodic accommodation, is to compare the interlocutors' degree of prosodic similarity during the first and second halves of an interaction, or between its first, second, and third parts (De Looze et al., 2014).

On the other hand, as mentioned earlier, the amount of similarity between speakers, related to certain linguistic or paralinguistic behaviors, may as well remain stable during the interaction (in this thesis, we refer to such phenomenon as *proximity*). This was case, for instance, in the Lelong and Bailly's (2011) study, in which no sign of an increase of resemblance during interactions was found (regarding renditions of vowels).

Moreover, considering that interlocutors pass through different phases during conversations (e.g. reflecting, arguing, giving feedback), the amount of linguistic resemblance between interlocutors may also fluctuate depending on their mental states and degree of involvement (De Looze et al., 2014; Edlund et al., 2009; Kousidis et al., 2009). For example, analyzing prosodic accommodation in Japanese dyadic telephone conversations, De Looze et al. (2014) found that prosodic resemblance did not continuously increase or decrease over the course of interactions. On the contrary, it varied several times during the conversations.

Similar results were found by Bonin et al. (2013), who analyzed telephone calls between unacquainted native Scottish English speakers. In this study, no significant differences were found between the first and second halves of the conversations (neither the first, second, and third parts), regarding prosodic and lexical similarity. The authors conclude that these types of similarity do not linearly increase over time, but rather fluctuate several times during interactions.

Further data indicating fluctuation during prosodic accommodation rather than linear increasing or decreasing have been reported for English by De Looze and Rauzy (2011), and Vaughan (2011). Nevertheless, according to the Bonin et al.'s (2013) results, fluctuations of similarity at prosodic and lexical levels are independent phenomena (i.e. they are not time aligned). The authors suggest that prosodic and lexical accommodation "evolve in a complementary manner. While lexicon accommodation would be augmented when new concepts are introduced, prosodic accommodation would serve the social function of communication and would be augmented at other instances of the conversation" (Bonin et al., 2013, p. 543).

Understanding speech accommodation as a linear or a dynamic phenomenon has led to different methods for measuring prosodic accommodation. These methods are discussed in detail by De Looze et al (2014).

# 2.6. Theoretical frameworks of accommodation

There are two predominant theoretical models regarding vocal accommodation: the *communication accommodation theory* (CAT) (Giles et al., 1991) and the *interactive alignment model* (IAM) (Pickering & Garrod, 2004)[20]. Both of them rely on extensive empirical support (some of which have been presented in previous sections). In the following, a brief description of these models is provided. An overview of the subject can be found in Ruch et al. (2017).

## 2.6.1. Communication accommodation theory (CAT)

[20] There are also researchers who propose that a combination of intentional-social factors (emphasized by the CAT) and automatic-unintentional conditions (emphasized by the IAM) may explain vocal accommodation phenomena (e.g. Babel, 2012; Pardo, 2006).

The CAT belongs to the social psychology and sociolinguistics tradition, and emphasizes the adaptive benefits of accommodation for survival and reproduction. It also entails the idea of a link between the perceived behavioral similarity of a person and the ascription of positive attributes to that person (Ruch et al., 2017). In this sense, speakers may promote social approval and efficient communication by adapting to their interlocutors' communicative behavior (Levitan et al., 2012).

The CAT was initiated in the 1970s by Howard Giles and his colleagues, under the name of *speech accommodation theory*. Initially, it was aimed at explaining language choices in communications within a group. Later on, the theory was refined and applied to a variety of communicative behaviors beyond language (Delvaux & Soquet, 2007).

At the beginning, Giles and colleagues established both convergence and divergence in the degree of regional accentedness of common English-speakers. This behavior was attributed to attempts to decrease or increase social distance between interlocutors (Pardo, 2013a). Overtime, the CAT has emphasized the communal role of communication, understanding speech as a way to attain acceptance within a social group, or to differentiate from such group. The degree of distance between members of a given group can be established by analyzing dimensions such as the dialectal variants of the speakers or the asymmetry in the social status of the interlocutors (Quezada et al., 2012).

Additionally, the CAT maintains that speakers "accommodate to their partners on an adaptation-maintenance-differentiation continuum, where at the extreme other end, they differentiate their behaviour to that of their interlocutor" (De Looze et al., 2014, p. 13). In this scenario, linguistic convergence and divergence are strategically used by speakers to minimize or maximize (respectively) the social distance with their interlocutors during conversations, and to reinforce their own social identity (Nguyen & Delvaux, 2015).

Convergence is thus seen as a result of the speaker's need for social identification and integration. In this sense, becoming more similar with one's interlocutor through behavioral resemblance aims at gaining the interlocutor's liking and expressing social closeness. Becoming less similar is seen as a way of generating and maintaining social distance (Ruch et al., 2017). For instance, as mentioned before, according to Giles et al.

(1991) the speech of individuals of lower social status tend to converge towards their interlocutor's speech, if such interlocutor is considered to be of a higher social status.

On the other hand, apart from converging in order to seek approval, it is thought that talkers can diverge with their interlocutor in order to increase the social distance between them, or to maintain their linguistic behavior in its "regular" fashion without accommodating with the other person (Abel & Babel, 2016).

Moreover, within the CAT framework accommodation not necessarily results exclusively from a conscious and intentional process. Speakers' goals may also be represented in different degrees in their speech. Consequently, changes in the communicative behavior cannot be taken as a direct index of intentions and orientations (Ruch et al., 2017).

In a recent development of the CAT, communicative behaviors have been divided in *accommodative* and *nonaccommodative* behaviors (Muir et al., 2017). The distinction between these types of behavior is thought to depend on subjective perceptions and evaluations of the recipient of the behavior in a particular context. In more detail:

> *Accommodative communications* are those that are perceived to be appropriate, desirable, or facilitating communication. Converging one's communication behaviors (...) to be similar to conversational partners is often perceived as accommodative and is positively received. *Nonaccommodative communications* are those perceived not to be adjusted appropriately for one or both individuals (...) Nonaccommodation can take the form of *overaccommodation*, if the extent of accommodation is perceived to be greater than desired (e.g., patronizing talk), whereas too little accommodation is perceived as *underaccommodation* (...) Although behaviors may be objectively accommodative (e.g., convergence in speech rate or word use), they may be subjectively evaluated by the recipient as nonaccommodative if inappropriate to the circumstances and social roles of the conversationalists (Muir et al., 2017, p. 4; italics added).

In sum, the CAT considers accommodation strategies as a way of expressing social closeness or distance, in a way that can be more or less voluntary (Ruch et al., 2017). However, in the opinion of Quezada et al. (2012), interlocutors' emotional and dispositional states are not taken into account in the CAT, and aspects such as gender, social role, or

dialect variants, are understood as the only ones responsible for the existence of convergence between speakers.

## 2.6.2. Interactive alignment model (IAM)

The IAM belongs to the cognitive psychology and psycholinguistics tradition, and emphasizes the causal mechanistic cognitive processes that result in accommodation (Ruch et al., 2017). According to this model, mutual understanding in dialogue relies in a variety of interconnected adaptation processes that occur at multiple *levels of linguistic representation* (such as lexical, syntactic, and semantic). The *alignment* at these levels leads to the *alignment* of the speakers' *situation models*, which, in turn, is the ultimate goal of a successful conversation (Xu & Reitter, 2016). In this context, *alignment* is understood as "a state in which two or more dialogue partners have an identical (or at least highly similar) representation at a particular linguistic level" (Oben & Brône, 2015, p. 550) (In the following we will elaborate the concepts of *level of linguistic representation* and *situation model*).

According to the IAM (Pickering & Garrod, 2004), the process of accommodation is essentially based on a *priming mechanism* that generates an automatic and bidirectional relationship between speech perception and production. An automatic priming account has been proposed for lexical, syntactic, and schematic parity in language use (Pardo, 2006). In this scenario, the perception of a linguistic unit automatically drives the production of that unit (cf. Pardo, 2006). The process of imitating the input is consequently automatic and uncontrollable. In the words of Ward and Litman (2007, p. 57):

> Hearing and decoding a speech unit, such as a certain word or syntactic structure for example, 'primes,' (ie: increases the activation of) corresponding internal representations. If these representations are still active during the next speech production, they are more likely to be used than alternatives which [sic] are less active.

Additionally, as stated in the IAM, during conversations the same mental representations are used to produce and understand speech. For instance,

> When Nicola says something to Harry, the utterance activates linguistic representations in Harry. Because the same representations are used in producing and understanding, Harry then has those same representations activated when he comes to speak, and he will therefore tend to use them. So Nicola's productions influence Harry's productions and their internal representations become aligned. (Garrod & Pickering, 2004, p. 9)

Moreover,

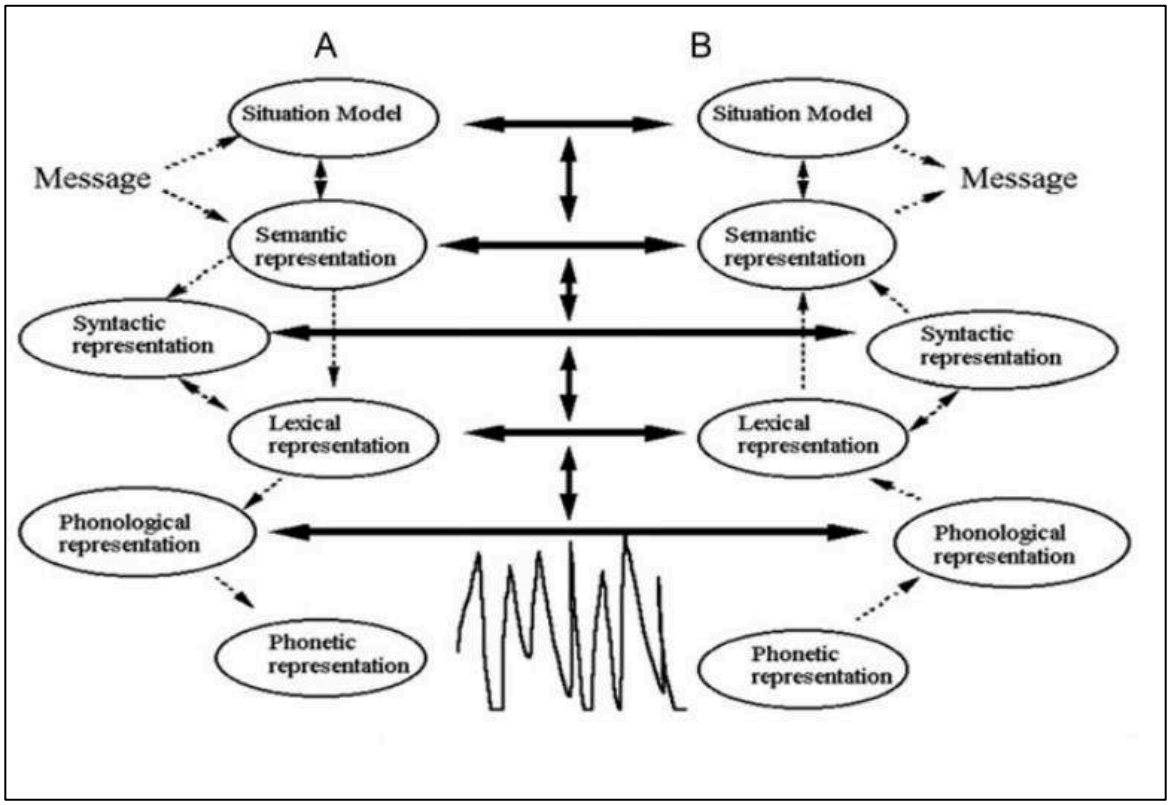
> Interlocutors tend to use many of the same computations in producing their utterances, which therefore tend to be similar at many different *linguistic levels* at the same time.... As the conversation proceeds, it will become increasingly common to use exactly the same set of computations. We call this process 'routinization'.... Such routinized expressions are similar to stock phrases and idioms, except that they only 'live' for the particular interaction. (Garrod & Pickering, 2004, p. 10, italics added)

In the Figure 1, taken from Oben and Brône (2015, p. 551), the different levels of linguistic representation proposed by the IAM are depicted.

To reach a common understanding, participants in a conversation must also align the mental representations of the things mentioned during the interaction, that is to say, their *situation models*. Such situation models consist in multi-dimensional spaces of representations that contain information about space, time, causality, and intentionality, regarding an interaction (Garrod & Pickering, 2004). In the IAM framework, conversations are successful mostly due to the persons sharing paired representations in their situation models.

Crucially, in the IAM, alignment at a lower level leads to alignment at a superior level, and alignment of superior levels facilitates further alignment of inferior levels. For instance, interlocutors align more strongly at the syntactic level when they have already aligned at the semantic level. Likewise, if interlocutors have aligned their mental models, their utterances will tend to be aligned too (phonetically, syntactically, and semantically) (Louwerse et al., 2012; Ward & Litman, 2007).

**Figure 1:** Levels of linguistic representation in the interactive alignment model (Oben & Brône, 2015, p. 551).




As proposed by Garrod and Pickering (2004, p. 9), "interlocutors become aligned at many different linguistic levels simultaneously, almost invariably without any explicit negotiation". In this sense, the process of alignment of the situation models is virtually unconscious, not requiring thus explicit negotiation. On the other hand, it has also been suggested that alignment can be inhibited when talkers' high-level goals override their low-level alignment plans (Abel & Babel, 2016).

The IAM relies also in the idea that an innate connection between perception and behavior allows the automatic and unintentional imitation of perceived actions (e.g. Chartrand & Bargh, 1999; a review in Abel & Babel, 2016). From this point of view, imitation depends on a correspondence between the perceptual and behavioral representations for actions. Phonetic convergence, for example, would rely in the active exploitation of a sensory-motor link to process speech sounds (Nguyen & Delvaux, 2015).

Motor representations of actions would be thus activated when those actions are observed in another persons (Abel & Babel, 2016).

In practical terms, during conversations one speaker perceives and comprehends what is been produced by her or his interlocutor, and the other way around. Consequently, the link between perception and production within a speaker is constantly activated (Ruch et al., 2017)[21]. In this sense, being able to adapt to another person's speech requires to be able to make a cross-modal correspondence between the acoustic information of the perceived sounds and the motor commands used to speak (Nguyen & Delvaux, 2015). However, it is also possible for a person to inhibit motor activation when factors discouraging imitation appear, or when the behavior to imitate is inconsistent with the actual objectives in a particular context (Dijksterhuis & Bargh, 2001).

Finally, it is worth noting that although the IAM was originally thought as a mechanistic model of dialogue, focusing principally on linguistic levels of representation, the strong and largely unconscious tendency of interlocutors towards convergence that it defends can also be observed in other non-linguistic behaviors, such as gestures, postures, and gaze (discussed in Section 2.4.6.) (Oben & Brône, 2015). In sum, from the IAM standpoint, convergence processes are mainly a byproduct of cognitive functioning: an automatic and unconscious phenomenon derived from the way people process information. Convergence is thus seen as a natural and inevitable result of the way in which the human cognitive system processes stimuli (Quezada et al., 2012).

Regarding the weak points of the IAM, following Quezada et al. (2012), understanding accommodation processes as an automatic and unconscious phenomenon implies neglecting affective and social dimensions of the human being. In effect, across conversations information is efficiently processed, but there are also pragmatic, emotional, and contextual factors that influence interlocutors' behaviors. Quezada et al. (2012) conclude that none of these factors is considered in the IAM approach.

[21] According to the *revised version of the motor theory of speech perception* (Galantucci, Fowler & Turvey, 2006), sensorimotor integration between perception and production mechanisms also underlies speech comprehension. This statement is supported, among others, by studies indicating that imitation improves comprehension when the incoming speech signal is unclear (e.g. background noise, unfamiliar accents) (See for instance Adank et al., 2010, and Kappes et al., 2009).

In this respect, Gambi and Pickering (2013) claim that social influence on accommodation can be understood in terms of a higher degree of exposure towards selected social partners. In this scenario, attitudes towards interlocutors only affect accommodation indirectly, depending on the amount of exposure to a particular person or way of speaking. Augmentation of exposure would generate more accurate forward-model predictions of the social partner or group (Ruch et al., 2017). On the other hand, on the latest developments of the IAM vocal accommodation is still understood as not due to intention or as a conversation strategy, but rather as a "by-product of the internal and automatic mechanisms of speech perception and comprehension" (Ruch et al., 2017, p. 3).

Additionally, according to Ruch et al. (2017, p. 4), the IAM "can not provide an explanation for divergent forms of vocal accommodation." In this sense, the automatic perception–production link scenario predicts that an increase in exposure automatically leads to higher levels of convergence. Nonetheless, a decrease of similarity may as well be the result of accommodation (as discussed in Section 2.2.). For instance, in 2006, Werlen and Schlegel (as cited in Ruch et al., 2017) reported that five out of 18 speakers from southern Switzerland who have moved to Berne used less Bernese pronunciation variants two years after moving than shortly afterwards, whereas only two out of 18 showed a clear increase in the use of Bernese pronunciation over time. Schweitzer and Lewandowski (2013), in turn, reported divergence effects in articulation rate in German spontaneous dialogues between unacquainted participants (as mentioned earlier). However, in this study the degree of convergence increased in accordance with the mutual liking scores. Considering this evidence, Ruch et al. (2017) conclude that vocal divergence cannot be explained by a simple perception-production connection and that further control mechanisms must be envisaged.

# 3. Speech rhythm

## 3.1. General

It has been suggested that periodicity of stimuli plays an important role in the perception and processing of sounds (e.g. Falk, Rathcke & Dalla Bella, 2014). However, there is an ongoing debate in the field of linguistics about whether periodicity is present in the acoustic signal, or it is only perceived. For instance, during sentence processing, listeners quickly anticipate details of the acoustic realization of forthcoming speech sounds, including information about projected phonetic characteristics of stressed and unstressed syllables (Brown, Salverda, Dilley & Tanenhaus, 2015). According to Brown et al. (2015), it is likely that perceived speech *rhythmicity* informs listeners about the location and acoustic characteristics of stressed syllables within upcoming words (understanding *rhythmicity* as a regular temporal organization, which in English and many other languages, centers around acoustically prominent or stressed syllables).

In addition, the ability to detect temporal patterns unfolding over time and project such patterns forward to predict upcoming information seems to facilitate stimulus processing in cognitive and perceptual domains (Brown et al., 2015). Rolke and Hofmann (2007), for example, found that the temporal uncertainty of the occurrence of visual events increases reaction times and decreases accuracy when compared to the processing of temporal predictable events. For their part, Jones, Moynihan, MacKenzie and Puente (2002) demonstrated that listeners are more accurate judging the pitch of rhythmically regular tones with respect to rhythmically irregular tones. However, the effect found by Jones et al. (2002) persisted over time (milliseconds), but disappeared when tones with irregular rhythms were interpolated within the sequences (see Brown et al., 2015 for an overview).

Regarding the influence of rhythmic speech properties on accommodation processes, Cummins (2009b) suggests that synchronization in parallel speech signals is unaffected by any modification to the speech that left gross temporal structure and intelligibility relatively intact. Furthermore, accommodation routines would be influenced by the speech rhythm in a yet unknown way, although Cummins (2009a) proposes that rhythm supports

conversational convergence by means of the *affordance*[22]. In any case, to correctly understand the role of speech rhythm in the processes of accommodation, the meaning of rhythm must be clarified first.

In this respect, for near 70 years, and more intensively during the past 30 years, linguists have been trying to establish what speech rhythm is, and how can it be quantified (Harris, 2015). Definition, validity, and underlying mechanisms of speech rhythm are subjects of an ongoing discussion in the academic community. For instance, in the opinion of some authors (e.g. Arvaniti, 2009; Cummins, 2009a), rhythm is not to be found in the acoustic speech signal or in the movement of the person who speaks. Rather, it is probably better understood as the link between the movement and the signal, a property of the act of speaking, and not of the speech signal itself. It is impossible, hence, in this scenario, to determine the core characteristics of the speech rhythm only with the analysis of some acoustic properties of the speech signal.

### 3.1.1. The isochrony hypothesis

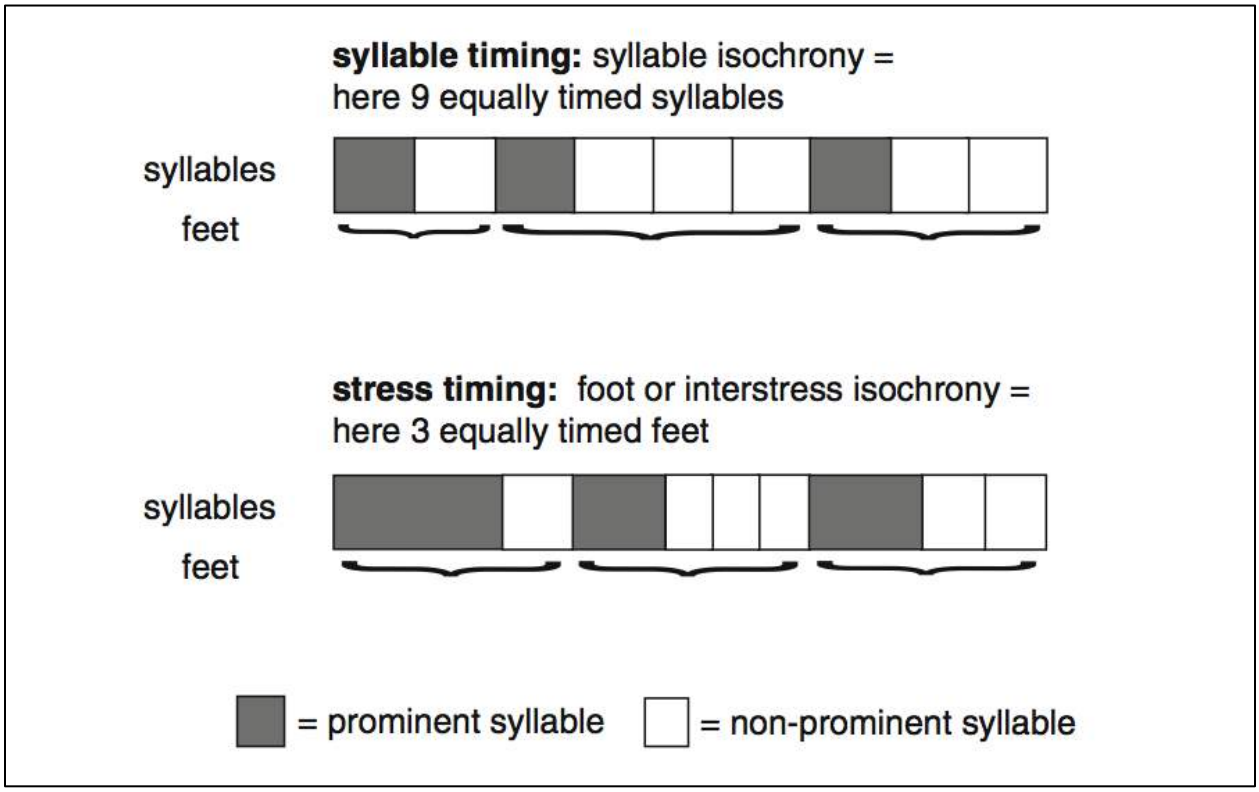
One of the most important theoretical frameworks for understanding speech rhythm is the *isochrony hypothesis* (an overview in Fuchs, 2016). This hypothesis implies that each spoken language possesses stable units of constant duration that occur regularly, making it sound in a particular way. Traditionally, three types of isochrony have been considered (according to the essential rhythmic element present in every language): *stress-timed*, *syllable-timed* and *mora-timed* languages (Gervain & Mehler, 2010). In this context, feet and syllables cannot be both isochronous at the same time, because feet consist of a varying number of syllables (Fuchs, 2016). Therefore, if syllables are isochronous, feet must be of unequal duration, and if feet are isochronous, syllables must differ in duration (see Figure 2).

The isochrony hypothesis predicts that Germanic languages, as stress-timed languages, exhibit equivalent foot lengths and syllabic compression. In contrast, Romance languages

[22] The term *affordance* refers to the properties of an object or environment that allow an organism to perform certain actions. The affordance, thus, implies the relation between the abilities of an organism and the specific properties of its environment with respect to a specific type of action (Cummins, 2009a).

are viewed as syllable-timed, therefore presenting equivalent syllable lengths and poor (or inexistent) syllabic compression (Dauer, 1983).

**Figure 2:** Relation feet-syllables according to the isochrony hypothesis (Fuchs, 2016, p. 36).




The reality of the isochrony hypothesis is supported by three sources of evidence: (1) studies of languages discrimination by infants and neonates, (2) studies of rhythmic classes discrimination (by adults) in degraded speech, and (3) studies indicating that motor control of timing is biologically 'hard coded' in speakers (Harris, 2015). As for the first source, Nazzi, Bertoncini and Mehler (1998) found that although five-day-old infants can differentiate between two unknown languages belonging to different rhythmic categories, they are unable to do it when such two languages belong to the same rhythmic category. In this study, using the *high-amplitude sucking procedure* (see Guasti, 2001), and phrases uttered by several monolingual speakers treated with a cutoff frequency of 400 Hz, it was found that five-day French infants can discriminate between English (stress-timed language) and Japanese (mora-timed language), but they are unable to differentiate

between English and Dutch (both stress-timed languages). See also similar studies conducted on fetuses (Kisilevsky et al., 2009) and a complete review on Gervain and Mehler (2010).

With respect to speech discrimination in degraded conditions, performed by adults, Ramus and Mehler (1999) found that French-speaking adults are capable of discriminate English from Japanese sentences by means only of syllabic rhythm. For arriving at this conclusion, the authors employed resynthesized English and Japanese natural sentences in three different conditions that preserved specific acoustic information: (1) rhythm and intonation, (2) intonation only, and (3) rhythm only.

Contrastingly, White, Mattys and Wiget (2012) propose that sensitivity to speech durational cues, rather than universal rhythmic classes, are responsible for the ability of adult speakers to distinguish between languages. The results of White et al.'s (2012) study showed that adult English listeners are able to distinguish between English and Spanish (different rhythmic classes), and between different accents of British English, when utterances are degraded in a way that only durational characteristics are preserved. In any case, there is no agreement regarding the pertinence of using speech degradation methods during studies of speech rhythm perception (cf. Arvaniti & Ross, 2012).

For the third type of evidence supporting the reality of the isochrony hypothesis, it has been proposed that speech patterns at the segmental level are very difficult to change. Consequently, speakers may modify their speaking rates altogether, but the relative duration of the segments (e.g. words) would remain mostly invariant for a given speech rate (Eriksson & Wretling, 1997; Wretling & Eriksson, 1998; both already discussed in Section 2.4.4.). It has to be noticed, however, that the two papers of Wretling and Eriksson that we have just mentioned are both based on the same experiment with only one participant. Should these results be confirmed in other studies, conclude the authors, "it would be an indication that segmental timing may be more or less hard coded in an individual speaker" (Eriksson & Wretling, 1997, p. 1046).

Additionally, Ramus, Nespor and Mehler (1999), based on the study of eight languages (perceptual and acoustic analyses), conclude that the proportion of vocalic intervals and the variability of intervals between consonants are consistent with the traditional isochrony

classification. These authors believe that rhythm (conceived as the alteration of vowels' and consonants' length) is an essential property that allows the differentiation of languages and the acquisition of the mother tongue. In general, Ramus et al.'s (1999) conclusions support the existence and relevance of rhythmic classes from a perceptual standpoint. Additional (not conclusive) evidence in favor of the reality of isochrony can be found in Ramus, Dupoux and Mehler (2003).

As for the methods employed to establish at which rhythmic class a particular language belongs, different rhythmic metrics have been proposed based on the isochrony hypothesis. Such metrics consist mostly in statistical measures of the duration of vowels, consonants, and syllables, and of the patterns of intervals between them (Lykartsis, Lerch & Weinzierl, 2015). Nonetheless, in the opinion of Lykartsis et al. (2015):

> Although [these rhythmic metrics] have been used extensively for speech rhythm description and the investigation of rhythmical differences between languages, those measures have also been criticized (...) for lack of robustness and for producing inconsistent results with respect to the rhythm class hypothesis, which states that languages belong either to a stress-timed or to a syllable-timed group (...) Further problems include (...) that the focus lies only on high-level language elements (such as syllables or consonants-vowels) and their duration patterns for rhythm description instead of examining directly measurable signal properties. (Lykartsis et al., 2015, p. 1)

Likewise, although the isochrony hypothesis is accepted be several researchers and it is used as a framework of reference and comparison, it has also been widely contested, and predicted effects such as those mentioned above have not been found in a systematic and stable manner among the languages of the world (Almeida, 1993, 1997; Gervain & Mehler, 2010; Lykartsis et al., 2015; Prieto, Vanrell, Astruc, Payne & Post, 2012). Roach (1982), for instance, found more variability with respect to foot duration in stress-timed languages (Russian, English, and Arabic) with respect to syllable-timed ones (French, Telugu, and Yoruba). There have even been reported a few languages that seem to belong to two categories at the same time (Gervain & Mehler, 2010).

## 3.1.2. Alternative views of speech rhythm

Several definitions of speech rhythm have been proposed (some of them already discussed in Section 2.4.4.). According to Arvaniti (2009), for instance, it is a mistake to characterize rhythm as a correlation between acoustic elements extracted from the speech signal. Conversely, rhythm can be seen as a psychological characteristic depending on speech stress patterns, with not significant differences between languages. From this stand, speech rhythm consists in a pattern of periodicities extracted from the durational characteristics of the speech signal (also known as *timing*) by means of a psychological process. Listeners can thus establish rhythm despite differences between speakers and disparities within the signal, detecting at the same time the nuances conveyed by the variations of speech.

Another view of speech rhythm, referred to as *the phonological approach*, implies that perception of different types of rhythm results from the presence or absence of specific phonological and phonetic properties in a particular language (e.g. syllable structure variety and complexity, vowel reduction, and phonetic realization of stress) (Prieto el al., 2012; see also Dauer, 1983). In this respect, Dauer (as cited in Prieto et al., 2012) claimed that:

> The coexistence of a certain set of ... phonological properties is responsible for promoting the perceptual prominence of stressed syllables in relation to other syllables – yielding 'stress-timed' perception – while a different set is responsible for the percept of equal salience between syllables – yielding 'syllable-timed' perception (Prieto et al., 2012, p. 682).

It has also been proposed that languages can be placed along an ideal scale going from perfect *stress-timing* to perfect *syllable-timing* (Hualde, 2012; Kohler, 2009). In this scenario, rhythm in stress-timing languages would be characterized by large differences between duration and intensity of prominent and nonprominent syllables, and great reduction and compression of unstressed syllables. Rhythm in syllable-timing languages, in turn, would correspond to languages in which all syllables have very similar structures.

On the other hand, as believed by Gibbon (2015), there is not such thing as speech rhythm, but speech rhythms, which consist in "temporally regular iterations of events which [sic] embody alternating strong and weak values of an observable parameter" (p.

115). In this definition, an *observable parameter* is a neutral concept between an auditory phenomenon and a set of acoustic factors. Alternating strong and weak values may also be referred to as loud-soft, stressed-unstressed, prominent-nonprominent, and even consonant-vowel. Following this logic, diverse variants of speech rhythms are possible. For instance (examples are taken from Gibbon, 2015, pp. 117-118; stressed syllables are represented by a number *1* and unstressed syllables by a number *0*):

(1) Singlets [1-1-1...]: This - fine - bear - swam - fast - near - Jane's - boat.

(2) Lambs [01-01-01]: And then - a car - arrived.

(3) Trochees [10-10-10]: This is - Johnny's - sofa.

(4) Dactyls [100-100-100]: Jonathan - Appleby - carried it - awkwardly.

(5) Anapests [001-001-001]: It's a shame - that he fell - in the pond.

(6) Amphibrachs [010-010-010...]: A lady - has found it - and Tony - has claimed it.

Based on the aforementioned definition of speech rhythm (or rhythms), proposed by Gibbon (2015), we suggest the following restricted definition in order to be used for the construction and implementation of the experiments conducted in this thesis: *the rhythm of speech consists in a temporally iteration of strong and weak values of lexical stress* (i.e. stressed and unstressed syllables)*, which is constituted, in turn, by an unknown combination of duration, intensity, and pitch* (see Section 3.2.1. for further details of lexical stress implementation in Spanish).

# 3.2. The rhythm of Spanish

## 3.2.1. General facts and stress patterns

More than 350 million people in the world speak Spanish, and it is a national language in Spain and more than a dozen Latin American countries (Beckman, Díaz-Campos, McGory & Morgan, 2002). In his *Manual de pronunciación española* (Handbook of Spanish pronunciation), Navarro-Tomás (2004/1918) argues that differences in pronunciation between Spanish speakers in Spain are deeper and larger with respect to Spanish speakers within Hispanic America. This would be due to a heavy bilingual influence in regions such as Catalonia, Valencia, and Galicia. Moreover, in regions like Aragon, Navarra, and Asturias,

several phonetic traits belonging to ancient dialects would have been incorporated into modern Spanish.

As for the differences between Spanish dialects, and referring specifically to lexical stress, Hualde (2007) states the following: "It is also largely unknown to what extent dialectal variation is found in this respect since, by and large, descriptive work has focused on standard Peninsular Spanish only" (pp. 63-64). On the other hand, several researchers, such as Beckman et al. (2002), assume a wide range of coverage instead of separate analyses for each of the different major dialect variations, aiming to understand the core phenomena shared across Spanish-speakers. In this thesis we adhere to the latter standpoint.

Spanish has been traditionally classified as a syllable-timed language. Its predominant syllable type is CV. It presents low degrees of syllable complexity, practically no signs of vowel reduction, and less final lengthening than English (Mooney & Sullivan, 2015; Prieto et al., 2012). "In Spanish all multisyllabic words have one syllable marked for primary stress.... [And in] around 75% of the words the second-to-last is the primary stressed syllable" (Toro, Rodríguez & Sebastián-Gallés, 2007, pp. 169, 174). In addition, all Spanish variations have lexical contrastive stress (Beckman et al., 2002). For example, the word *nu**me**ro* (I assign numbers) contrast both with ***nú**mero* (number) and with *nume**ró*** (she or he assigned numbers) (Stressed syllables are presented in bold).

Spanish is a free stress language, that is to say, there cannot be predicted on which syllable within a word the stress will fall (Alcoba & Murillo, 1998). Words within an utterance can be classified as stressed or unstressed depending on the presence or absence of a stressed syllable. Words used isolated or metalinguistically become stressed, if it is not already the case (Hualde, 2012).

Spanish words are classified into three groups according to their stress pattern: *oxytones* (final stress; e.g. *ca**mión**, **él**, **hoy*** [truck, he, today]), *paroxytones* (penultimate stress; e.g. ***ga**to, her**ma**na, par**ti**do* [cat, sister, match]) and *proparoxytones* (antepenultimate stress; e.g. ***pá**gina, po**lé**mica, **tí**tulo* [page, polemic, title]) (Stressed syllables are presented in bold). There are also *proparoxytones*, which are in fact compound words comprising a verb form and two clitics added at the end (e.g. ***dí**gaselo* [tell that to him or her] = ***di**ga* + *se* + *lo*;

*can**tán**dotelo* [singing it to you] = *can**tan**do* + *te* + *lo*) or adverbs comprising an adjective form followed by the suffix *-mente* (e.g. ***fran**camente* [frankly]; **rá**pidamente [quickly]) (Ohannesian, 2004; Planas, 2013).

In Spanish, content words (aka lexical words), such as nouns, verbs, adjectives, and adverbs, are always stressed on one of their last three syllables (this principle is known as the *three-syllable window*; Hualde, 2012). Function words (aka grammatical words), instead, may be stressed or unstressed when used within an utterance (Hualde, 2007). For instance, definite articles (e.g. *el, la, los...*), pronominal possessives placed before a noun (e.g. *mi, tu, su...*), and all prepositions except "*según,*" are unstressed. Indefinite articles (e.g. *un, una, alguien...*), demonstratives (e.g. *este, ese, aquel...*), and pronominal possessives not placed before a noun (e.g. *mío, tuyo, suyo...*), are stressed (Hualde, 2009; Real Academia Española, 2014). For a complete list see Real Academia Española (2014).

There are several cases in Spanish in which unstressed words may become stressed or vice versa, including: contrastive focus, nominalization, citation, parentheticals, exclamatory and interrogative sentences, compound words, and appearing after a modifying numeral. There are also words that may be pronounced with two stresses, such as adverbs finishing in *-mente*, which historically derive from compounds (Face, 2003; Hualde, 2007, 2009) (Note that this is a case of *lexical secondary stress*, not be confused with the *rhythmic secondary stress*, discussed in section 3.2.5.; see Ohannesian, 2004 for a review of the differences between them). As explained later, in order to maintain only two levels of prominence (stress / no stress), all the instances mentioned in this paragraph were avoided when creating the experimental task employed in this thesis.

In Spanish, the *nuclear stress* tends to fall on the last word of the utterance; that is to say, speakers tend to perceive the last lexically stressed syllable within an utterance as more prominent than the rest (understanding *nuclear stress* as the most prominent stress in an utterance; as stated in Section 1.1.) (Hualde, 2012). Moreover, according to Beckman et al. (2002), the choice of words order in Spanish seems to be closely related to the metrical organization of the sentence, and, under regular circumstances, sentences final positions are metrically strong.

Regarding stress perception studies in general, Lieberman and Blumstein (1988, p. 154) conclude that "human listeners make seemingly 'simple' stress distinctions by taking into account the total fundamental frequency contour of the utterance, the amplitude of syllabic 'peaks,' the relative durations of segments of the utterance and the range of formant frequency variations." More specifically, and with respect to both production and perception studies, lexical stress in Spanish relies mainly on three acoustic features: *F0*, *duration*, and *intensity*. There are also secondary features related to stress, such as spectral characteristics (timbre) and reduction or elision of unstressed vowels (Candia, Urrutia & Fernández, 2006). In any case, the relative weight of each one of these factors in the perception of lexical stress in Spanish is not easy to establish.

It has been claimed, for instance, that a single unique factor is mostly responsible for lexical stress in Spanish, depending on the author: *F0* (Bolinger & Hodapp, 1961; Quilis, 1981), *duration* (Díaz-Campos, 2000; Garrido, Llisterri, de la Mota & Ríos, 1995; Ortega & Prieto, 2007), or *intensity* (Navarro-Tomás, 2004/1918). In addition, it has been proposed that at least two of these three variables interact to generate lexical stress (Llisterri, Machuca, de la Mota, Riera & Ríos, 2003). Moreover, Solé (1984) found that the interaction of these three variables underlies lexical stress, but that F0 has the fundamental role, and can even generate the stress alone when opposed to duration and intensity. An overview of this subject can be found in Candia et al. (2006); see also Alcoba and Murillo (1998), and Urrutia (2007).

Of course, only studies carried on since the 1990s have benefit from digital techniques applied in phonetics laboratories. For instance, the research of Candia et al. (2006) was conducted on a digital corpus of read-aloud materials uttered by monolingual Spanish speakers of northern Spain. Their results indicate a complex correlation between F0, intensity, and duration underlying lexical stress. However, also according to their results, intensity is more important that F0 and duration for marking lexical stress as well as phrasal stress.

Llisterri et al. (2003), in turn, analyzed isolated meaningful three-syllable words uttered by a native speaker of Castilian Spanish. Recordings of the words were manipulated to establish the role of the three aforementioned acoustic cues in the perception of lexical

stress. According to their results, neither F0 alone, duration alone, nor intensity alone, are enough to identify the stressed syllable within a word. Only the combination of F0 with duration, or intensity, or both, constitutes a relevant acoustic cue for the perception of lexical stress. Nonetheless, according to Ortega and Prieto (2007), in Llisterri et al.'s (2003) study all stressed syllables were also pitch-accented, while unstressed ones were de-accented due to the intonation patterns that they used. In consequence, the results would rather indicate that a pitch movement is the main factor underlying lexical stress in Spanish.

Furthermore, Ortega and Prieto (2007) conducted their own study with educated bilingual speakers of Spanish (L1) and Catalan (L2) from Barcelona. In this study, participants were instructed to utter four-syllable verbs embedded in utterances with either a declarative intonation or a flat intonation. Resulting data showed that stressed syllables have longer durations than unstressed ones, and that such differences in duration are sufficient to distinguish stressed from unstressed syllables. The authors conclude that duration is the main cue to lexical stress in Spanish, and that successful classification of stressed syllables drawn from duration differences occurs independently of the presence or absence of a pitch accent. In accordance with these results, spectrographic analyses of read-aloud speech tokens, uttered in Peninsular Spanish, led Garrido et al. (1995) to conclude that both in reading and spontaneous speech, duration is the fundamental acoustic correlate of lexical stress, when compared to F0 and intensity.

Regarding the relation between pitch accents and lexical stress in Spanish, Hualde (2012) concludes:

> Most words in the discourse bear a pitch accent, and these pitch accents are aligned with respect to the lexically-stressed syllables, which act as designated anchoring points for these intonational elements. We find deviations from this rule in contexts where secondary stresses are employed.... Notice that the observation that stressed syllables tend to bear pitch accents does not mean that stressed syllables will have higher pitch than other syllables. Since the most common pitch accents in Spanish declaratives are rising, with the tonal rise continuing onto the following syllable,

> very often the posttonic will have higher pitch than the stressed syllable. (Hualde, 2012, p. 164)

In accordance with Hualde's (2012) statement above, it is believed that both in isolated sentences and in sentences included in a context, higher pitch values tend to fall in the posttonic syllable (Urrutia, 2007). Likewise, Garrido (2012) agrees with the fact that in Spanish a tonal increase (also referred to as a F0 peak) is often found around the stressed syllable within a word. In the words of Hualde (2007):

> Typically, in a pragmatically neutral declarative sentence, every content word will be (pitch-)accented in Spanish. The [unstressed] function words ... have less prominence than other words in the same phrase in pragmatically neutral contexts ... they are totally unstressed.... In a neutral context, a prepositional phrase such as *para dos* 'for two', for instance, perceptually has the same prominence contour as the single word *parador* 'inn', both with a single stress on the final syllable. (Hualde, 2007, p. 60)

Likewise, as reported by Ortega and Prieto (2007, p 158), "in declarative sentences, stressed syllables ... bear a pitch-accent while unstressed syllables remain de-accented." Face (2003) and Hualde (2009) agree with this statement. According to Hualde (2007):

> In Spanish, in sentences or phrases under neutral focus (uttered in an out-of-the-blue context), all content words receive stress on a lexically-designated syllable ... and will normally be accompanied by a pitch excursion (i.e. will be accented, see, for instance, Sosa 1999), the last of those stresses being perceived as most prominent, that is, as the nuclear or main stress.... In pragmatically neutral sentences all content words will typically be pitch-accented. (Hualde, 2007, p. 62)

We emphasize this matter because, as explained later, for the purposes of the experimental task employed in this thesis, pitch accents and stressed syllables must be ideally aligned.

## 3.2.2. Phonological phrasing

In this section we discuss briefly, from the theoretical stand, the two most studied forms of phonological phrasing in Spanish (excluding the syllable): *accentual feet* and *accentual groups* (in the next Section [3.2.3.] we present the related empirical evidence).

## Accentual feet

Within an utterance, *accentual feet* comprise from the first stressed syllable (or vowel) to the syllable (or sound) immediately behind the next stressed syllable (or vowel) (Almeida, 1993, 1997; Cantero 2002). According to Grau (2013), the minimal foot in Spanish is a monosyllable, as long as it is bimoraic, and any monosyllabic content word forms a foot.

On the other hand, in the opinion of Hualde (2012), since Spanish can have long stretches of unstressed syllables between stressed ones, the notion of accentual foot (in Spanish) is hard to justify. For instance, the following utterance (adapted from Hualde, 2012) has eight unstressed syllables between two stressed ones: *los es.tu.**dian**.tes de nues.tras u.ni.ver.si.**da**.des* (the students of our universities; syllables are separated by a period and stressed syllables are presented in bold). Note that this example is not atypical and can be found easily in everyday speech.

## Accentual groups

*Accentual groups* have been studied by several researchers of the Spanish language, sometimes using different terms for referring to it. Almeida (1993, 1997), for instance, defines an accentual group as a lexical unit together with all the grammatical words related to it. Mora, Villamizar, Blondet and López (1999) agree with this definition and propose the next example (synalephas are marked with an asterisk [*]; stressed syllables are represented by a number *1* and unstressed syllables by a number *0*; accentual groups are shown in brackets):

(1) El viento norte y el sol (*The northern wind and the Sun*)

(2) [El viento] [norte] [y el sol]

(3) [010] [01] [0*01]

Note that, when establishing accentual groups, two syllables linked by a synalepha count as one.

In addition, as showed by Alcoba (2007) in the next example, accentual groups may also be constituted by a function word such as the indefinite article "*un"*, alone, or with other words. Note also that in Spanish indefinite articles are stressed whereas definite articles are unstressed (Real Academia Española, 2014).

(1) En un lugar de la Mancha vivía un caballero (*A gentlemen lived in a place in la Mancha*)

(2) [En un] [lugar] [de la Mancha] [vivía] [un] [caballero]

(3) [01] [01] [0010] [010] [1] [0010]

For his part, Cantero (2002, 2003) uses the terms *rhythmic group* and *phonic word* (*palabra fónica*) to refer to the same phrasing principle. In his terms, a rhythmic group (or phonic word) is constituted by one stressed word and all the grammatical unstressed elements that are pronounced together with it. Toledo (1994) also uses the term *rhythmic group* to refer to the same linguistic arrangement. We offer an example based on the one provided by Planas (2013, p. 68) (Stressed syllables are represented by a number *1* and unstressed syllables by a number *0*; rhythmic groups are shown in brackets):

(1) Ellas comerán el pescado por la noche (*They will eat the fish at night*)

(2) [Ellas] [comerán] [el pescado] [por la noche]

(3) [10] [001] [0010] [0010]

Hualde and Nadeu (2014), in turn, refer to this type of phrasing principle as *phonetic group*, which they define as a morphological word and any unstressed functional words preceding it. For instance:

(1a) Amistad (*Friendship*)

(2a) [001]

(1b) Abadesa (*Abbess*)

(2b) [0010]

(1c) Contra lo tratado (*Against the agreement*)

(2c) [000010]

On the other hand, it is worth noting that the limits of an accentual group may vary between speakers and between renditions, therefore it is not possible to predict them with

certainty (Cantero, 2002; Toledo, 1994). Cantero (2002, p. 53) offers the following utterance as an example, with two possible compositions of the resulting accentual groups (examples are adapted to our scheme of presentation):

(1) La casa estaba vacía y triste (*The house was empty and desolate*)

(2a) [La casa] [estaba] [vacía] [y triste]

(3a) [010] [010] [010] [010]

(2b) [La casa es] [taba] [vacía y] [triste]

(3b) [0100] [10] [010*0] [10]

One more example, this time with three possible compositions of the resulting accentual groups (Cantero, 2002, p. 53):

(1) Bueno, ese es mi primo (*Well, that is my cousin*)

(2a) [Bueno e] [se es] [mi primo]

(3a) [10*1] [0*1] [010]

(2b) [Bueno] [ese es] [mi primo]

(3b) [10] [10*0] [010]

(2c) [Bueno ese es] [mi primo]

(3c) [10*10*0] [010]

In the preceding examples, Cantero (2002) envisages even three stressed syllables within a single cluster, which is contrary to the definitions of accentual group that we have presented. Presumably, in such type of rendition only one syllable would be actually stressed. Additionally, in these examples a single word is separated into two different clusters, as may be the case in accentual feet. Please note also that some of the aforementioned possible arrangements of accentual groups are subject to issues related to resyllabification and *sirremas* (both subjects are discussed later).

Finally, the term accentual group must not be confused with the term *phonic group* (*grupo fónico*), proposed by Quilis (1993). According to this author, the phonic group is a discourse segment comprised between two pauses, consisting of one or more rhythmic groups arranged around a phrasal stress. Moreover, the definition and examples of accentual group that we have discussed here are quite similar, if not the same, to the

phonological units termed *phonological words*, *prosodic words*, or *pwords* (see Aronoff & Fudeman, 2011).

### 3.2.3. Rhythmic classification of Spanish

Following authors such as Pike, Abercrombie, and Hockett, Spanish has been traditionally considered as a syllable-timed language (Fuchs, 2016; Prieto et al., 2012; Toledo, 2009). In contrast, the works of Navarro-Tomás in the first half of the twentieth century indicate that duration of syllables in Spanish is uneven, whereas accentuated feet tends to remain stable (see Navarro-Tomás, 2004/1918; an overview in Pamies, 1999). Furthermore, after analyzing the preexistent experimental work on rhythm of spoken Peninsular Spanish, Pointon (1980) concludes that Spanish is neither stress-timed nor syllable-timed, given that its rhythmic patterns' length depend on the phonetic context. In the words of the author himself: "Spanish has no regular rhythm in the sense of an isochronous sequence of similar events, be they syllables or stresses" (Pointon, 1980, p. 302).

In addition, several researchers have analyzed the rhythm of Spanish, mostly from the isochrony hypothesis standpoint. For instance: Almeida (1997), Toledo (1994, 1989), Dauer (1983), Borzone and Signorini (1983), Mora et al. (1999), and Pamies (1999). In the following we comment briefly the most relevant results of these works.

In an early study, Dauer (1983) analyzed feet between one and six syllables in Castilian and Cuban Spanish. The results showed a positive correlation between number of syllables and feet's length. However, no evidence was found of Spanish belonging to the syllable-timed or stress-timed categories. These results are partially consistent with the research conducted by Toledo in the 1980s (see Toledo 2010a and 2010b for an overview). In these studies, the author reports accentual isochrony between small feet (1 and 2 syllables), but not between larger feet (3 to 5 syllables), in texts read by an Argentinian speaker. Toledo (2010a, 2010b) also found a general tendency for syllabic isochrony, independent of feet's length, in texts read by a Colombian speaker, and both syllabic and accentual isochrony tendencies in texts read by a Cuban speaker and in Argentinian spontaneous speech.

For their part, Borzone and Signorini (as cited in Toledo, 1994, 2010b) analyzed the Spanish spoken in Buenos Aires, finding a significant degree of isochrony regarding inter-stress intervals, and an average feet length between 447 and 467 ms. A variable amount of time regarding syllables' length was also reported (Prieto et al., 2012; Toledo, 2009). Taken together, these data suggest that the Spanish spoken in Buenos Aires resembles more a stress-timed language than a syllable-timed one. On the contrary, Mora et al. (1999) suggest that in the city of Buenos Aires, Spanish tends to present a syllabic rhythm, whereas in the region of Tucumán the tendency is to an accentual rhythm.

Moreover, in a series of studies, Pamies (1999) measured the length of feet up to six syllables in Spanish sentences, which were uttered by speakers of different regions of Spain and Argentina. The author found notable inequalities between feet's durations, including feet three or four times longer than other feet within the same sentence. Additionally, Pamies' (1999) results suggest a positive correlation between feet length and number of syllables per foot.

Regarding Canarian Spanish, Almeida (1997) reports a stronger tendency to temporal regularity in syllables and accentual groups compared to accentual feet. The author also reports the existence of feet between one and five syllables, with an average length of 373.2 ms and a high standard deviation of 148.2 ms.

It has also been found that in Venezuelan Spanish, accentual groups exhibit a greater temporal regularity than accentual feet and syllables (Mora et al., 1999). In this study, four recordings of Venezuelan speakers coming from different regions and reading the same text were acoustically analyzed. The results also showed that stressed syllables were significantly longer that unstressed ones. Interestingly, Mora et al. (1999) analyzed accentual feet with head to the left, as it is usually done, but also with head to the right, without noticing a pattern of temporal regularity in both cases.

Regarding South American Spanish in general, Toledo (1988) reports the existence of both accentual and syllabic isochrony. The author also reported, in line with the results of Mora et al. (1999), that accentual groups are more isochronous than syllables and accentual feet.

Taken together, the aforementioned studies do not present a clear trend towards a rhythmic classification of Spanish, at least within the isochrony hypothesis framework. In this respect, according to Gil & Llisterri (2004), even during a discourse uttered in Spanish, a speaker may alternate between isochronous and unequal sequences, of both feet and accentual groups. Moreover, drawn form phonological analysis of poetic discourse, Flores (Flores, 2004; Flores & Horne, 2003) concludes that Spanish has stress-timed tendencies and can be characterized as a language with a mixed rhythm structure.

Likewise, after more than 30 years of research in Spanish phonetics, Guillermo Toledo concludes that Spanish is in principle a language between the syllable-timed and stress-timed categories (personal communication, April 30, 2016). The author also indicates that research on this topic may be influenced by several factors, such as social and linguistic context of the studied sample, stylistic conventions, and type of speech. For instance, Toledo explains, cultured speech tends to be syllable-timed whereas popular speech tends to be stress-timed (see Toledo, 1989, 1994, 2009, 2010a, 2010b). How variations between syllable-timed and stress-timed rhythms can take place in the same language is beyond the scope of this thesis, but a possible explanation may be found in Gibbon (2015).

## 3.2.4. Resyllabification and sirrema

### Resyllabification

The term *resyllabification* implies the division of words into syllables in a different way with respect to the standard division of the isolated word. It happens automatically and unconsciously under special circumstances (listed below), but it may not happen when emphasizing a word, making a pause within an utterance, or when ambiguity can result from it (Hualde, 2014; Navarro-Tomás, 2004/1918; Ulloa, 2011). In Spanish, resyllabification may occur in three different cases (almost all examples are adapted from Planas, 2013; syllables are separated by a period):

1. Syllabic contraction resulting from the merging of the vowel at the end of one word and the vowel at the beginning of the following word (also known as a *synalepha*). This could result in:

1a. Reduction to a single vowel (e.g. *una animalada* [one stupidity] = u.na.ni.ma.la.da).

1b. Diphthong formation (e.g. *la imaginación* [the imagination]) = lai.ma.gi.na.ción).

1c. Triphthong formation (e.g. *imperio inglés* [English empire] = im.pe.rioin.glés).

If the contraction occurs within a word, it is known as a *syneresis* (e.g. *poeta* [poet] poe.ta).

2. Syllabic contraction resulting from the merging of the consonant at the end of one word and the same consonant at the beginning of the following word (e.g. *ciudad dormitorio* [Commuter town] = ciu.da.dor.mi.to.rio).

3. Resyllabification resulting from the merging of the consonant at the end of one word and the vowel at the beginning of the following word (e.g. *el oso* [the bear] = e.lo.so) or a decreasing diphthong at the beginning of the following word (e.g. *con autoridad* [with authority] = co.nau.to.ri.dad). Nonetheless, resyllabification does not occur when an increasing diphthong is located at the beginning of the second word (e.g. *quieren huevos* [they want eggs] = quie.ren.ue.vos).

## Sirrema

As proposed by Antonio Quilis, a *sirrema* is a group of two or more words that constitutes a grammatical, melodic, and semantic unit. Under normal circumstances there is not pause between the words composing a sirrema when they are being uttered (Quilis, 1993; Spang, 1983). The following combinations are proposed to form a *sirrema*:

1. Article + noun (e.g. *el perro* [the dog]).

2. Noun + adjective (and vice versa) (e.g. *perro blanco, blanco perro* [white dog]).

3. Noun + determiner (e.g. *el perro de Juan* [John's dog]).

4. Verb / adjective / adverb + adverb (e.g. *ven aquí* [come here], *muy inteligente* [very intelligent], *demasiado tarde* [too late]).

5. Conjunctions and the element they connect to (e.g. *y Pedro* [and Peter]).

6. Prepositions and the element they connect to (e.g. *con Juan* [with John]).

7. Unstressed pronouns and the element they connect to (e.g. *te lo dije* [I told you so]).

8. Compound tenses (e.g. *habían sido advertidos* [they have been warned]).

9. Elements constituting a periphrasis (e.g. *iba a llamarte* [I was going to call you]).

Note that combinations such as number two and three are contrary to the phrasing of accentual groups, given that both the noun and the adjective (#2), and the two nouns (#3), are lexical stressed words, which should not be phrased together in the same group.

## 3.2.5. (Rhythmic) Secondary stress

(Not to be confused with the *lexical secondary stress*; see section 3.2.1.)

Despite how words are combined within an utterance in Spanish, they maintain the stress in the same syllable in which they have it individually. However, there may be differences in the degree of intensity carried by the different stressed syllables within an utterance, resulting in a nuclear (or main) stress, which depends on contextual factors (Navarro-Tomás, 2004/1918).

Moreover, it has been proposed that, in Spanish, a secondary stress may be present within an accentual group. According to Navarro-Tomás (in the words of Hualde, 2010):

> Spanish secondary stress is essentially a rhythmic phenomenon operating at the level of prosodic words, which include proclitics such as definite articles, prepositions and conjunctions. [Navarro-Tomás] states that, generally, secondary stress falls on alternating syllables from the lexically stressed or *tonic* syllable. (Hualde, 2010, p. 11)

In fact, Navarro-Tomás (2004/1918) proposes that Spanish speakers perceive an increase-decrease effect in accentual groups of more than two syllables. When one, or two, secondary stresses are present within an accentual group, the intensity responsible for the syllabic prominence would be lower than the one of the syllables bearing lexical stress, but higher with respect to the rest of unstressed syllables (note that for Navarro-Tomás intensity is the main factor underlying lexical stress). The next examples help to illustrate the general alternating patterns of stress levels within an accentual group. All of them are adapted from Navarro-Tomás (2004/1918) (Syllables are separated by a period; primary, lexical stress, is represented by a number *1* and in bold within accentual groups; secondary stressed syllables are represented by a number *2* and are underlined within accentual groups; unstressed syllables are represented by a number *0*):

(1) 2 - 0 - 1 = <u>re</u>.pe.***tir***, <u>com</u>.pa.***rar***, <u>a</u>.mis.***tad*** (to repeat, to compare, friendship).

(2) 1 - 0 - 2 = ***rá.****pi.*<u>*do*</u>*,* ***tí.****mi.*<u>*do*</u>*,* ***pá.****ni.*<u>*co*</u> (quick, shy, panic).

(3) 0 - 1 - 0 - 2 = *re.****tó.****ri.*<u>*ca*</u>*, fo.****né.****ti.*<u>*ca*</u>*, la* ***mú.****si.*<u>*ca*</u> (rhetoric, phonetics, the music).

(4) 2 - 0 - 1 - 0 = <u>*ca*</u>*.ri.****ño.****so,* <u>*la*</u> *ma.****ña.****na,* <u>*en*</u>*.tre* ***to.****dos* (affectionate, the morning, among all).

(5) 2 - 0 - 2 - 0 - 1 - 0 = <u>*con*</u>*.tra.*<u>*pro*</u>*.du.****cen.****te,* <u>*lo*</u> *que* <u>*pro*</u>*.me.****tie.****ron,* <u>*con*</u>*.tra* <u>*lo*</u> *tra.****ta.****do* (counterproductive, what they promised, against the agreement).

However, this alternation between primary and secondary stressed syllables would have an exception: "In groups of four or five syllables with primary stress on the fourth, the secondary stress does not fall on the second syllable as would be expected from the alternation principle, but on the first" (Navarro-Tomás; as cited in Hualde, 2010, p. 11). A couple of examples to illustrate the situation (adapted from Hualde, 2010):

(6) 2 - 0 - 0 - 1 = <u>*em*</u>*.pe.ra.****dor****,* <u>*con*</u>*.ver.sa.****ción****,* <u>*re*</u>*.con.quis.****tar*** (emperor, conversation, reconquer).

(7) 2 - 0 - 0 - 1 - 0 = <u>*ex*</u>*.pli.ca.****cio.****nes,* <u>*por*</u> *la ma.****ña.****na,* <u>*en*</u> *la co.****rrien.****te* (explanations, in the morning, in the stream).

Hualde (2010) summarizes Navarro-Tomás' proposal for secondary stress placement in three constraints that must be satisfied in a hierarchical order (being the number one the most important): (1) *no stress clash*: stresses may not fall in adjacent syllables; (2) *initial stress*: assign a secondary stress to the initial syllable; (3) *alternating stress*: assign stress to every other syllable counting for the primary (*clashes* and *gaps*, or *lapses*, are discussed later).

In Hualde's (2010) opinion, most researchers after Navarro-Tomás have also identified "either alternating syllables from the tonic or the word-initial syllable as (potential) locations of secondary stress, subject to the non-adjacency condition; [and] several other authors have made explicit statements regarding avoidance of clash as an overriding constraint" (Hualde, 2010, p. 12). That is indeed the case of James Harris and Iggy Roca (reviewed in Díaz-Campos, 2000).

In any case, the prohibition against stresses on adjacent syllables (lexical stresses, not pitch accents) would apply only to secondary stresses, because lexical stresses may occur

on adjacent syllables, both across word boundaries (e.g. ***sé có.****mo* [I know how]) and in compound words (*so.****fá-ca.****ma****,*** *nor.****mal.men.****te* [sofa bed, normally]). Consequently, Hualde (2010) proposes a fourth constraint that would outrank the three aforementioned Navarro-Tomás' constraints: *lexical stresses must be preserved*.

On the other hand, some authors find that rules for determining Spanish secondary stress location are unclear (e.g. Stockwell, Bowen & Silva, 1956; Quilis, 1993; Wallis, 1951). It have even been claimed that no rule can be established regarding such matter. Bolinger (1962), for instance, analyzed secondary stress placement in American Spanish (Spanish spoken in the Americas), concluding that the cases discussed by Navarro-Tomás (2004/1918) are just a fraction of the possible universe, and that the location of secondary stress within words vary between speakers and even in the same speaker in different realizations.

In fact, there is a fundamental disagreement with respect to Spanish secondary stress patterns: some authors claim that this type of stress seems to characterize all phrases, independent of the register (Navarro-Tomás, 2004/1918; Roca, 1986; Stockwell et al., 1956; in these texts, secondary stress may be referred to with other names, such as *rhythmic stress* or *medial stress*). Wallis (1951), in turn, analyzed natural speech tokens of highly educated Mexicans living in or near Mexico City, concluding that Spanish secondary stress occurs in an apparently automatic manner only in oratory or narrative style. Other authors, such as Wallis (1951), and more recently Hualde (2007, 2010), claim that secondary stress is an optional phenomenon used for rhetorical purposes, but generally absent in less emphatic styles.

More specifically, Hualde (2007, 2010, 2012) concludes that secondary stress in Spanish is a rhythmic optional phenomenon operating at the level of accentual groups (prosodic words, in his terms) and used for rhetorical purposes. It may be frequent in oratorical style but is generally absent in less emphatic registers. Secondary stress would thus not appear in conversational speech or experimental reading materials, but in public discourses, lectures, broadcastings, and such types of speech.

Another example of empirical evidence regarding Spanish secondary stress can be found in Hualde (2010). In this study, in which a reading task was employed, native speakers of

Peninsular Spanish did not avoid stress clash in their renditions, systematically placing secondary stress on the initial syllable of words with lexical stress on the second one. Furthermore, prominence was realized in different ways on the primary and secondary stress: secondary stressed syllables exhibited a pitch accent whereas lexically stressed ones exhibited durational stress. In any case, it should be noted that participants in this study had to read three types of materials recorded by Hualde himself; two of these materials, the ones that resulted in the stress clash rendition by the participants, were purposely given stress in the prepretonic syllable.

In addition, according to Hualde (2007, p. 79), "experimental work has generally failed to confirm the existence of rhythmic secondary stress in Spanish; that is, no clear phonetic correlates distinguishing secondarily-stressed from unstressed syllables have been identified." For example, Díaz-Campos' (2000) data on Peninsular Spanish does not support the idea that duration, F0, or intensity can mark the difference between secondary stressed syllables and primarily stressed or unstressed ones. Likewise, Prieto and van Santen (1996), and Scharf, Hertrich, Roca and Dogil (1995), did not find acoustic correlates of secondary stress in Spanish. However, Piña and Díaz-Campos (2005) affirm that there is a problem with the pitch measurement technique used in the Díaz-Campos' (2000) study that invalidates resulting data. Piña and Díaz-Campos (2005) conducted thus a study with speakers of Peninsular Spanish. Preliminary findings of this research indicate that duration and pitch are indeed phonetic correlates of secondary stress in Spanish. Unfortunately, to our best knowledge, no further reports of this particular study have been made so far.

More recently, Hualde and Nadeu (2014) proposed that secondary stress patterns in Spanish rely on the *rhetorical stress*, which is "the optional appearance of readily identifiable stress prominence in certain speech styles on syllables that are not lexically designated to carry stress" (Hualde & Nadeu, 2014, p. 229). Such rhetorical stress would be a feature of public speech (e.g. radio and television announcers, speech of politicians, preachers, and lecturers) intended to mark the whole discourse rather than specific words. Additionally, the rhetorical stress would be commonly placed two syllables before the lexically stressed syllable in words with two and three pretonic syllables, or on the initial syllable of words with lexical stress on the second one, generating thus an stress clash

(Hualde, 2010). Moreover, rhetorical stress, also referred to as *overstressing, emphatic stress,* and *insistence stress,* would be normally absent in conversational speech, although it could be part of the personal style of some speakers.

Following this logic, secondary stress placement in Spanish would follow a rule such as the following "add prominence (i.e. a pitch accent) two syllables before the lexical stress. If there is only one syllable before the lexical stress, add prominence to that syllable" (Hualde & Nadeu, 2014, p. 245). This rule would generate, in some occasions, secondary stress patterns different of those proposed by other authors, but would coincide in the case of words with two pretonic syllables. We exemplify with accentual groups of three, four, and five syllables (examples are adapted from Hualde & Nadeu, 2014; notation as usual; HU stands for Hualde, NT for Navarro-Tomás, and HA for Harris):

(Three syllables) HU 2 - 1 - 0 / NT 0 - 1 - 0 / HA 0 - 1 - 0

(Four syllables) HU 2 - 0 - 1 - 0 / NT 2 - 0 - 1 - 0 / HA 2 - 0 - 1 - 0

(Five syllables) HU 0 - 2 - 0 - 1 - 0 / NT 2 - 0 - 0 - 1 - 0 / HA 2 - 0 - 0 - 1 - 0 or 0 - 2 - 0 - 1 - 0

# 3.2.6. Rhythmicity and rhythmic alternation principle

## Rhythmicity

As cited in Shih, Grafmiller, Futrell and Bresnan (2015, p. 403), Abercrombie defines *rhythmicity* as "the periodic occurrence of some sort of movement, [which produces] an expectation that the regularity of succession will continue." On this basis, Shih et al. (2015) argue that speakers optimize their utterances seeking a fundamental contrast between stressed and unstressed syllables, which is–"one of the most desired rhythmic states in language" (Shih et al., 2015, p. 403). This desired rhythmic state would be present, for instance, in an utterance in which exactly one unstressed (weak) syllable appears between each stressed syllable (e.g. *ellos toman vino* [they drink wine] = 101010; notation as usual).

At this regard, it has been proposed that addition of secondary stresses in public speech in Spanish creates greater rhythmic regularity (with respect to informal speech), improving thus the communicative function of the speaker (Hualde, 2012). Additionally, rhythmic regularity in speech is thought to facilitate information processing (Kohler 2009).

## Rhythmic alternation principle

The *rhythmic alternation principle* (or *alternation principle*, or *rhythmic alternating stress principle*) states that, in order to maintain the equal distribution of stress, languages avoid production of sequences with two or more consecutives syllables with the same degree of articulatory tension: [+strong] [+strong] (known as *syllable clash*) or [+weak] [+weak] (known as *syllable gap* or *lapse*) (Almeida, 1993). Alternation could occur in words, syntagms, and phrases. In syntagms and phrases a reordering of primary and secondary stresses of individual words could take place to avoid the clash or gap. In the words of Elizabeth Selkirk (as cited in Shih et al., 2015, p. 403), this principle is "a sort of Platonic ideal to which the rhythmic structure, grounded in syllables, tones, and syntactic structure, aspires."

Nevertheless, Almeida (1993) and Toledo (1989) did not find evidence of clash or gap avoidance at phrase level in both Canarian and American Spanish. Almeida (1993) concludes that, at phrase level, the rhythmic alternation principle is not an attribute of Spanish.

In any case, not all speech is equally rhythmic and speakers may talk in a more or less rhythmic fashion depending on the context. Moreover, the way to make speech more or less rhythmic varies among languages (Hualde, 2012). "In English, greater rhythmicity is obtained by the temporal spacing of lexically stressed syllables.... In Spanish, rhythm is created by assigning pitch accents [secondary stresses] to lexically unstressed syllables in a regular fashion" (Hualde, 2012, p. 167). In effect, Hualde proposes that the rhetorical stress (which underlies secondary stress patterns) can be characterized "as the anchoring of a pitch-accent on a syllable preceding the lexical stress ... whereas the lexically stressed syllable retains durational cues of prominence" (Hualde & Nadeu, 2014, p. 244).

# 4. Experiments

# 4.1. Hypotheses

## Expected results

Given that the experimental task implemented in this thesis was applied to both mixed-gender and same-gender dyads (female / female and male / male), and also involves reading and repeating utterances, we expect to find some gender effects, and some reading / repeating effects, which have been reported in the academic literature.

*1A. Considering that:*

• According to Traunmüller and Eriksson (1994), and Simpson (2009), the average pitch and the pitch range of a person's voice are normally measured in hertz (e.g. Kousidis et al., 2009). However, it is also common, from a perceptual point view, to conduct the measurement in hertz, and then convert the data into semitones (Pépiot, 2014). In the words of Simpson (2009, p. 623):

> The way in which the ear analyses sound, is non-linear. Put simply, what we perceive as equal jumps in pitch, are physically different. The higher the pitch becomes the greater the physical difference between two tones has to be for the perceived difference to remain the same. For this reason, Hertz measurements are often converted into other units of measurement that more appropriately reflect our perception of frequency.

• In general terms, measurements in hertz indicate that the average F0 is higher in women than in men, and that the F0 range is broader in women than in men[23]. However, when these data are converted from hertz to semitones the results exhibit inconsistent patterns (especially those related to the range of F0) (Chen, 2007; Hudson & Holbrook, 1982; Pépiot, 2014; Traunmüller & Eriksson, 1994). For example, it has been reported that

[23] Distinctions between sexes related to F0 may be related to several factors, including: (a) vocal fold anatomy and physiology configurations, (b) racial types, (c) sociophonetic differences, and (d) practices such as smoking. These topics are discussed in detail in Chen (2007), and Simpson (2009).

men's F0 ranges are broader than women's F0 ranges, and also that they are roughly equal (Hudson & Holbrook, 1982; Simpson, 2009 and references therein).

• Reading F0 is higher than spontaneous speaking F0 (we assume that it is also higher than repeating F0) in both men and women, measured in both hertz and tones (Hudson & Holbrook, 1982; Horii, 1982; Mysak, 1959; Snidecor, 1943).

• Reading has a F0 range broader than speaking (we assume that it has also a F0 range broader than repeating) in both men and women, measured in tones (we assume that also measured in hertz) (Mysak, 1959; Snidecor, 1943; cf. Hudson & Holbrook, 1982 and references therein) (Note that although Hudson & Holbrook [1982] mention a few studies in line with this statement, actually, their study showed the opposite results, i.e. reading exhibited a F0 range narrower than speaking).

*We expect that***:**

(*E* stands for *expectation*)

• **$E_1$:** Average F0, measured in hertz, will be higher in women than in men**.**

• **$E_2$:** F0 range, measured in hertz, will be broader in women than in men.

• **$E_3$:** Average F0, measured in hertz, will be higher in reading than in repeating, in both men and women.

• **$E_4$:** F0 range, measured in hertz, will be broader in reading than in repeating, in both men and women.

*1B. Considering that:*

• Speech rate (aka speaking rate) can be measured in several ways, including syllables per second and words per minute (Schultz et al., 2015). These two types of calculations are influenced by pauses within utterances and speech errors. On the other hand, speech rate can be measured in terms of articulation rate (e.g. Winter & Grawunder, 2012). However, as mentioned earlier, this type of calculation (articulation rate) eliminates pauses and silences, removing thus important temporal information that contributes to speakers' speech rate and prosody (Schultz et al., 2015). Speech rate has also been measured in terms of mean word duration (Pépiot, 2014).

• In general terms, men tend to speak slightly faster than women during spontaneous, telephonic, and read speech (Binnenpoorte, Van Bael, den Os & Boves, 2005; Byrd, 1994; Cohen et al., 2017; Weirich & Simpson, 2014; Yuan, Liberman & Cieri, 2006). However, several authors have reported no significant cross-gender differences related to the speech rate (Pépiot, 2014 and references therein), both for spontaneous speaking and reading (Lass & Sandusky, 1971). For instance, no gender differences in speaking rate or articulation rate were found for New Zealand English or American English (Robb, Maclagan & Chen, 2004). Note that in several of the studies that we have just mentioned, aspects such as resyllabification and sentence duration played an important role in the results (as will be described later, these factors were controlled in the experimental task employed in this thesis).

• Spontaneous speaking is slower than reading in both women and men (Lass & Sandusky, 1971; Snidecor, 1943), and listening is (slightly) slower than reading (Rayner & Clifton, 2009), therefore it is possible to assume than repeating an utterance is also slower (to a certain amount) than reading.

*We expect that:*

• **$E_5$:** Speech rate, measured in syllables per second, will be faster in men than in women, in both modes of rendition (reading / repeating).

• **$E_6$:** Speech rate, measured in syllables per second, will be faster in reading than in repeating, in both men and women.

## Hypothesized results

Given that one of the aims of this thesis is to help to determine the influence of sentence-level rhythmic regularity and phonological phrasing on linguistic accommodation during conversational interactions, we propose the following hypotheses:

*2A. Considering that:*

• Listeners extract beats from their interlocutors' utterances during conversational interactions, and those perceived beats affect in turn the listeners' vocal productions when it is their turn to speak (Schultz et al., 2015).

• Behavioral synchronization between individuals requires an external timing signal, which can be derived from partners or from other sources (Merker et al., 2009). Moreover,

> [Such signal] must allow the time lag between stimulus and response introduced by sensorimotor reaction time to be eliminated through predictive timing. For this, a signal consisting of a repeating unit duration affords unique informational economy by making the very next beat in the sequence perfectly predictable (Merker et al., 2009, p. 5).

• "Aperiodic utterances are processed in a way that somehow involves their reinterpretation as periodic" (Mooney & Sullivan, 2015, p. 130). Consequently, periodic utterances are arguably processed faster than aperiodic utterances.

• "Listeners are faster to process information within syllables that are expected to bear stress based on intonational or metrical patterns in the preceding context" (Brown et al., 2015, p. 4).

• Repetitions of a speech stimulus can reduce the time and neural activation needed for its processing, and multiple repetitions significantly enhance memory and learning (Falk et al., 2014).

• Späth et al. (2016, discussed in Section 2.4.4.) found that speech rhythm resemblance is greater in sentences with a metrically regular structure with respect to sentences with an irregular structure.

• According to the interactive alignment model (Section 2.6.2.), as interlocutors align their linguistic representations, other linguistic behaviors also align in form (phonetically, syntactically, and semantically) (Louwerse et al., 2012; Ward & Litman, 2007).

*We hypothesize that:*

• **$H_1$:** Rhythmic distance between speakers will be closer during conversational interactions involving regular rhythmic sentences compared to interactions involving irregular rhythmic sentences.

• **$H_2$:** Interval time (delay) between hearing and repeating an utterance will be shorter during conversational interactions involving regular rhythmic sentences compared to interactions involving irregular rhythmic sentences.

• **$H_3$:** F0 range distance between speakers will be closer during conversational interactions involving regular rhythmic sentences compared to interactions involving irregular rhythmic sentences.

• **$H_4$:** F0 mean distance between speakers will be closer during conversational interactions involving regular rhythmic sentences compared to interactions involving irregular rhythmic sentences.

• **$H_5$:** Speech rate distance between speakers will be closer during conversational interactions involving regular rhythmic sentences compared to interactions involving irregular rhythmic sentences.

• **$H_6$:** Lexical repetitions will be greater in number during conversational interactions involving regular rhythmic sentences compared to interactions involving irregular rhythmic sentences.

*2B. Considering that* (together with the considerations made in 2A)**:**

• In general terms, different units of phonological phrasing have been proposed as responsible for accommodation behaviors, including: prosodic feet (Goswami & Leong, 2013), syllables (Schultz et al., 2015, referring to speech rate accommodation), and pitch accents (Couper-Kuhlen, 1993, referring to turn-taking behavior).

• During a discourse uttered in Spanish, a speaker may alternate between isochronous and unequal sequences, of both feet and accentual groups (Gil & Llisterri, 2004).

• Regarding Canarian Spanish, Almeida (1997) reports a stronger tendency to temporal regularity in syllables and accentual groups compared to accentual feet.

• In Venezuelan Spanish accentual groups exhibit a greater temporal regularity than accentual feet and syllables (Mora et al., 1999).

• Regarding South American Spanish in general, Toledo (1988) reports that accentual groups are more isochronous than syllables and accentual feet.

*We hypothesize that:*

• **$H_7$:** Rhythmic distance between speakers will be closer during conversational interactions involving sentences arranged in accentual groups compared to interactions involving sentences arranged in accentual feet.

• **$H_8$:** Interval time (delay) between hearing and repeating an utterance will be shorter during conversational interactions involving sentences arranged in accentual groups compared to interactions involving sentences arranged in accentual feet.

• **$H_9$:** F0 range distance between speakers will be closer during conversational interactions involving sentences arranged in accentual groups compared to interactions involving sentences arranged in accentual feet.

• **$H_{10}$:** F0 mean distance between speakers will be closer during conversational interactions involving sentences arranged in accentual groups compared to interactions involving sentences arranged in accentual feet.

• **$H_{11}$:** Speech rate distance between speakers will be closer during conversational interactions involving sentences arranged in accentual groups compared to interactions involving sentences arranged in accentual feet.

• **$H_{12}$:** Lexical repetitions will be greater in number during conversational interactions involving sentences arranged in accentual groups compared to interactions involving sentences arranged in accentual feet.

## Possible results

Given that, to the best of our knowledge, there are no studies in Spanish comparing the effects of employing regular versus irregular rhythmic structures, and accentual feet versus accentual group phrasing, on accommodation between interlocutors, the experiments presented in this thesis function as an exploratory study regarding the following subjects:

• **Results related to F0 range, F0 mean, and speech rate with respect to rhythmic regularity and phonological phrasing:** We have no grounds to hypothesize about the behavior of these acoustic-prosodic features (F0 range and mean, and speech rate) with respect to the rhythmic regularity and phonological phrasing manipulated in our experiments. Each one of them may prove to be higher (or broader), lower (or narrower), or even equal, in the regular rhythmic condition, or in the accentual group phrasing condition, with respect to the opposite conditions.

• **Results related to the mode of rendition (reading / repeating) with respect to rhythmic regularity and phonological phrasing:** As is the case of the acoustic-prosodic

features just mentioned, we have no grounds to hypothesize about the possible differences between reading and repeating with respect to the manipulated rhythmic regularity and phonological phrasing.

In addition, given the already mentioned inconsistencies between acoustic and perceptual data in studies of linguistic accommodation (Pardo, 2013b; Ruch et al., 2017; see Section 2.4.1.), and considering that the perceptual task applied in this thesis was specifically tailored to determine the capacity of listeners to assess rhythmic resemblance, it should be envisaged that the results of the acoustic measures of rhythmic resemblance may not be consistent with the results of the perceptual judgments of such resemblance.

# 4.2. Materials and methods

## 4.2.1. Experiment 1: Acoustic evaluation

### Participants

Twenty-four native Spanish speakers were tested: 12 females and 12 males, ranging from 18 to 28 years of age. Participants were unacquainted before the experiment and have lived in Bogotá (Colombia) the majority of their lives, so no different dialects were involved in the experiment. Participants had to fulfill three conditions to be part of the experiment: (1) not to present speech or hearing disorders; (2) not to have learned a second language during childhood; and (3) not to have formal musical training. All participants were university students and they were compensated with extra course credit for participation.

### Stimuli

Sixty-four sentences of nine syllables each were created for this study (the complete list of sentences can be consulted in Appendix 1). Each sentence is comprised of six words belonging to the 5,000 most frequent words in Spanish (see Real Academia Española, 2008). All 64 sentences have the same syntactic structure: subject + verb + complement.

Four blocks of 16 sentences were created, each one with a particular rhythmic structure. These structures were obtained through the arrangement of different types of words (oxytones, paroxytones, proparoxytones, and unstressed words) in *accentual feet* (Cantero, 2002) or *accentual groups* (Hualde & Nadeu, 2014) (See Section 3.2.2.). Rhythmic structures

were composed as follows (unstressed syllables are represented by a lowercase *x* and stressed syllables by an uppercase *X* and in uppercase within sentences):

**1. Regular feet (RF):** Xxx-Xxx-Xxx (e.g. MA-rio te VIO sin la MÁ-qui-na) [Mario saw you without the machine].

**2. Regular groups (RG):** xXx-xXx-xXx (e.g. la CA-sa se VEN-de por PAR-tes) [the house is sold by parts].

**3. Irregular feet (IF):** Xx-Xxxx-Xxx (e.g. SOL me CUEN-ta de su SÁ-ba-do) [Sol tells me about her Saturday].

**4. Irregular groups (IG):** Xx-xxxX-xXx (e.g. MA-rio se nos que-DÓ sin NO-via) [Mario ended up without a girlfriend].

All 64 sentences are comprised of three accentual feet or accentual groups, each one with a single strong syllable and without punctuation. Sentences within the regular feet and regular groups blocks have the same syllabic distribution (3 - 3 - 3). The same stands for sentences within the irregular feet and irregular groups blocks (syllabic distribution: 2 - 4 - 3). In all four blocks the last foot or group has three syllables, allowing further comparisons. These specific syllabic distributions were created from partial results of two previous pilot studies with seven distinct types of distributions, in order to improve the resulting differences between blocks of stimuli (please note that the pilot studies served as a calibration mechanism for the actual experiments, and some aspects such as the stimuli presentation technique and the construction of the sentences varied between pilots, therefore, the partial results obtained with them are not discussed in this document).

In addition, in all 64 sentences synalephas and other kinds of resyllabification between words' limits were avoided (see Section 3.2.4. for details of resyllabification). Voiced stops, /b/, /d/, and /g/, were also avoided in sentence initial position to ease measurement of onsets (in a similar way to Späth et al.'s [2016] procedure). Neither content words, nor particular combinations of functional words, are repeated within each block of stimuli.

To maintain only two levels of prominence (stress / no stress), we avoided all instances in which unstressed words might become stressed, and vice versa; that is: focus, nominalization, citation, parentheticals, and exclamatory and interrogative sentences (Face,

2003; Hualde, 2009). Compound words with two lexical stresses were also avoided. Further details of the stimuli are presented in Table 2.

**Table 2:** Number of words and letters in the four blocks of stimuli.

| | **Regular feet** | **Irregular feet** | **Regular groups** | **Irregular groups** |
|---|---|---|---|---|
| **Number of letters for sentence** | | | | |
| Average | 22.69 | 23.13 | 22.75 | 22.00 |
| SD | 1.40 | 1.63 | 1.88 | 1.93 |
| **Number of words by stress position** | | | | |
| Proparoxytones | 16 | 16 | 0 | 0 |
| Paroxytones | 16 | 16 | 48 | 32 |
| Oxytones | 16 | 16 | 0 | 16 |
| Unstressed words | 48 | 48 | 48 | 48 |
| **Number of words by syllable number** | | | | |
| 1-syllable words | 64 | 64 | 48 | 48 |
| 2-syllable words | 16 | 16 | 48 | 48 |
| 3-syllable words | 16 | 16 | 0 | 0 |
| **Total of words** | 96 | 96 | 96 | 96 |

# Procedure

Participants were divided in 12 dyads: four female / female (FF), four female / male (FM), and four male / male (MM). Each dyad was tested separately in a quiet room, in which the two participants were facing each other, with a 13-inch PC screen in front of each of them (one MacBook Pro Mid 2012 and one MacBook White Mid 2010 were used).

The first member of the dyad was asked to read aloud the sentence that appeared in his or her screen, exactly as it was written and with a declarative intonation and neutral focus (please note that, in Spanish, in sentences uttered in such way stressed syllables are usually accompanied by a pitch accent, while unstressed syllables are not; Face, 2003; Hualde, 2007; Ortega & Prieto, 2007; see Section 3.2.1.) [24] (This part of the procedure was implemented considering the possible role of pitch accents combined with duration, as an additional acoustic correlate of speech rhythm). Next, the second member of the dyad was

[24] It should be noted also the existence of subtle differences across Spanish variations in the tunes for declarative utterances (see Beckman et al., 2002 for a review).

asked to repeat the sentence exactly as listened, as is usually done in shadowing tasks (a drawing of an ear appeared in that moment in his or her screen). Participants were instructed to avoid speech overlaps and repairs. Reading and repeating the 16 sentences of each block of stimuli was alternated between the two members of the dyad. Moreover, the first turn of reading in each of the four blocks was also alternated. Consequently, all 64 sentences (four experimental blocks) were uttered twice (one time by each member of the dyad), and the reading and repetition were balanced within dyads and blocks.

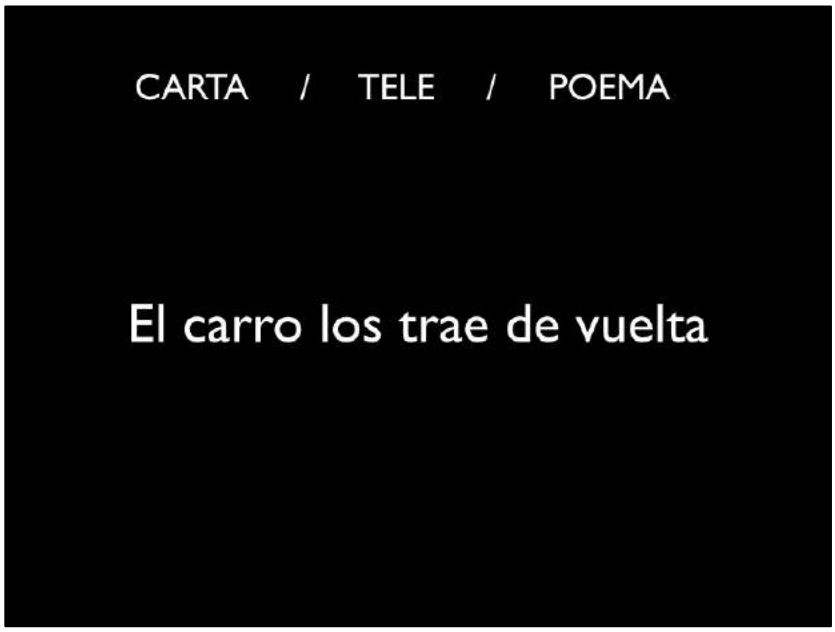
In order to maintain actual sentences' rhythms as similar as possible as intended, participants who mispronounced sentences during reading were asked to read again the complete sentence once (and in rare cases twice) before the repetition made by the other member of the dyad. This was the case in 88 sentences, approximately 5.7% of the total number of sentences. These reading repetitions were distributed as follows: in terms of phrasing, sentences arranged by feet 49% and sentences arranged by groups 51%, in terms of regularity, regular sentences 44% and irregular sentences 56%.

Additionally, when the second member of the dyad mispronounced a sentence during repetition, the complete sequence (reading + repetition) was repeated once (and in rare cases twice). This was the case in 76 sentences, approximately 4.9% of the total number of sentences. These sequence repetitions were distributed as follows: in terms of phrasing, feet 42% and groups 58%, in terms of regularity, regular 45% and irregular 55%. Both reading repetitions and sequence repetitions were taken into account during the statistical treatment of response times.

Furthermore, at the top of each participant's screen, above the sentence to be pronounced or the drawing of the ear indicating that a sentence had to be repeated, three *category-words* were presented, always in the same order: *poema* (poem), *tele* (short for television), and *carta* (letter) (See Graphics 1a and 1b). After each sentence was uttered by one participant and repeated by the other, both participants were asked to categorize the sentence as related to a letter, a poem, or a TV show, by saying aloud one of the three category-words (this was made in the same order in which sentences were uttered). That is to say: participant A read a sentence -> participant B repeated the sentence -> participant A categorized the sentence -> participant B categorized the sentence. This process was

followed for each of the 16 sentences in each of the four experimental blocks, alternating the participant that initiated the sequence. Each new sentence was presented in the computer screen as soon as both participants finished the categorization task.

**Graphic 1a:** Example of a screen showed to participants in Experiment 1.




**Graphic 1b:** Example of a screen showed to participants in Experiment 1.


All four blocks of stimuli (RF, IF, RG, IG) were presented to all three subgroups of dyads (FF, FM, MM). The order of presentation of stimuli blocks to each one of the four dyads

within each subgroup was arranged according to the following three conditions: (1) each one of the four blocks had to be presented in the first, second, third, and fourth position; (2) regular blocks had to be presented both after and before irregular blocks (applies also to the distinction group - foot); and (3) regular blocks had to be presented both following each other and separately regarding irregular blocks (applies also to the distinction group - foot) (Note the similarity with a balanced Latin Square). The order of presentation of the 16 sentences within each block was randomized. See Table 3 for a schematic description.

**Table 3:** Schematic description of the presentation of stimuli in Experiment 1.

| | | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|
| Female / Female dyads | **A** | RF | IF | RG | IG |
| | **B** | IF | IG | RF | RG |
| | **C** | RG | RF | IG | IF |
| | **D** | IG | RG | IF | RF |
| Female / Male dyads | **A** | RF | IF | RG | IG |
| | **B** | IF | IG | RF | RG |
| | **C** | RG | RF | IG | IF |
| | **D** | IG | RG | IF | RF |
| Male / Male dyads | **A** | RF | IF | RG | IG |
| | **B** | IF | IG | RF | RG |
| | **C** | RG | RF | IG | IF |
| | **D** | IG | RG | IF | RF |
| | | | | | |
| **1...4:** Order of presentation | | **RF:** Regular feet | | **IF:** Irregular feet | |
| **A...D:** Dyad | | **RG:** Regular groups | | **IG:** Irregular groups | |

At the begging of each experimental session a written informed consent was obtained from every participant. Participants were also told that they would take part in a study of conversations in Spanish, but no further information about the aims of the study or specific characteristics of the stimuli were provided. Beginning and ending of the experimental sessions were indicated verbally and in the screen of each participant. Also, beginning and ending of each block of stimuli were indicated verbally and in the screen, and a pause of two minutes was made. Presentation of the four blocks of stimuli was preceded by a familiarization block, comprised of six sentences, which did not belong to any of the

experimental blocks, nor did they have any particular rhythmic pattern. Presentation of stimuli and messages on the screens was managed by the experimenter, who shared the room with the participants, with the help of a wireless remote control linked to the two computers in a way that the screens' content (synchronized visual presentations) changed at the same time.

Experimental sessions were recorded with an iPad Air 2 (running on iOS 11.4), equipped with a dual-microphone system that registers speakers' voices from different directions. Audio files were recorded at 44,100 Hz - 24 bits - stereo, using the software Voice Record Pro 3.3.6. At the end of each experimental session, participants completed a brief survey indicating if they had perceived any particular difference between the blocks of stimuli.

## 4.2.2. Experiment 2: Perceptual evaluation

### Participants

Twenty-four native Spanish speakers participated in this experiment: 12 females and 12 males, ranging from 17 to 28 years of age. Participants had to fulfill three conditions to be part of the experiment: (1) not to present hearing disorders; (2) not to have learned a second language during childhood; and (3) not to have formal musical training. All participants were university students and they were compensated with extra course credit for participation (none of them took part in Experiment 1).

### Stimuli

Audio files corresponding to the pair of sentences presented in the first, second, eighth, ninth, fifteenth, and sixteenth positions were extracted from each block of stimuli (RF, IF, RG, IG), of each one of the 12 dyads' recordings of Experiment 1. In this way, we acquired tokens of the beginning, middle, and final parts of all dyads' interactions in all conditions: twenty-four tokens altogether for each dyad. Each token consisted then in the same sentence repeated by the two participants of the dyad. Half of the tokens taken from each dyad were read by participant A and repeated by participant B; the inverse situation occurred in the other half.

Using Audacity 2.0.2 for Mac, the sound of each token was individually normalized, setting the amplitude peak at -1.0 dB. This process allowed us to present all the stimuli at a similar volume while maintaining signal-to-noise ratio and relative dynamics unchanged. Half a second of total silence was inserted between each pair of sentences, replacing the original interval time between them. Half a second of silence was also inserted at the beginning and at the end of each token.

## Procedure

To keep the task to a manageable length for the participants, presentation of stimuli was distributed in a way that the 24 tokens of each one of the 12 dyads (Experiment 1) were rated by eight evaluators (four females and four males). Consequently, each one of the 24 evaluators rated a total of 96 pairs of sentences, corresponding to one subgroup of dyads (FF, FM, or MM). The order of presentation of the dyads to each evaluator followed the same logic of the order of presentation of stimuli blocks in Experiment 1. The order of presentation of blocks (RF, IF, RG, IG) corresponding to each dyad, and the order of tokens within each block, were randomized. See Table 4 for a schematic description.

**Table 4:** Schematic description of the presentation of stimuli in Experiment 2.

| | **1** | **2** | **3** | **4** | |
|---|---|---|---|---|---|
| **M1 & M7** | A | B | C | D | **Female / Female dyads** |
| **M2 & M8** | B | D | A | C | |
| **F1 & F7** | C | A | D | B | |
| **F2 & F8** | D | C | B | A | |
| **M3 & M9** | D | C | B | A | **Female / Male dyads** |
| **M4 & M10** | C | A | D | B | |
| **F3 & F9** | B | D | A | C | |
| **F4 & F10** | A | B | C | D | |
| **M5 & M11** | A | B | C | D | **Male / Male dyads** |
| **M6 & M12** | C | A | D | B | |
| **F5 & F11** | B | D | A | C | |
| **F6 & F12** | D | C | B | A | |
| | | | | | |
| **M1…M12:** Male 1 to 12 (evaluating) | | | **A…D:** Dyad (to evaluate) | | |
| **F1…F12:** Female 1 to 12 (evaluating) | | | **1…4:** Order of presentation | | |

Stimuli were presented through a pair of headphones (Sony MDR-ZX110) connected to a PC (MacBook Pro Mid 2012). Volume was adjusted by each participant. Participants were instructed to rate the *rhythmic similarity* between the two sentences within each token on a five-point Likert scale, with 1 being "not similar at all" and 5 being "extremely similar." Participants were informed that the second sentence was a repetition of the first one, and they were instructed to focus on the rhythmic aspect of the sentences, leaving aside the tone of voice, the volume, and the speech rate of both members of the dyad. In order to make instructions clear, a familiarization block of six tokens was presented. These tokens were extracted from two previous pilot studies, and consisted in pairs of sentences with high scores, medium scores, and low scores with respect to their acoustic rhythmic resemblance (see the data analysis section). This familiarization block was scored by each participant under the guide of the experimenter.

During the experiment, each token (reading and repetition of a sentence) was presented as soon as the participant finished rating the previous one. Stimuli presentation was controlled by the experimenter using the PC keyboard. At the end of each experimental session, participants completed a brief survey indicating if they had perceived any particular difference between the blocks of stimuli.

# 4.3. Data analysis

## Preparation of participants' recordings

In the first place, recordings were segmented into pairs of sentences and pairs of category-words. One text grid for each pair of sentences (including the silence between them) was then created using Praat 6.0.20 (Boersma & Weenink, 2017). Next, text grids and corresponding audio files were aligned using the software SPPAS, version 1.7.6 (Bigi, 2015). Lastly, correction and adjustment of aligned text grids, as well as all subsequent acoustic analyses, were hand-made by the experimenter using Praat 6.0.20 (Boersma & Weenink, 2017).

A total of 1536 sentences were analysed, and six acoustic-prosodic features were determined, as follows:

**1. Total length of each sentence (seconds).** Given the characteristics of the stimuli, no "abnormal" silences within sentences were expected. However, an arbitrary lapse of maximum 0.2 seconds of silence within a sentence was established. Consequently, silences longer than 0.2 seconds were discarded when establishing the total length of the sentence and the length of each one of the three feet or groups. This was the case in 27 sentences, approximately 1.8% of the total number of sentences. These discarded silences were distributed as follows: in terms of phrasing, sentences arranged by feet 41% and sentences arranged by groups 59%, in terms of regularity, regular sentences 67% and irregular sentences 33%.

**2. Length of each one of the three feet or groups composing each sentence (seconds).** Accentual groups were established in accordance with the definitions provided by Almeida (1993, 1997), Hualde and Nadeu (2014), and Mora et al. (1999). Accentual feet, in turn, were established in accordance with the definitions provided by Almeida (1993, 1997) and Cantero (2002) (See Section 3.2.2.). In both cases, following Almeida (1997), vocalic transitions were considered as a part of the vowel, and stops' release periods as a part of the consonant, when it was the case.

**3. Time elapsed between the end of participant A's sentence rendition and the beginning of participant B's sentence repetition (seconds)**. Being participants A and B the two members of each dyad**.**

**4. Average pitch of each sentence (hertz).**

**5. Minimum pitch of each sentence (hertz).**

**6. Maximum pitch of each sentence (hertz).**

## Determination of dependent variables

Using the aforementioned six acoustic-prosodic features, and the count of the 1536 category-words, ten dependent variables were determined, as follows:

**1. IT (Interval time):** Time elapsed between the end of participant A's sentence rendition and the beginning of participant B's sentence repetition (seconds).

**2. F0M (F0 mean):** Average pitch of each sentence (hertz).

**3. F0R (F0 range):** Pitch range of each sentence (hertz). It was obtained by subtracting the pitch minimum from the pitch maximum in each sentence. Note that the pitch range, also referred to sometimes as *pitch span*, has been established in at least two different ways: (1) F0 maximum minus F0 minimum (De Looze et al., 2011; Pépiot, 2014), and (2) F0 standard deviation (Traunmüller & Eriksson, 1994). It has also been subcategorized into *overall pitch level* and *pitch span* (Patterson & Ladd, 1999). A detailed commentary on the terminological differences between *pitch level, pitch span* and *pitch range* can be found in Patterson and Ladd (1999).

**4. SR (Speech rate):** Syllables per second (as measured in Wang, Kong, Zhang, Wu & Li, 2018). Obtained from the total length of each sentence (note that all 64 sentences used in this experiment have the same number of syllables; see Table 2 for the number of letters in the sentences comprising each block of stimuli). Please note also that speech rate may differ during reading and during speaking / repeating (see Section 4.1.). This difference is taken into account during the statistical analysis of data.

**5. RD (Rhythmic distance):** First of all, it must be noted that, though several methods exist, there is no standard for measuring prosodic accommodation (Thomason et al., 2013)[25]. In our experiment, following Späth et al. (2016), a *rhythmic distance score* (RDS) was used to determine the degree of rhythmic resemblance between each reading sentence and its repetition. The RDS consists in comparing the Euclidean distances of the relative duration of metrical units within each sentence (in our case, feet or groups). This means that for each foot / group, from each sentence, for each participant, there is a single distance measure that aids to determine the RDS of each pair of sentences.

Following the procedure detailed in Späth et al. (2016), each metrical unit (foot / group) was divided by the total length of its corresponding sentence before computing the RDS (in such a way, each pair of sentences is normalized for speech rate). Next, the following formula was used to compute the RDS of each pair of sentences:

*sqrt((a1 - b1)^2 + (a2 - b2)^2 + (a3 - b3)^2)*

[25] In this respect, some methods to evaluate behavioral accommodation between humans are revised in Duranton and Gaunet (2016). Additionally, although acoustic-prosodic accommodation is usually measured in dyadic interactions, Rahimi, Litman and Paletz (2019) discuss the correct methodology for analyzing this phenomenon at a group level.

Where:

*a1, a2, a3* = relative feet / groups durations of the sentence uttered by the first member of the dyad (reading).

*b1, b2, b3* = relative feet / groups durations of the same sentence spoken by the second member of the dyad (repetition).

A resulting RDS value of 0 indicates the exact equivalency between participants' metrical timing patterns (rhythms), independent of the absolute speaking rate. The more the resulting RDS values distance from 0, the more the metrical timing patterns distance from each other. Note that Euclidean distances used in this way do not have obvious bound values for the maximum distance, but some maximum possible discrepancy value that remains unknown until specifically computed (Barrett, 2005). Nonetheless, as stated by authors such as Gessinger et al. (2018) and Späth et al. (2016), convergence is quantified as a decrease in Euclidean distance, and hence and increase in similarity.

**6. F0MD (F0 mean distance):** Considering that Euclidean distances allow the comparison of any two vectors taken across the same variable (Barrett, 2005), and because of that they have been used to establish the degree of distance / similitude of tokens of several linguistic features, such as height and slope of pitch accents (Gessinger et al., 2018), words within a sentence (Ferrer, 2004), vowels' formant frequencies (Babel, 2010; Pardo et al., 2010), and mean pitch, mean intensity, and duration (in seconds) of speech preceding backchannels (Levitan et al., 2011), we propose a *F0 mean distance score* (F0MDS) to determine the degree of distance between the average values of F0 of a reading sentence and its repetition. In this case, as in the case of the RDS discussed above, we consider a decrease in Euclidean distance as an increase in similarity between the measured tokens (see Gessinger et al. [2018] and Späth et al. [2016]).

Raw Euclidean distances, however, are sensitive to the scaling of each constituent variable. Consequently, mischievous data may result from the comparison of variables whose score ranges are quite different (in our case, acoustic-prosodic features of women and men) (Barrett, 2005). In the case of the rhythmic distance score (RDS), this problem was solved dividing each metrical unit (foot / group) by the total length of its corresponding sentence (therefore, normalizing it). However, whereas the RDS is

established in a three-dimensional Euclidean space, the F0 mean distance score (F0MDS) that we propose should be established in a Euclidean space of only one dimension, rendering impossible to normalize both of them in the same manner (as is the case of the distance scores proposed later for both pitch range and speech rate).

To overcome this difficulty, we normalized (z-scored) the F0MDS of each one of the twelve dyads, separately, using the function *scale()* in the software R (this function calculates the general mean and standard deviation of a set of values, and then subtracts the mean and divides by the standard deviation each one of those values). Next, the following formula was used:

*sqrt((a - b)^2)*

Where:

*a* = z-scored F0 mean of the sentence uttered by the first member of the dyad (reading).

*b* = z-scored F0 mean of the same sentence spoken by the second member of the dyad (repetition).

As well as in the RDS, in the F0MDS a resulting value of 0 indicates an exact equivalency between the measured units (in this case, F0 means of two sentences). The more the resulting value distances from 0, the more the measured units distance from each other.

**7. F0RD (F0 range distance):** Following the same logic of the F0 mean distance score (F0MDS) described above, we propose a *F0 range distance score* (F0RDS) to determine the degree of distance between the F0 range value of a reading sentence and the F0 range value of its repetition. The following formula was used:

*sqrt((a - b)^2)*

Where:

*a* = z-scored F0 range of the sentence uttered by the first member of the dyad (reading).

*b* = z-scored F0 range of the same sentence spoken by the second member of the dyad (repetition).

As in the case of the F0MDS, F0RDS values were z-scored by dyad. Also, resulting scores of 0 indicate an exact equivalency between the measured units (in this case, F0 ranges of

two sentences). The more the resulting value distances from 0, the more the measured units distance from each other.

**8. SRD (Speech rate distance):** Following the same logic of the F0 mean and range distance scores, we propose a *speech rate distance score* (SRDS) to determine the degree of distance between the speech rate value of a reading sentence and the speech rate value of its repetition. The following formula was used:

$$sqrt((a - b)^2)$$

Where:

*a* = z-scored total length of the sentence uttered by the first member of the dyad (reading).

*b* = z-scored total length of the same sentence spoken by the second member of the dyad (repetition).

Given that the sentences being compared are identical in terms of the syllables and words within them, we assume that the distance between their total duration corresponds to the distance between the speech rates of the members of the dyad that uttered them. This was done in order to avoid the double transformation of data from total length to syllables per second to the distance score. As in the case of the F0 mean and range distance scores, SRDS values were z-scored by dyad. Also, resulting scores of 0 indicate an exact equivalency between the measured units (in this case, the speech rates of participants uttering the sentences). The more the resulting value distances from 0, the more the measured units distance from each other.

**9. LR (Lexical repetitions):** Considering that, as seen in Section 2.4.3., the recurrent use of the same term made by different persons during a conversation in order to refer to an object or situation is considered an instance of lexical convergence (e.g. Brennan & Clark, 1996), and that in some scenarios the potential for enormous variability in people's lexical choices makes unlikely that two persons use the exactly same term during a conversation to refer to some specific action or subject (e.g. when deciding between, *delete, erase, kill, omit, destroy, lose,* or *change* when removing a file from a computer) (Brennan, 1996), we decided to establish a small and explicit list of lexical choices in order to asses if any of the

experimental conditions would lead to a greater amount of lexical repetitions between the two members of each dyad.

The LR variable indicates thus the number of times in which both participants classified a sentence, after reading and repetition, with the same category-word: (a) *poema* (poem), (b) *tele* (short for television), or (c) *carta* (letter) (See Section 4.2.1.). One point (1) was coded if both participants used the same category-word. Zero points (0) were coded if each participant used a different category-word.

**10. PR (Perceptual rating):** This variable indicates the ratings of rhythmic similarity obtained through the five-point Likert scale used in Experiment 2 (with 1 being "not similar at all" and 5 being "extremely similar").

## Statistical approach

All statistics were performed using R (R Core Team, 2017), with the R packages *lmerTest* (Kuznetsova, Brockhoff & Christensen, 2017), *influence.ME* (Nieuwenhuis, Grotenhuis & Pelzer, 2012), and *MASS* (Venables & Ripley, 2002). Following the procedure conducted by Späth et al. (2016), a series of linear mixed effects models were calculated in order to estimate the factors influencing resulting data (except for the lexical repetition variable [LR], discussed below). P-values and Satterthwaite approximations for degrees of freedom (rounded to integers) are reported for each model (for the validity of linear mixed effects models in Likert-scale data analysis see Kizach [2014] and Norman [2010]; an example in Gibson, Piantadosi & Fedorenko [2011]).

A design-driven model selection was conducted to determine the fixed and random effects in the models (see Cardwell, 2016). The use of subject- (dyad or participant) or item-specific slopes was determined case by case finding the largest model that converged and comparing it to the intercept-only model, as described in Barr (2013).

Five variables were entered as fixed effect factors: (1) regularity of the sentences, coded as **STRCT**, with two levels: regular / irregular; (2) type of phrasing, coded as **UNIT**, with two levels: groups / feet; (3) type of dyad, coded as **SEX**, with three levels: FF (female / female), FM (female / male), and MM (male / male); (4) half of the test (i.e. first or last half of the 16 sentences comprising each block of stimuli), coded as **HALF**, with two levels: first

/ last, and (5) mode of rendition, coded as **MODE**, with two levels: reading / repeating. Note that, given that our analytical approach of accommodation relies on the comparison of degrees of similarity rather than patterns of synchronization, the variable HALF was entered in order to assess a basic effect of convergence between the speakers (as mentioned in Section 2.5.4., one way to assess the linear progression of prosodic accommodation is to compare the interlocutors' degree of similarity during the first and second halves of an interaction; e.g. De Looze et al., 2014).

To account for the independence of the observations, two variables were used as crossed random effects (see Baayen, 2008). In the models with a dependent variable obtained through paired measures (interval times and distance scores), and in the perceptual rating model: (1) the 64 sentences used as stimuli, coded as **ITEM**, with one level for each sentence; and (2) the 12 dyads, coded as **DYAD**, with one level for each dyad. In the models with a dependent variable obtained through individual measures (F0 range, F0 mean, and speech rate): (1) the 64 sentences used as stimuli, coded as **ITEM**, with one level for each sentence, same as the rest of the models; and (2) the 24 individual participants, coded as **PTCP**, with one level for each one.

On the other hand, given that the raw data did not indicate considerable differences in the lexical repetitions count in terms of STRCT or UNIT (see Table 6), a binomial logistic regression was conducted to evaluate if the observed difference between types of dyad (SEX) was significant (note that the data of the three types of dyad: FF, FM, and MM, are constituted by independent observations). In this case, P-values are determined by likelihood ratio chi-square tests.

For each model (except the one for lexical repetitions), Cook's distances were evaluated by subject (dyad or participant) and by item (sentence) separately (Baayen, 2008). A cut-off value for Cook's distance of $D_i > 4/n$ (where $n$ is the number of items, dyads, or participants), was used to identify highly influential points (Nieuwenhuis et al., 2012). Items or subjects found to be influential points in the model were reanalyzed, measuring a second time the related scores of the dependent variable at least two standard deviations higher or lower than the general average, in order to find anomalies in the recordings or in

the measurement procedure. Results of these tests are presented with each model in the following section.

In addition, a graphical diagnostic of each model was conducted in accordance with the guidelines provided by Gelman and Hill (2007). When the inspection of residual plots revealed a significant deviation from homoscedasticity, or lack of linearity, five data transformations of the dependent variable were considered: square root (*sqrt*), cube root (*cbrt*), log transformation (*log*), Tukey ladder of powers (*T_tuk*), and Box Cox transformation (*T_box*). The most suitable transformation with respect to the model assumptions was chosen. Data that did not present significant deviation from homoscedasticity, or lack of linearity, were treated raw. Details are offered with each model in the following section. Furthermore, following Gelman and Hill (2007), and Norman (2010), normality of the residuals was not considered as an assumption of the models. Finally, as commented above, the independence assumption was satisfied using random intercepts for subjects (dyads or participants) and items (sentences) in each model. In the case of the binomial logistic regression for the lexical repetition variable, residual plots and outliers were analyzed following the procedure described by Zhang (2016).

# 4.4. Results

**Table 5:** Groups' and feet's lengths (average and standard deviations [SD]), quantity, and number of syllables.

| | | | Length (sec) | |
|---|---|---|---|---|
| | **Syllables** | **n** | **Average** | **SD** |
| **Feet** | 2 | 384 | 0.475 | 0.135 |
| | 3 | 1536 | 0.586 | 0.143 |
| | 4 | 384 | 0.728 | 0.160 |
| **Groups** | 2 | 384 | 0.403 | 0.120 |
| | 3 | 1536 | 0.571 | 0.131 |
| | 4 | 384 | 0.728 | 0.199 |

**Table 6**: General results (average values are presented first, followed by standard deviations in parenthesis; symbols representing levels of significance are explained at the bottom of the table). Non-transformed data are presented. Levels of significance and

interactions are based on the statistical models employed, in which some data were transformed, as explained in the next section.

| | Rhythmic regularity (STRCT) | | Phonological phrasing (UNIT) | |
|---|---|---|---|---|
| | Regular | Irregular | Groups | Feet |
| **RD - Rhythmic distance** | 0.053 (0.025) *** 0.065 (0.037) | | 0.060 (0.034) / 0.058 (0.030) | |
| **IT - Interval time (sec)** | 0.613 (0.294) / 0.664 (0.320) | | 0.660 (0.299) / 0.617 (0.316) | |
| **F0R - F0 range (hertz)** | 78.96 (32.08) ** 88.30 (38.21) | | 80.36 (32.53) ** 86.91 (38.11) | |
| **F0RD - F0 range distance** | 1.076 (0.798) / 1.201 (0.955) | | 1.076 (0.863) * 1.201 (0.896) | |
| **F0M - F0 mean (hertz)** | 181.56 (53.03) ** 183.62 (52.23) | | 180.77 (52.17) ** 184.42 (53.04) | |
| **F0MD - F0 mean distance** | 1.561 (0.711) / 1.629 (0.751) | | 1.583 (0.716) / 1.607 (0.747) | |
| **SR - Speech rate (syllables*sec)** | 5.42 (0.87) / 5.28 (0.94) | | 5.45 (0.96) / 5.25 (0.85) | |
| **SRD - Speech rate distance** | 0.94 (0.67) / 1.00 (0.75) | | 0.98 (0.72) / 0.96 (0.71) | |
| **LR - Lexical repetitions (count)** | 158 / 163 | | 164 / 157 | |
| **PR - Perceptual rating (Likert)** | 3.22 (1.32) / 3.20 (1.33) | | 3.22 (1.32) / 3.20 (1.33) | |

| | Type of dyad (SEX) | | |
|---|---|---|---|
| | Female / Female | Female / Male | Male / Male |
| **RD - Rhythmic distance** | 0.053 (0.028) / 0.064 (0.035) / 0.060 (0.033) | | |
| **IT - Interval time (sec)** | 0.600 (0.258) / 0.575 (0.250) / 0.740 (0.375) | | |
| **F0R - F0 range (hertz)** | 103.13 (30.80) ** 86.42 (38.48) ** 61.35 (22.03) | | |
| **F0RD - F0 range distance** | 0.967 (0.773) / 1.316 (0.943) / 1.132 (0.889) | | |
| **F0M - F0 mean (hertz)** | 227.64 (18.65) *** 189.78 (53.49) *** 130.35 (17.02) | | |
| **F0MD - F0 mean distance** | 1.371 (0.896) / 1.943 (0.313) / 1.472 (0.722) | | |
| **SR - Speech rate (syllables*sec)** | 4.93 (0.68) * 5.39 (0.98) * 5.73 (0.86) | | |
| **SRD - Speech rate distance** | 0.97 (0.75) / 1.04 (0.69) / 0.90 (0.69) | | |
| **LR - Lexical repetitions (count)** | 116 * 91 * 114 | | |
| **PR - Perceptual rating (Likert)** | 3.07 (1.34) ** 3.56 (1.32) ** 3.01 (1.25) | | |

| | Half of the test (HALF) | | Mode of rendition (MODE) | |
|---|---|---|---|---|
| | First | Last | Reading | Repeating |
| **RD - Rhythmic distance** | 0.058 (0.030) / 0.060 (0.034) | | | |
| **IT - Interval time (sec)** | 0.642 (0.311) / 0.635 (0.305) | | | |
| **F0R - F0 range (hertz)** | 84.94 (36.01) *** 82.33 (35.10) | | 91.43 (37.88) *** 75.84 (31.23) | |
| **F0RD - F0 range distance** | 1.123 (0.841) / 1.153 (0.922) | | | |
| **F0M - F0 mean (hertz)** | 182.57 (52.71) / 182.61 (52.58) | | 185.88 (52.86) *** 179.30 (52.21) | |
| **F0MD - F0 mean distance** | 1.587 (0.736) / 1.603 (0.728) | | | |
| **SR - Speech rate (syllables*sec)** | 5.40 (0.94) *** 5.30 (0.88) | | 5.17 (0.90) *** 5.53 (0.89) | |
| **SRD - Speech rate distance** | 0.97 (0.71) / 0.97 (0.71) | | | |
| **LR - Lexical repetitions (count)** | 150 / 171 | | | |
| **PR - Perceptual rating (Likert)** | 3.29 (1.30) * 3.14 (1.34) | | | |

| | Interactions |
|---|---|
| **RD - Rhythmic distance** | STRCT*UNIT |
| **IT - Interval time (sec)** | — |
| **F0R - F0 range (hertz)** | HALF*SEX |
| **F0RD - F0 range distance** | STRCT**SEX |
| **F0M - F0 mean (hertz)** | MODE***UNIT |
| **F0MD - F0 mean distance** | — |
| **SR - Speech rate (syllables*sec)** | SEX*HALF |
| **SRD - Speech rate distance** | — |
| **LR - Lexical repetitions (count)** | — |
| **PR - Perceptual rating (Likert)** | UNIT*SEX |
| | |
| **Levels of significance:** *** (0.001) ** (0.01) * (0.05) / (Not significant) | |

# 4.4.1. Experiment 1: Acoustic evaluation

## Rhythmic distance (RD)

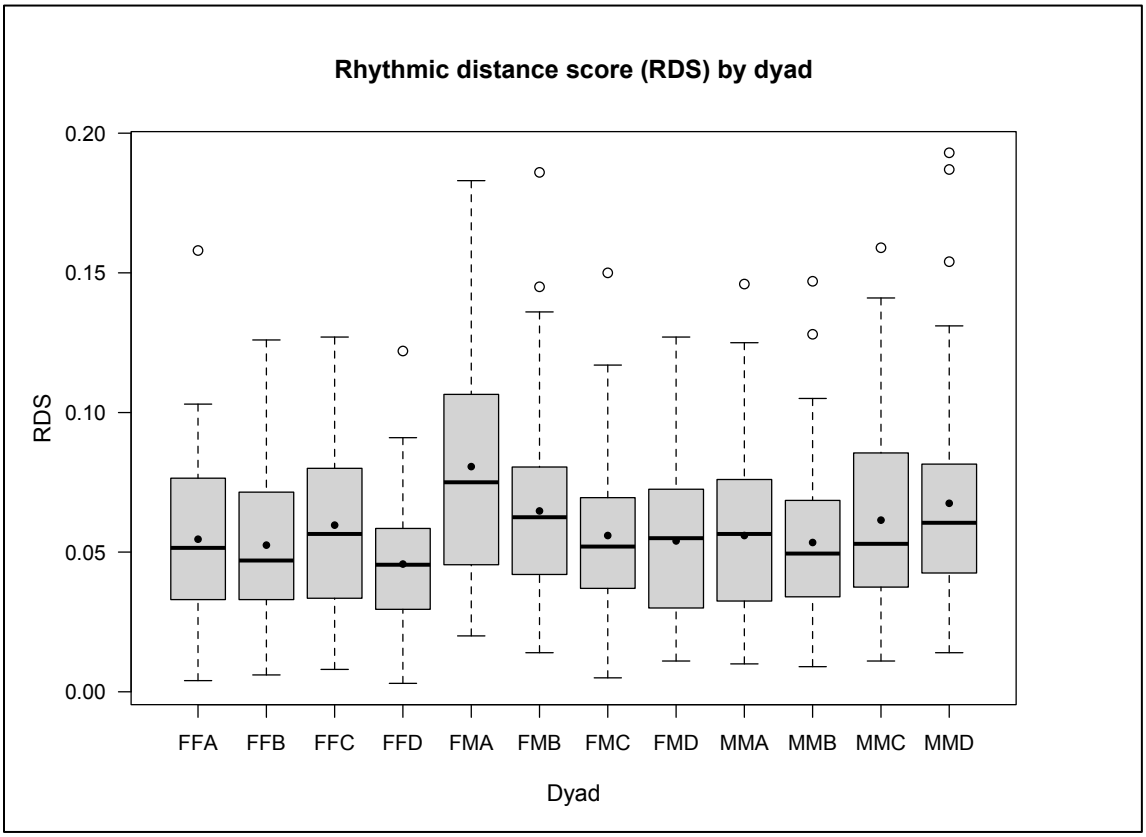
*Influential points*

None of the Items was found to exceed the established cut-off value for Cook's distance ($D_i > 4/n$), but the dyad FMA did exceed it. The recordings of this particular dyad with scores of RD at least two standard deviations higher or lower than the general average were measured a second time. This was the case in 11 pairs of sentences, approximately 1.4% of the total number of sentences pairs. No evidence was found for errors in the recordings or in the measurement procedure. Rather, it was established that the female member of the concerned dyad tended to speak in a leisurely pace, while also lengthening the words (See Graphic 2). Consequently, all points of analysis were retained in the model.

*Transformations*

Inspection of residual plots in this particular model revealed a significant deviation from homoscedasticity. As mentioned above, five transformations were considered. The square root transformation (*sqrt*) managed to solve the problem.

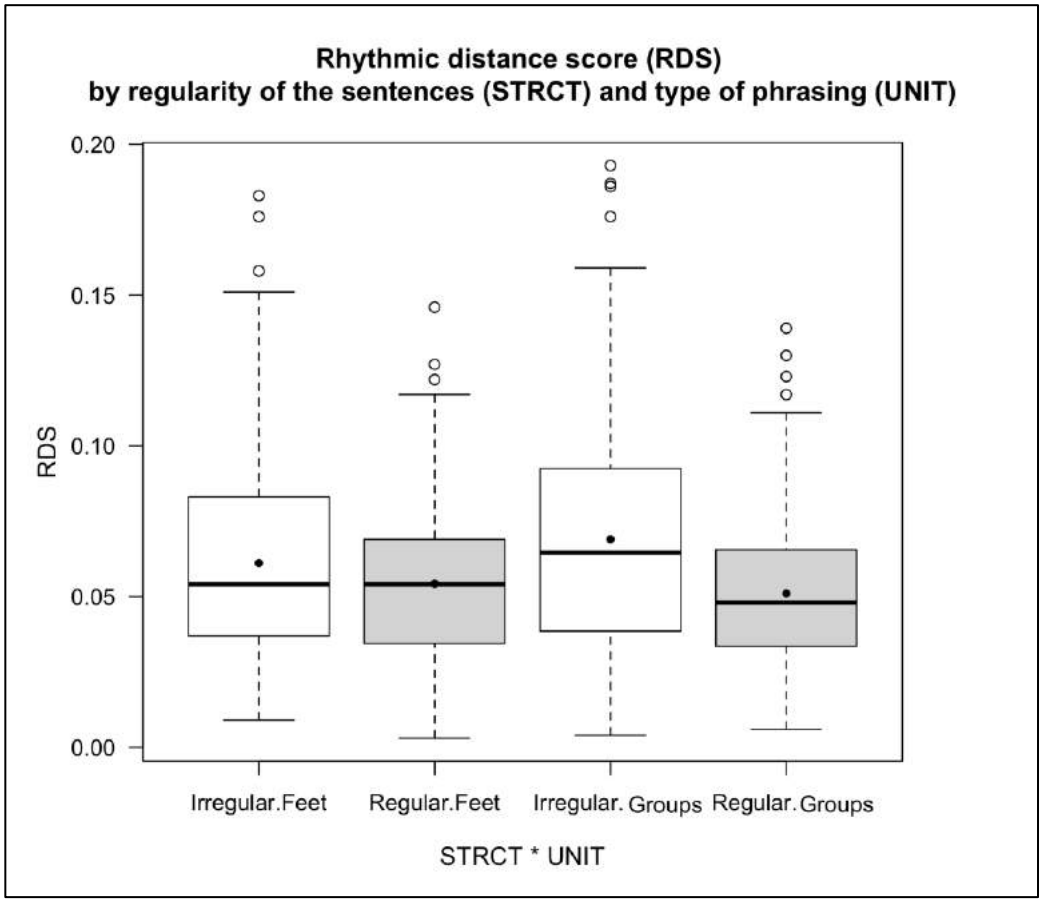
**Graphic 2:** Rhythmic distance score (RDS) by dyad (**FF:** Female / Female; **FM**: Female / Male; **MM:** Male / Male; **A-D:** Dyads).




*Results*

There was a significant effect of the factor **STRCT** ($F(1,60) = 22.59$; $p < 0.001$), but not of **UNIT** ($F(1,60) = 0.62$; $p = 0.433$), **SEX** ($F(2,9) = 1.61$; $p = 0.253$), or **HALF** ($F(1,749) = 0.93$; $p = 0.336$). Additionally, there was a significant interaction effect for **STRCT × UNIT** ($F(1,60) = 4.50$; $p < 0.05$). After the square root data transformation a significant interaction of **SEX × HALF** ($F(2,750) = 3.54$; $p = 0.029$) was lost, and was replaced by ($F(2,750) = 2.67$; $p = 0.069$). Hence, rhythmic distance was closer in metrically regular compared to irregular sentences, and this effect was greater in sentences arranged by groups than in sentences arranged by feet (See Graphic 3).

**Graphic 3:** Rhythmic distance score (RDS) by regularity of the sentences (STRCT) and type of phrasing (UNIT).




# Interval time (IT)

*Influential points*

According to Baayen (2008), data points that are suspect for experimental reasons should be removed. In our case, the reading and sequence repetitions due to mispronounced sentences, discussed above (see *Procedure* in Section 4.2.1.), may have influenced the lapse of time between reading and repetition of the implicated or adjacent sentences. In this respect, considering that shadowing of complete sentences has been reported within a typical range between 500 and 1,500 ms, with exceptional "close shadowers" with mean latencies of roughly 250 ms (Marslen, 1973, 1985), and that distributions of response times in psycholinguistic research tend to be positively skewed

(Baayen & Milin, 2010), a conservative threshold between 200 and 2,000 ms was established. Accordingly, 25 values lower than 200 ms (3.3% of the total number of data) and two values higher than 2,000 (0.3% of the total number of data) were replaced with the general average IT score. Nonetheless, no difference in the (lack of) significance of the predictors was found comparing the complete and the trimmed models.

Moreover, none of the dyads was found to exceed the established cut-off value for Cook's distance ($D_i > 4/n$), but item number 11 did exceed it. The recordings of this item corresponding to scores of IT at least two standard deviations higher or lower than the general average were measured a second time. This was the case in three pairs of sentences, approximately 0.4% of the total number of sentences pairs. One miscalculation was found and corrected. After that, all points of analysis were retained in the model (excluding, of course, the data trimmed before).

*Transformations*

Inspection of residual plots in this model revealed a significant deviation from homoscedasticity. The log transformation (*log*) managed to solve the problem.

*Results*

There were no significant effects of any of the factors: **STRCT** ($F(1,30) = 2.09$; $p = 0.159$), **UNIT** ($F(1,23) = 2.57$; $p = 0.123$), **SEX** ($F(2,15) = 0.86$; $p = 0.442$), or **HALF** ($F(1,716) = 0.08$; $p = 0.779$). Also, no significant interactions between predictors were found.

## F0 range (F0R)

*Influential points*

None of the Items was found to exceed the established cut-off value for Cook's distance ($D_i > 4/n$), but the participants FFB1 (female) and FMC2 (male) did exceed it. The recordings of these participants with scores of F0R at least two standard deviations higher or lower than the female or male average, respectively, were measured a second time (i.e. influential points of FFB1 were measured with respect to the rest of female participants' F0R, and FMC2 with respect to the rest of male participants). This was the case in 3 sentences, approximately 0.2% of the total number of sentences. One miscalculation was found and corrected. After that, all points of analysis were retained in the model.

*Transformations*

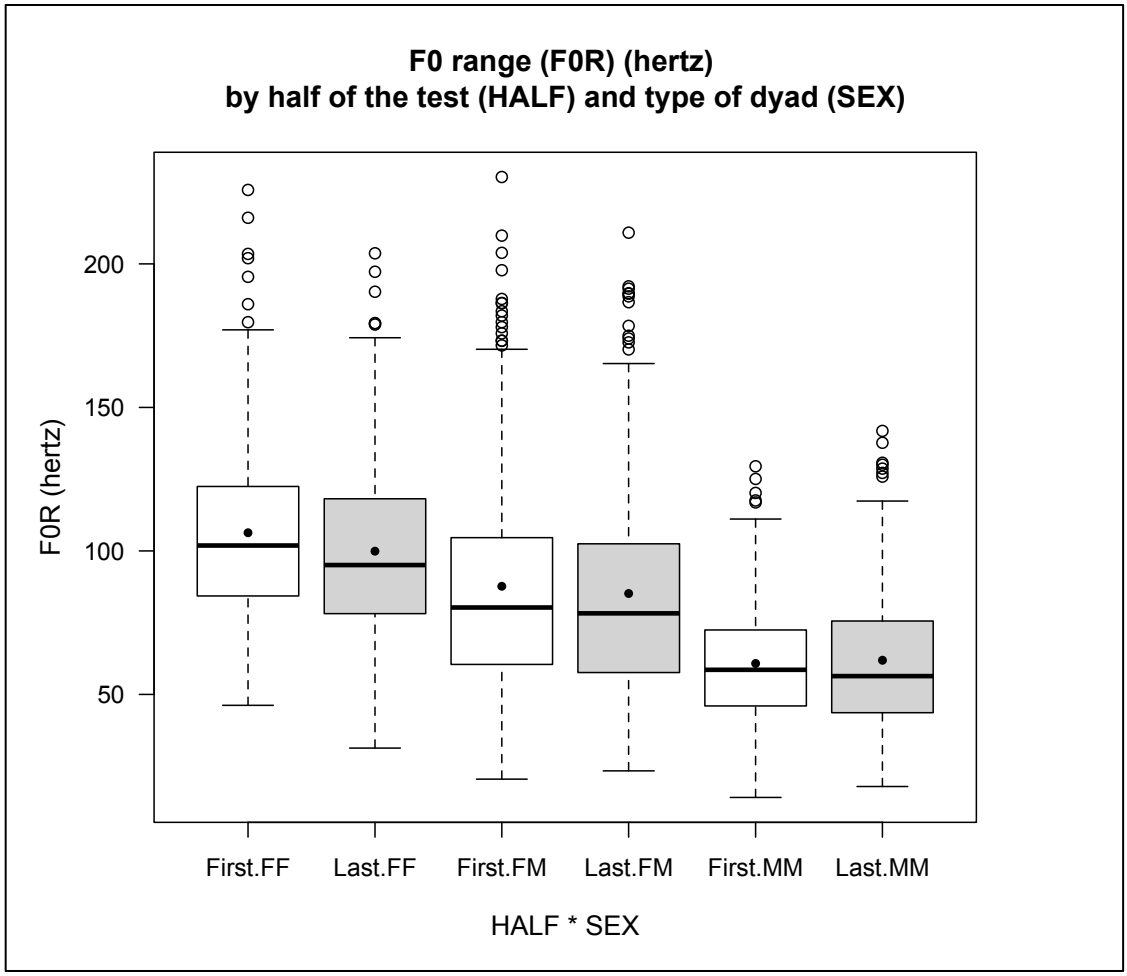
Inspection of residual plots in this particular model revealed a significant deviation from homoscedasticity. The Box Cox transformation (*T_box*) managed to solve the problem.

*Results*

There were significant effects of the factors **STRCT** ($F(1,48) = 10.45$; $p < 0.01$), **UNIT** ($F(1,57) = 7.81$; $p < 0.01$), **SEX** ($F(2,21) = 7.87$; $p < 0.01$), **HALF** ($F(1,1421) = 11.32$; $p < 0.001$), and **MODE** ($F(1,1376) = 288.55$; $p < 0.001$). Additionally, there was a significant interaction effect for **HALF x SEX** ($F(2,1424) = 3.04$; $p < 0.05$). Consequently, participants' F0 ranges were narrower in metrically regular sentences compared to metrically irregular sentences, in sentences arranged by group compared to sentences arranged by feet, and during repetition of sentences compared to reading of sentences. Also, F0 ranges were narrower in male only dyads compared to female only dyads. In this regard, given that female / male dyads' F0 ranges are an average between a female participant and a male participant, mixed dyads scores are around the middle between female / female and male / male dyads (see Table 6, above).

Additionally, participants' F0 ranges were narrower during the last half of the interaction compared to the first half, especially in the female / female dyads, and, to a lesser extent, in the female / male dyads. Conversely, participants' F0 ranges were slightly wider during the last half of the interaction compared to the first half in male / male dyads (See Graphic 4). In this respect, when the model was calculated without the female / female dyads, the significance of **SEX** disappeared ($F(1,14) = 3.39$; $p = 0.087$), as well as the interaction effect for **SEX × HALF** ($F(1,955) = 1.96$; $p = 0.162$) (All the other interactions remaining). This fact indicates that the **SEX × HALF** interaction must only be considered for female / female dyads (F0 ranges narrower during the last half of the interaction), and that the main difference in **SEX** respect to the range of F0 applies to female only dyads regarding men only dyads.

**Graphic 4**: F0 range (F0R) (hertz) by half of the test (HALF) and type of dyad (SEX) (**FF:** Female / Female; **FM**: Female / Male; **MM:** Male / Male).




# F0 range distance (F0RD)

*Influential points*

Given that influential points on F0R have already been examined, all points of analysis were retained in this model.

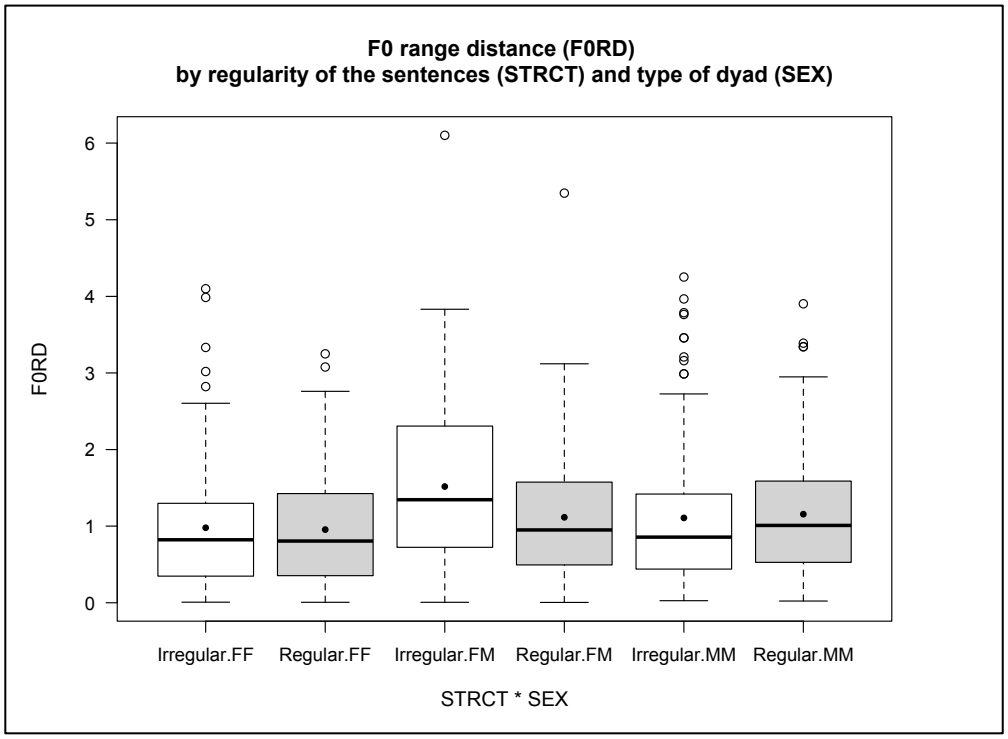
*Transformations*

Inspection of residual plots in this model revealed a significant deviation from homoscedasticity. The cube root transformation (*cbrt*) managed to solve the problem.

*Results*

There was a significant effect of the factor **UNIT** ($F(1,61) = 4.98$; $p < 0.05$), but not of **STRCT** ($F(1,61) = 1.79$; $p = 0.187$), **SEX** ($F(2,9) = 1.76$; $p = 0.226$), or **HALF** ($F(1,747) = 0.01$; $p = 0.939$). Additionally, there was a significant interaction effect for **STRCT × SEX** ($F(2,690) = 5.56$; $p < 0.01$). Hence, the distance between interlocutors' F0 ranges was closer in sentences arranged by groups than in sentences arranged by feet. Additionally, F0 range distances were closer in metrically regular sentences compared to metrically irregular sentences, especially in the female / male dyads, and, to a lesser extent, in the female / female dyads. Conversely, metrically irregular sentences were slightly closer in terms of F0 range distances with respect to regular sentences in male / male dyads (See Graphic 5). In this respect, when the model was calculated without the female / male dyads, the interaction effect for **STRCT × SEX** disappeared ($F(1,439) = 0.57$; $p = 0.45$) (With the **UNIT** interaction remaining). This fact indicates that the **STRCT × SEX** effect must only be considered for mixed dyads (F0R distances closer in metrically regular sentences).

**Graphic 5:** F0 range distance (F0RD) by regularity of the sentences (STRCT) and type of dyad (SEX) (**FF:** Female / Female; **FM**: Female / Male; **MM:** Male / Male).



## F0 mean (F0M)

*Influential points*

None of the Items was found to exceed the established cut-off value for Cook's distance ($D_i > 4/n$), but the participants FMC1 (female), FMB2 (male), and FMC2 (male), did exceed it. The recordings of these participants with scores of F0M at least two standard deviations higher or lower than the female or male average, respectively, were measured a second time (i.e. influential points of FMC1 were measured with respect to the rest of female participants' F0M, and FMB2 and FMC2 with respect to the rest of male participants). This was the case in 12 sentences, approximately 0.8% of the total number of sentences. No evidence was found for errors in the recordings or in the measurement procedure. Consequently, all points of analysis were retained in the model.

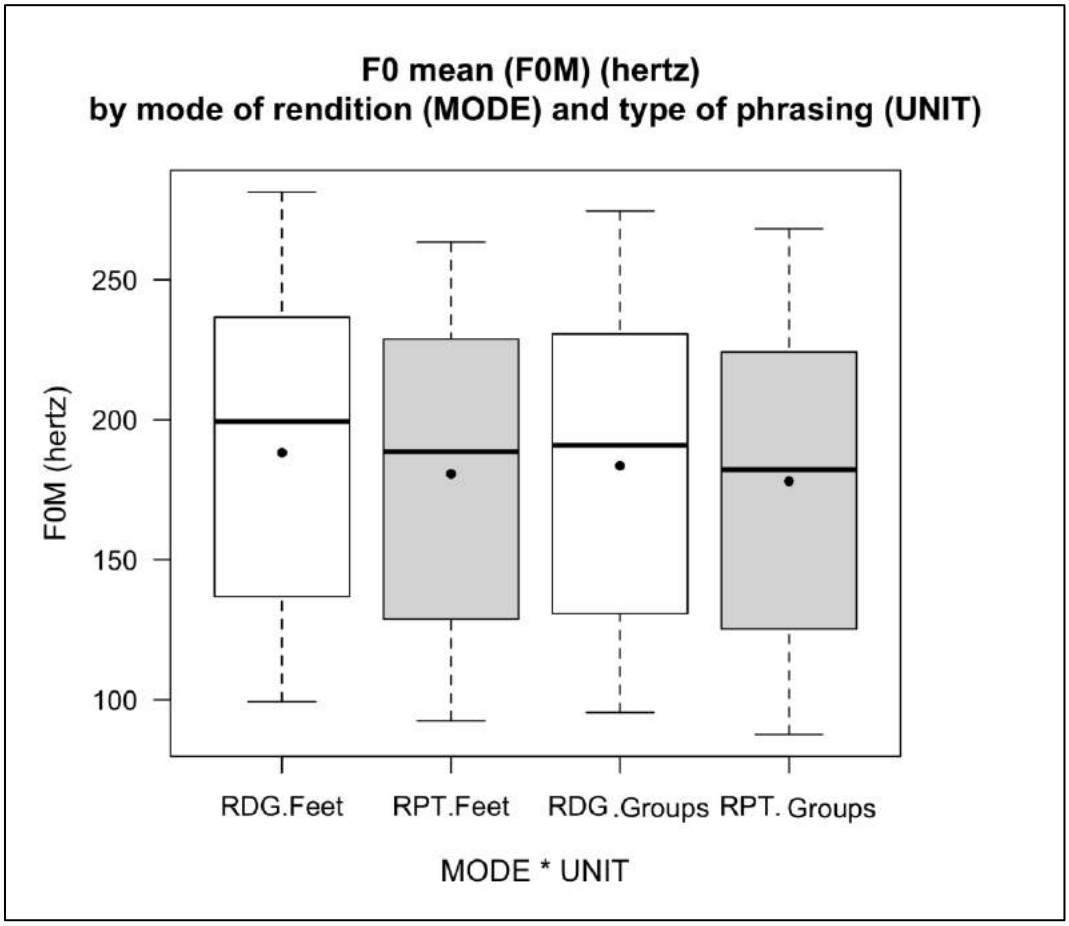
*Transformations*

Inspection of residual plots in this model revealed a significant deviation from homoscedasticity. The square root transformation (*sqrt*) managed to solve the problem.

*Results*

There were significant effects of the factors **STRCT** ($F(1,50) = 7.45$; $p < 0.01$), **UNIT** ($F(1,47) = 11.13$; $p < 0.01$), **SEX** ($F(2,21) = 13.84$; $p < 0.001$), and **MODE** ($F(1,1378) = 478.48$; $p < 0.001$), but not of **HALF** ($F(1,1418) = 0.07$; $p = 0.794$). Additionally, there was a significant interaction effect for **MODE × UNIT** ($F(1,1378) = 10.99$; $p < 0.001$). Consequently, participants' F0 means were lower in metrically regular sentences compared to metrically irregular sentences, and in sentences arranged by group compared to sentences arranged by feet. Also, participants' F0 means were lower in the male / male dyads compared to the female / female dyads. Furthermore, given that female / male dyads' F0 means are an average between a female participant and a male participant, mixed dyads scores are around the middle between female / female and male / male dyads (see Table 6, above). Additionally, participants' F0 means were lower during repetition of sentences compared to reading of sentences, and this effect was greater in sentences arranged by feet than in sentences arranged by groups (See Graphic 6).

**Graphic 6:** F0 mean (F0M) (hertz) by mode of rendition (MODE) and type of phrasing (UNIT) (**RDG:** Reading; **RPT:** Repeating).




## F0 mean distance (F0MD)

*Influential points*

Given that influential points on F0M have already been examined, all points of analysis were retained in this model.

*Transformations*

Inspection of residual plots in this model revealed a significant deviation from homoscedasticity. The square root transformation (*sqrt*) managed to solve the problem.

*Results*

There was a significant effect of the factor **SEX** ($F(2,13) = 7.68$; $p < 0.01$), but not of **STRCT** ($F(1,12) = 0.13$; $p = 0.726$), **UNIT** ($F(1,15) = 0.00$; $p = 0.953$), or **HALF** ($F(1,725) = 0.01$; $p = 0.911$). No significant interactions between predictors were found. Hence, the distance between interlocutors' F0 means was closer in female / female and male / male dyads with respect to female / male dyads. As mentioned earlier, this is an expected result because mixed dyads' scores are calculated comparing F0 means of a woman with F0 means of a man, which are "far" between each other, in terms of Euclidean distances, and "further away" with respect to same sex dyads. Moreover, when the model was calculated without the female / male dyads, the significant effect of the factor **SEX** was lost ($F(1,9) = 0.58$; $p = 0.465$). This fact indicates that there was no difference between same sex dyads with respect to F0 mean distance.

## Speech rate (SR)

*Influential points*

Influential points for this model were calculated from the total length of each sentence, before converting it to syllables per second. None of the Items was found to exceed the established cut-off value for Cook's distance ($D_i > 4/n$), but participants FFB1 and FMA1, both females, did exceed it. The recordings of these participants with scores of SR at least two standard deviations higher or lower than the female average were measured a second time. This was the case in 34 sentences, approximately 2.2% of the total number of sentences. A closer look at these recording revealed a particularly slow pace of pronunciation of both participants FFB1 and FMA1 (represented as longer sentences' durations). Moreover, 28 out of the 34 influential sentences were uttered during the first half of the turn: that is to say, they were read. Other than that, no evidence was found for errors in the recordings or in the measurement procedure.

Given that the speech rate of participants FFB1 and FMA1 was fairly below the women's average, the model was calculated three times: (1) without modifications; (2) converting scores two standard deviations higher than the female average to the female average (34 sentences, approximately 2.2% of the total number of sentences); and (3) converting scores three standard deviations higher than the female average to the female average (11 sentences, approximately 0.7% of the total). All three models present significant effects for

**HALF** and **SEX**, and a significant interaction for **HALF × SEX** (see details below). However, in both trimmed models the significant interaction **UNIT × HALF** is lost with respect to the complete model. Therefore, given that the **UNIT × HALF** interaction seems to depend on the particularly slow reading pace of two participants out of 24, and that there are not other major difference between the three versions of the model, it was decided to keep the third version (speech rate of participants FMA1 and FFB1 three standard deviations higher than the female average converted to the female average).

*Transformations*

Inspection of residual plots did not reveal significant deviations from considered assumptions. Thus, the model is based on non-transformed data.

*Results*

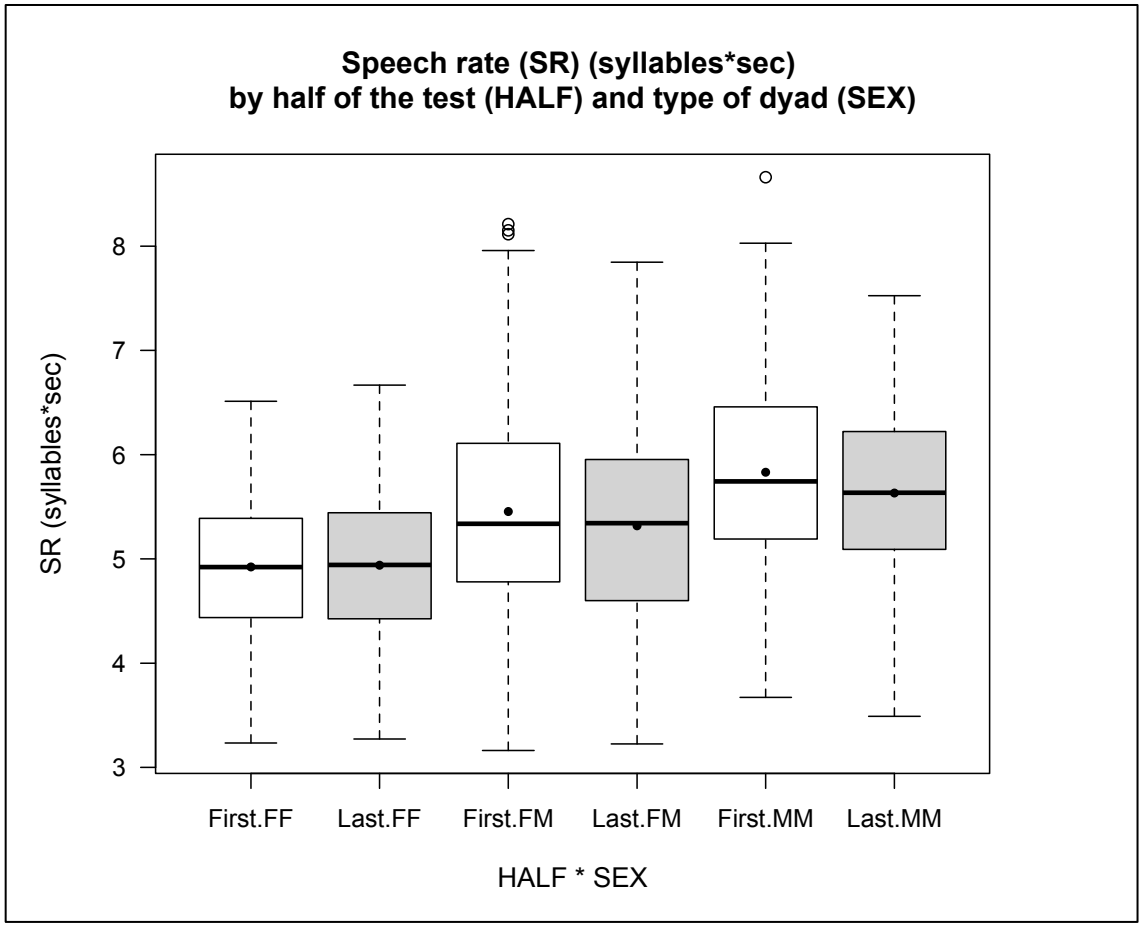
There were significant effects of the factors **SEX** ($F(2,21) = 4.23$; $p < 0.05$), **HALF** ($F(1,1388) = 11.11$; $p < 0.001$), and **MODE** ($F(1,1376) = 251.55$; $p < 0.001$), but not of **STRCT** ($F(1,72) = 1.65$; $p = 0.203$) or **UNIT** ($F(1,70) = 3.33$; $p = 0.072$). Additionally, there was a significant interaction effect for **HALF × SEX** ($F(2,1389) = 3.93$; $p < 0.05$). Hence, speech rate was higher (participants spoke faster) during repetition of sentences than during reading of sentences, and in the male / male dyads compared to the female / female dyads. In this regard, given that female / male dyads' speech rates are an average between a female participant and a male participant, mixed dyads scores are around the middle between female / female and male / male dyads (see Table 6, above). Additionally, speech rate was higher during the first half of the interaction compared to the last half, especially in the male / male dyads, and, to a lesser extent, in the female / male dyads. Conversely, speech rate was slightly lower during the first half of the interaction compared to the last half in female / female dyads (See Graphic 7).

## Speech rate distance (SRD)

*Influential points*

Given that influential points on SR have already been examined, all points of analysis were retained in this model.

**Graphic 7:** Speech rate (SR) (syllables*sec) by half of the test (HALF) and type of dyad (SEX) (**FF:** Female / Female; **FM**: Female / Male; **MM:** Male / Male).




*Transformations*

Inspection of residual plots in this model revealed a significant deviation from homoscedasticity. Tukey ladder of powers (*T_tuk*) managed to solve the problem.

*Results*

There were no significant effects of any of the factors: **STRCT** ($F(1,61) = 0.50$; $p = 0.484$), **UNIT** ($F(1,61) = 0.12$; $p = 0.730$), **SEX** ($F(2,9) = 0.50$; $p = 0.620$), or **HALF** ($F(1,749) = 0.01$; $p = 0.919$). Also, no significant interactions between predictors were found.

# Lexical repetitions (LR)

*Influential points*

The procedure to detect influential data in logistic regressions discuted in Zhang (2016) was followed. No influential points were found. All points of analysis were retained in the model.

*Transformations*

Inspection of residual plots did not reveal significant deviations from considered assumptions. Thus, the model is based on non-transformed data.

*Results*

The factor **SEX** was statistically significant ($\chi 2 = 6.2778$, $df = 2$, $p < 0.05$), indicating that there were more lexical repetitions between the participants of same sex dyads with respect to the participants of mixed sex dyads.

## Survey

In the following we present briefly the most important participants' remarks. Each remark is followed by a number in parenthesis indicating the number of participants sharing that particular thought. We refer to the blocks of stimuli using their actual name, but, of course, participants did not know such names and referred to each block as "the first one", or "the one in the middle", or something similar:

- Articles such as "*las*", and pronouns such as "*les*" and "*nos*", were difficult to utter (1) / to read (8) / to repeat (3). [Please note that such articles and pronouns were present in all blocks of stimuli, but within irregular groups and regular feet there were combinations of two of them that seem to have hindered the task.]
- Regular feet were more difficult to read (3).
- Irregular groups were more difficult to read (2) / to repeat (1).
- Regular groups were more difficult to read (2).
- Irregular feet were more difficult to read (2).
- Irregular groups were easier to utter (in general; not distinction between reading and repeating was made by the participants) (2) / to read (1).
- Regular groups were easier to utter (1) / to read (1).
- Irregular feet were easier to read (1).

- Sentences within the Regular feet block were more common (in informal speech) (2).
- There were repeated sentences (2). [In fact, there were not.]
- It was easier to repeat than to read (2).
- Sentences within the Irregular groups block were longer (1).
- There were no semantic coherence in some sentences (1).
- There were sentences with the same meaning (1). [In fact, there were not.]
- Accent (or diacritic) marks (specifically, the acute mark [´]) made sentences difficult to read (1). [Note that accent marks were present in all blocks of stimuli, except in the Regular groups block.]

# 4.4.2. Experiment 2: Perceptual evaluation

## Perceptual rating (PR)

*Influential points*

None of the Items was found to exceed the established cut-off value for Cook's distance ($D_i > 4/n$), but the evaluator F8's rating reports did exceed it. These rating reports were reviewed a second time but no evidence was found for errors in the rating procedure. Consequently, all points of analysis were retained in the model.

*Transformations*

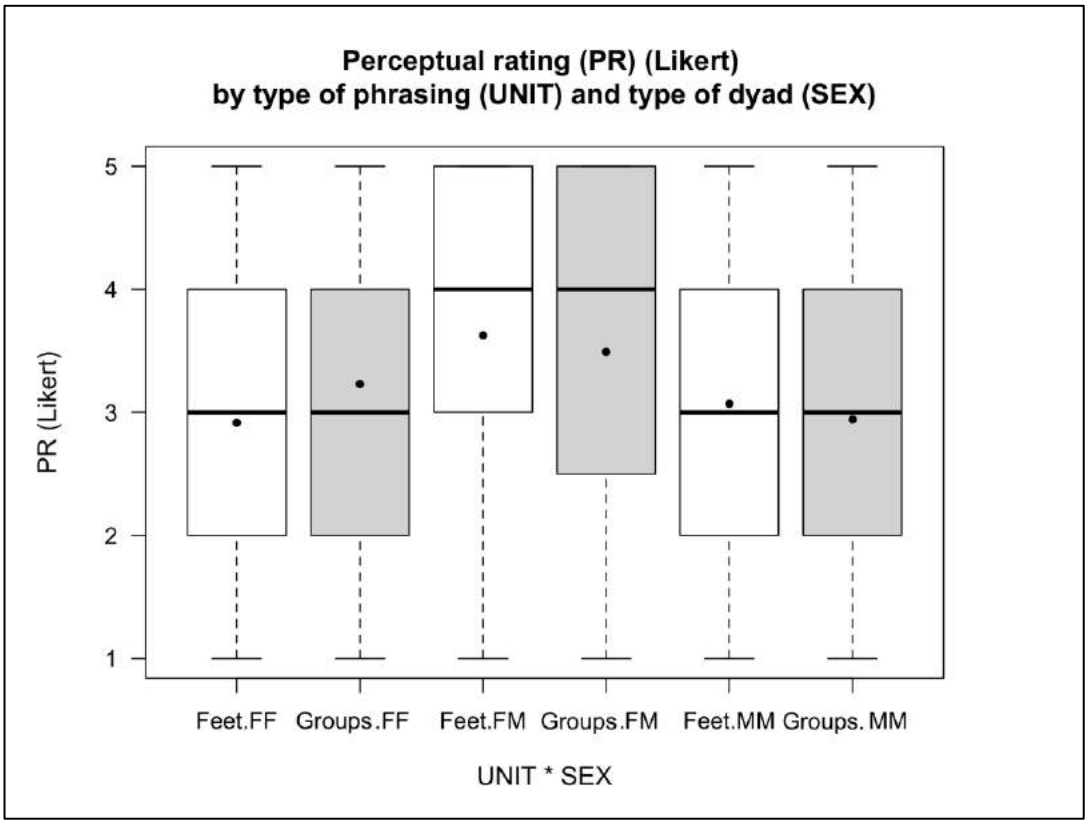
Inspection of residual plots did not reveal significant deviations from considered assumptions. Thus, the model is based on non-transformed data.

*Results*

There were significant effects of the factors **SEX** ($F(2,10) = 11.71$; $p < 0.01$) and **HALF** ($F(1,301) = 5.86$; $p < 0.05$), but not of **STRCT** ($F(1,16) = 0.47$; $p = 0.503$) or **UNIT** ($F(1,11) = 0.11$; $p = 0.751$). Additionally, there was a significant interaction effect for **UNIT × SEX** ($F(2,11) = 4.58$; $p < 0.05$). Hence, sentences of mixed dyads were rated more similar to each other with respect to sentences of female only and male only dyads, and sentences uttered during the first half of the interaction were rated more similar to each other with respect to sentences uttered during the last half. Additionally, sentences were rated more similar to each other when they were arranged by feet than when they were arranged by groups, in

female / male and male / male dyads. Conversely, sentences were rated more similar to each other in terms of rhythm when they were arranged by groups than when they were arranged by feet in female / female dyads (See Graphic 8).

**Graphic 8:** Perceptual rating (PR) (Likert) by type of phrasing (UNIT) and type of dyad (SEX) (**FF:** Female / Female; **FM**: Female / Male; **MM:** Male / Male).




## Survey

In the following we present briefly the most important participants' remarks. Each remark is followed by a number in parenthesis indicating the number of participants sharing that particular thought. Note that participants in this perceptual task were not aware that the first sentence of each pair was read. There are some remarks that we do not include, such as "there were differences in the tone of voice of some of the dyads", because participants made them even after being asked not to consider such type of differences when scoring the sentences:

- Some persons omitted the final “s” in some words (5) / added a final “s” in some words (1) / lengthened the final “s” in some words (1).
- There were different paces of speech (6). [In this case participants seem to refer to differences between persons rather than differences between types of sentences.]
- Some words were changed during repetitions (4).
- Some participants made some mistakes (3).
- There were some utterances more difficult to utter than others (2).
- Articles such as “las”, and pronouns such as “les” and “nos”, made the sentences different between each other (1).
- Sometimes there were pauses within sentences (1).

# 5. Discussion and conclusions

With respect to the expected results, the first two of them ($E_1$ and $E_2$; see Section 4.1.), related to the differences between women and men in terms of pitch mean and range, resulted to be accurate. That is to say, in line with the data provided by Chen (2007), Hudson and Holbrook (1982), Pépiot (2014), and Traunmüller and Eriksson (1994), in our experiment, pitch means of Spanish speakers were higher in women than in men, and pitch ranges were broader in women than in men. Note that, in our research, men and women were tested speaking to both other men and women. Additionally, the intonation contours were controlled, as well as (arguably) the quantity and location of pitch accents (which were aligned with lexical stresses).

$E_3$ and $E_4$ also resulted to be correct. In other words, in consonance with the data indicating that reading F0 is higher than spontaneous speaking F0 (Hudson & Holbrook, 1982; Horii, 1982; Mysak, 1959; Snidecor, 1943), and that reading has a F0 range broader than speaking (measured in tones) (Mysak, 1959; Snidecor, 1943), in our experiment F0 ranges of Spanish speakers were broader in reading of sentences than in repetition, and F0 means were higher during reading of sentences than during repetition. Additionally, this difference in F0 means was greater in sentences arranged by feet than in sentences arranged by groups.

Likewise, $E_5$ was no mistaken either. In keeping with studies indicating that men tend to speak slightly faster than women during spontaneous and read speech (Binnenpoorte et al., 2005; Byrd, 1994; Cohen et al., 2017; Weirich & Simpson, 2014; Yuan et al., 2006), our data showed that speech rate was, in general terms, faster in men than in women. However, regarding $E_6$, in contrast to research indicating that spontaneous speaking is slower than reading in both women and men (Lass & Sandusky, 1971; Snidecor, 1943), we found that participants spoke faster during repetition of sentences than during reading of sentences. Note that in our study all sentences (spoken and read) had the same length, and also they did not have punctuation or any indication of pauses within them (nonetheless, “irregular” pauses were made within some sentences, as showed by the acoustic analyses and as reported by one of the participants in the survey following perceptual evaluation).

We also found that, for female / female dyads, F0 ranges were narrower and speech rate was higher during the last half of the interaction, compared to the first half. This behavior constitutes one of the few time effects of accommodation found in our research, and may be seen as an effect of fatigue generated by the task. In this sense, speech rate may have become faster towards the end of the task in order to finish as soon as possible, and the range of F0 may have become flatter as usually seen in bored people' speech. However, with respect to the speech rate, the opposite trend was observed in male / male and mixed dyads, which tended to speak slower towards the end of the task.

As for the hypothesized results, in two words, we put forward the idea that conversational interactions involving regular rhythmic sentences would generate greater amount of lexical repetitions, a shorter delay of response, and more similarity (in terms of closer Euclidean distances) regarding rhythm, average pitch, pitch range, and speech rate, compared to interactions involving irregular rhythmic sentences ($H_1$ to $H_6$). We also hypothesized the same results for conversational interactions involving sentences arranged in accentual groups compared to interactions involving sentences arranged in accentual feet ($H_7$ to $H_{12}$).

In our opinion, the most important result of this research work was the fact that sentence-level temporal regularity allowed (or generated) further rhythmic resemblance between Spanish speakers (at least under the terms and conditions established in this thesis). This result is in line with the findings that Späth et al. (2016) presented for German healthy speakers and patients with Parkinson's disease (Section 2.4.4.). Of the same importance is the fact that the effect just mentioned (rhythmic distance being closer in metrically regular compared to irregular sentences) was greater in sentences arranged by groups than in sentences arranged by feet. This result (considering that temporal regularity enables rhythmic resemblance) is consistent with the reports of several authors indicating a greater amount of temporal regularity of accentual groups, regarding accentual feet, in Spanish (Almeida, 1997; Mora et al., 1999; Toledo, 1988).

One possible explanation of the accentual group allowing closer rhythmic distances than the accentual foot relies on the concept of sirrema (discussed in Section 3.2.4.). According to this concept, structural relations between the components of a sentence determine to a

certain extent the way in which such sentence is phrased. Unlike accentual feet, which depend mostly in metric relations at the phonological level, accentual groups also rely on syntactic connections between words. In this scenario, the typical phrasing of the accentual group, deep-rooted in Spanish speakers, would have hindered the "correct" phrasing of the Regular feet block of stimuli (XxxXxxXxx), generating instead an accentual non-regular phrasing, and consequently less rhythmic resemblance.

In addition, the fact that the effect of metrical regularity was greater in accentual groups (regarding feet), which depend on relations at different linguistic levels, added to the fact that the surveys conducted after both experiments did not suggest that participants were aware of the dichotomy of phrasing and regularity, indicates that rhythmic accommodation relies on a multilevel non-conscious mechanism, such as the one described by the interactive alignment model (IAM; see Section 2.6.2.). Following the postulates of the IAM, the fact that the effect of regularity allowing rhythmic resemblance was greater in accentual groups (regarding feet) may have facilitated the greater resemblance between interlocutors' F0 ranges found also in accentual groups, compared to accentual feet. The same logic may be extended to the regular / irregular distinction, but in this case only mixed dyads exhibited closer pitch range distances in metrically regular sentences.

Apart from the results just mentioned, no other hypotheses regarding the effects generated by conversational interactions involving regular rhythmic sentences and sentences arranged in accentual groups were confirmed. Naturally, this could be due to at least two reasons: (1) there were in fact no differences between the experimental conditions with respect to such matters, and, more likely, (2) the methodological approach used in this research was unable to determine further differences between conditions. For example, no differences whatsoever were found regarding response times (interval times between hearing and repeating sentences). This fact could indicate that processing different types of rhythmic regularity and phonological phrasing implies a low-level intellectual activity that does not entail a different amount of cognitive workload.

This, of course, would be conflicting with the facts mentioned in Section 4.1., indicating that periodic utterances are processed faster than aperiodic utterances (Mooney & Sullivan, 2015), and that listeners are faster to process information within syllables that are

expected to bear stress based on metrical patterns in the preceding context (Brown et al., 2015). We think it is more likely that the difficulties in reading and pronunciation caused by the presence of articles such as "las", and pronouns such as "les" and "nos", within the utterances, reported by some of the participants in both experiments (see the surveys in Sections 4.4.1. and 4.4.2.), have influenced the overall times of response, despite the precautions taken during the statistical analysis of the data. In this respect, we envisage the employ of model talkers in future research, in a similar form to that of Späth et al.'s (2016) study. In this way, it is possible that we will be able to keep the "pure" rhythmic patterns that we intended to establish in this research work. Another possibility to attain and test such pure rhythmic patterns would be to test participants with respect to a computer-created, or theoretical, perfectly regular rhythm consisting in temporally equal segments (syllables, feet, or accentual groups) within sentences.

Note also that no convergence effects were found during the experiments. On the contrary, sentences uttered during the first half of the interaction were perceptually rated more similar to each other with respect to sentences uttered during the last half. At this regard, it must be noted that our approach to the measurement of accommodation was based on the concept of similarity rather than the concept of convergence. There are, conversely, a few approaches to the measurement of accommodation that focus on the temporal development of the phenomenon (see De Looze et al., 2014 for an overview), some of which we will definitely consider for further research. In any case, the results related to the effect of HALF (half of the test) in our research, added to visual inspections of the evolution of the levels of similarity between speakers during each experimental block, did not indicate an increase of similarity in any of the variables that were tested.

Another drawback of our approach consists in the restricted definition of rhythm that we established; specifically, rhythm as a temporally iteration of strong and weak values of lexical stress (i.e. stressed and unstressed syllables), which is constituted, in turn, by an unknown combination of duration, intensity, and pitch. In this sense, our approach is unable to manage different "types" of rhythm, for instance, speech rhythms in which pitch accents are not aligned with lexical stresses due to specifics patterns of intonation or the presence of secondary rhythm or lexical stresses (see Sections 3.2.1. and 3.2.5.). At this

respect, we envisage also the analysis of different kinds of speech rhythm in future research.

Furthermore, our results also indicate that sentences were perceptually rated more similar to each other when they were arranged by feet than when they were arranged by groups, in female / male and male / male dyads. Conversely, in agreement with the acoustic analyses, sentences were rated more similar to each other in terms of rhythm when they were arranged by groups than when they were arranged by feet in female / female dyads. These mixed results somehow confirm the already mentioned inconsistencies between acoustic and perceptual data in studies of linguistic accommodation (Pardo, 2013b; Ruch et al., 2017; see Section 2.4.1.). In our case, even though participants were specifically instructed to focus on rhythmic differences, their judgments were presumably based on a series of acoustic characteristics beyond the intended rhythmicity.

On the subject of gender differences, we found more lexical repetitions between participants of same sex dyads with respect to participants of mixed sex dyads. This result is partially in accordance with the study conducted by Street (1984; see Section 2.5.3.), who found that male / male dyads tended to converge with respect to turn duration, whereas mixed female / male dyads tended to diverge. On the other hand, in our perceptual evaluation experiment, sentences of mixed dyads were rated more similar to each other with respect to sentences of female only and male only dyads. In sum, as in some other studies (Kawasaki et al., 2013; Thomson et al., 2001 and references therein), we did not find significant differences between men and women in terms of accommodation.

Concerning the results that we termed possible in Section 4.1., our data showed that pitch ranges were narrower, and average pitch was lower, in metrically regular sentences compared to metrically irregular sentences and in sentences arranged by group compared to sentences arranged by feet. As we mentioned before, we had no grounds to hypothesize about the behavior of these acoustic-prosodic features (F0 range and mean) with respect to the rhythmic regularity and phonological phrasing manipulated in our experiments, and, to our best knowledge, we have also no grounds to explain these results. In any case, we think that it is possible that precisely the greater temporal regularity of the sentences arranged in accentual feet produced an effect of habituation that made the speech less dynamic and

flatter in terms of pitch range. The flatter pitch range, in turn, would be responsible for the lower F0 mean.

Finally, we hope that the overview of linguistic accommodation and of the rhythmic characteristics of Spanish, which we have provided in this research work, will be of help to investigators and academics interested in the subject. We also hope that we have not entirely failed in the endeavor of shedding some light on the knowledge about speech rhythm and linguistic accommodation in Spanish, and in general.

# 6. References


Abel, J., & Babel, M. (2016). Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech, 60*(3), 1-24. doi:10.1177/0023830916665652

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. *Psychological Science*, *20*(10), 1-7. doi:10.1177/0956797610389192

Alcoba, S. (2007). Usos de *cual*, grupo acentual y unidad melódica [Uses of "cual", accentual group and melodic unit]. *Moenia*, *13,* 39-68.

Alcoba, S., & Murillo, J. (1998). Intonation in Spanish. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems* (pp. 152-166). Cambridge: Cambridge University Press.

Almeida, M. (1993). Alternancia temporal y ritmo en español [Temporal alternation and rhythm in Spanish]. *Verba, 20,* 433-443.

Almeida, M. (1997). Organización temporal del español: El principio de isocronía [Temporal organization of Spanish: The isochronous principle]. *Revista de Filología Románica*, *14*(1), 29-40.

Aronoff, M., & Fudeman, K. (2011). *What is morphology?* (2nd ed.). Oxford: Wiley-Blackwell.

Arriaga, G., Zhou, E., & Jarvis, E. (2012). Of mice, birds, and men: The mouse ultrasonic song system has some features similar to humans and song-learning birds. *PLoS ONE*, *7*(10), 1-15. doi:10.1371/journal.pone.0046610

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*, 46-63. doi:10.1159/000208930

Arvaniti, A., & Ross, T. (2012). Rhythm classes and speech perception. *Phonetica, 66*(1-2), 1-15.

Aubanel, V., & Nguyen, N. (2010). Automatic recognition of regional phonological variation in conversational interaction. *Speech Communication*, *52*(6), 577-586. doi:10.1016/j.specom.2010.02.008

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. New York: Cambridge University Press.

Baayen, R., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12-28. doi:10.21500/20112084.807

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*, 437-456. doi:10.1017/s0047404510000400

Babel, M. (2011). Imitation in speech. *Acoustics Today*, *7*(4), 16-22. doi:10.1121/1.3684224

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*, 177-189. doi:10.1016/j.wocn.2011.09.001

Babel, M., & Bulatov, D. (2011). The role of fundamental frequency in phonetic accommodation. *Language and Speech, 55*(2), 231-248. doi: 10.1177/0023830911417695

Barón, L. (2016). Animal communication and human language: An overview. *International Journal of Comparative Psychology*, *29*, 1-27.

Barr, D. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. 1-2. doi:10.3389/fpsyg.2013.00328

Barrett, P. (2005). *Euclidean distance: Raw, normalized, and double-scaled coefficients* [White paper]. Retrieved from http://www.pbarrett.net/techpapers/euclid.pdf

Beckman, M., Díaz-Campos, M., McGory, J., & Morgan, T. (2002). Intonation across Spanish, in the Tones and Break Indices framework. *Probus*, *International Journal of Romance Linguistics, 14*(1), 9-36. doi:10.1515/prbs.2002.008

Beebe, B., Knoblauch, S., Rustin, J., & Sorter, D. (2003). A comparison of Meltzoff, Trevarthen, and Stern. *Psychoanalytic Dialogues*, *13*(6), 777-804. doi:10.1080/10481881309348768

Berwick, R., & Chomsky, N. (2011). The biolinguistic program: The current state of its evolution and development. In A. di Sciullo & C. Boeckx (Eds.), *The Biolinguistic Enterprise: New Perspectives on the Evolution and Nature of the Human Language Faculty* (pp. 19-41). Oxford: Oxford University Press.

Bigi, B. (2015). SPPAS - multi-lingual approaches to the automatic annotation of speech. *The Phonetician: A publication of ISPhS / International Society of Phonetic Sciences*, *111-112*, 54-69.

Binnenpoorte, D., Van Bael, C., den Os, E., & Boves, L. (2005). Gender in everyday speech and language: a corpus-based study. In *Proceedings of Interspeech 2005*, Lisbon, PT, 1-4.

Bloom, K. (1998). The missing link's missing link: Syllabic vocalizations at 3 months of age. *Behavioral and Brain Sciences*, *21*(4), 514-515. doi:10.1017/s0140525x98251260

Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*(3), 355-387. doi:10.1016/0010-0285(86)90004-6

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer* [Computer program]. Version 6.0.20, retrieved from http://www.praat.org/

Bolinger, D. (1962). *"Secondary stress" in Spanish*. *Romance Philology, 15*(3), 273-279.

Bolinger, D. & Hodapp, M. (1961). Acento melódico. Acento de intensidad [Melodic stress. Intensity stress]. *Boletín de Filología de la Universidad de Chile, 13*, 33-48.

Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., & Campbell, N. (2013). Investigating fine temporal dynamics of prosodic and lexical accommodation. In *Proceedings of Interspeech 2013*, Lyon, FR, 539-543.

Borrie, S., & Liss, J. (2014). Rhythm as a coordinating device: Entrainment with disordered speech. *Journal of Speech, Language, and Hearing Research, 57*(3), 815-824. doi:10.1044/2014_jslhr-s-13-0149

Borrie, S., Lubold, N., & Pon-Barry, H. (2015). Disordered speech disrupts conversational entrainment: A study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. *Frontiers in Psychology*, *6:1187*, 1-8. doi:10.3389/fpsyg.2015.01187

Borzone, A., & Signorini, A. (1983). Segmental duration and rhythm in Spanish. *Journal of Phonetics*, *11*(2), 117-128.

Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication, 47*, 80-86. doi:10.1016/j.specom.2005.05.009

Branigan, H., Pickering, M., & Cleland, A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*, B13-B25. doi:10.1016/s0010-0277(99)00081-5

Branigan, H., Pickering, M., Pearson, J., McLean, J., & Nass, C. (2003). Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*, Boston, US, 186-191.

Brennan, S. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the 1996 International Symposium on Spoken Dialogue*, Philadelphia, US, 41-44.

Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482-1493.

Brown, M., Salverda, A., Dilley, L., & Tanenhaus, M. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 306-323. doi:10.1037/a0038689

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*, 1435. doi:10.1126/science.1069587

Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication, 15*(1-2), 39-54. doi:10.1016/0167-6393(94)90039-6

Candia, L., Urrutia, H., & Fernández, T. (2006). Rasgos acústicos de la prosodia acentual del español [Acoustic features of Spanish accentual prosody]. *Boletín de Filología*, *41*, 11-44.

Cantero, F. (2002). *Teoría y análisis de la entonación* [Theory and analysis of intonation]. Barcelona: Ediciones Universitat de Barcelona.

Cantero, F. (2003). Fonética y didáctica de la pronunciación [*Phonetics and teaching of pronunciation*]. In A. Mendoza (Coord.), Didáctica de la Lengua y la Literatura [*Teaching of Language and Literature*] (pp. 545-572). Madrid: Prentice Hall.

Cardwell, R. (2016). *Specifying the random effect structure in linear mixed effect models for analyzing psycholinguistic data* (Master's thesis), Retrieved from

http://digitool.library.mcgill.ca/webclient/StreamGate?folder_id=0&dvs=1530864118342~119

Chartrand, T., & Bargh, J. (1999). The Chameleon Effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893-910. doi:10.1037/0022-3514.76.6.893

Chen, S. (2007). Sex differences in frequency and intensity in reading and voice range profiles for Taiwanese adult speakers. *Folia Phoniatrica et Logopaedica, 59*(1), 1-9. doi:10.1159/000096545

Chomsky, N. (2010). Some simple evo-devo theses. How true might they be for language? In R. Larson, V. Deprez & H. Yamakido (Eds.), *The Evolution of Human Language* (pp. 45-62). Cambridge: Cambridge University Press.

Clark, H., & Wilkes, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1-39.

Cohen, U., Edelist, L., & Gleason, E. (2017). Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America, 141*(5), 2989-2996. doi:10.1121/1.4982199

Collins, B. (1998). Convergence of fundamental frequencies in conversation: If it happens, does it matter?. In *Fifth International Conference on Spoken Language Processing*, Sydney, AU, 1-4.

Condon, W., & Sander, L. (1974). Synchrony demonstrated between movements of the neonate and adult speech. *Child Development*, *45*(2), 456-462. doi:10.2307/1127968

Coulston, R., Oviatt, S., & Darves, C. (2002). Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, US, 2689-2692.

Couper-Kuhlen, E. (1993). *English speech rhythm: Form and function in everyday verbal interaction*. Amsterdam: John Benjamins Publishing.

Cummins, F. (2009a). Rhythm as an affordance for the entrainment of movement. *Phonetica*, *66*(1-2), 15-28. doi:10.1159/000208928

Cummins, F. (2009b). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*(1), 16-28. doi:10.1016/j.wocn.2008.08.003

Dauer, R. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*, 51-62.

Davis, C., & Kim, J. (2018). Characterizing rhythm differences between strong and weak accented L2 speech. In *Proceedings of Interspeech 2018*, Hyderabad, IN, 2568-2572.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *Neuropsychiatrie de l'Enfance et de l'Adolescence*, *60*(5S), 1-20. doi:10.1016/j.neurenf.2012.05.105

De Looze, C., & Rauzy, S. (2011). Measuring speakers' similarity in speech by means of prosodic cues: Methods and potential. In *Proceedings of Interspeech 2011*, Florence, IT, 1393-1396.

De Looze, C., Oertel, C., Rauzy, S., & Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *17th International Congress of Phonetic Sciences*, Hong Kong, CN, 1294-1297.

De Looze, C., Scherer, S., Vaughan, B., & Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, *58*, 11-34. doi:10.1016/j.specom.2013.10.002

Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, *64*(2-3), 145-173. doi:10.1159/000107914

Díaz-Campos, M. (2000). The phonetic manifestation of secondary stress in Spanish. In *Papers from the 3rd Hispanic Linguistic Symposium*, Somerville, US, 49-65.

Dijksterhuis, A., & Bargh, J. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. In M. Zanna (Ed.), *Advances in Experimental Social Psychology* (pp. 1-40). San Diego: Academic Press.

Doupe, A., & Kuhl, P. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience, 22*, 567-631. doi:10.1146/annurev.neuro.22.1.567

Duran, N., & Fusaroli, R. (2017). Conversing with a devil's advocate: Interpersonal coordination in deception and disagreement. *PLoS ONE, 12*(6), 1-25. doi:10.1371/journal.pone.0178140

Duranton, C., & Gaunet, F. (2016). Behavioural synchronization from an ethological perspective: Overview of its adaptive value. *Adaptive Behavior*, *24*(3), 1-11. doi:10.1177/1059712316644966

Edlund, J., Heldner, M., & Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009*, Brighton, UK, 2779-2782.

Eriksson, A., & Wretling, P. (1997). How flexible is the human voice? – A case study of mimicry. In *Proceedings of Eurospeech '97*, Rhodes, GR, 1043-1046.

Evans, B., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America, 121*(6), 3814-3826. doi:10.1121/1.2722209

Face, T. (2003). Intonation in Spanish declaratives: Differences between lab speech and spontaneous speech. *Catalan Journal of Linguistics*, *2*, 115-131.

Falk, S., Rathcke, T., & Dalla Bella, S. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1491-1506. doi:10.1037/a0036858

Feldman, R., Mayes, L., & Swain, J. (2005). Interaction synchrony and neural circuits contribute to shared intentionality. *Behavioral and Brain Sciences, 28*(5), 23-24. doi:10.1017/s0140525x0529012x

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477-501.

Ferrer, R. (2004). Euclidean distance between syntactically linked words. *Physical Review, E70*, 1-5. doi:10.1103/physreve.70.056135

Fine, A., Jaeger, T., Farmer, T., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*(10), 1-18. doi:10.1371/journal.pone.0077661

Fitch, W. (2005). The evolution of language: A comparative review. *Biology and Philosophy*, *20*, 193-230. doi:10.1007/s10539-005-5597-1

Fitch, W. (2010). *The evolution of language*. New York: Cambridge University Press.

Flores, M. (2004). *The role of catalexis in Spanish rhythm structure: A phonological and phonetic study of catalexis in Spanish syllable-timed poetry* (Master's thesis), Retrieved from https://lup.lub.lu.se/student-papers/search/publication/1330190

Flores, M., & Horne, M. (2003). Evidence for the metrical foot in Spanish. In *34th Poznań Linguistic Meeting,* Poznań, PL, 16-17.

Freud, D., Ezrati, R., & Amir, O. (2018). Speech rate adjustment of adults during conversation. *Journal of Fluency Disorders, 57*, 1-10. doi:10.1016/j.jfludis.2018.06.002

Fuchs, R. (2016). *Speech rhythm in varieties of English. Evidence from educated Indian English and British English*. Singapore: Springer. doi:10.1007/978-3-662-47818-9

Galantucci, B., Fowler, C., & Turvey, M. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361-377. doi:10.3758/bf03193857

Gambi, C., & Pickering, M. (2013). Prediction and imitation in speech. *Frontiers in Psychology*, *4*, 1-9. doi:10.3389/fpsyg.2013.00340

Garrido, J. (2012). Análisis fonético de los patrones melódicos locales en español: Patrones acentuales [Phonetic analysis of the local melodic patterns in Spanish: Accentual patterns]. *Revista Española de Lingüística*, *42*(1), 79-107.

Garrido, J., Llisterri, J., de la Mota, C., & Ríos, A. (1995). Estudio comparado de las características prosódicas de la oración simple en español en dos modalidades de lectura [Comparative study of the prosodic features of Spanish simple sentences in two reading modalities]. In A. Elejabeitia & A. Iribar (Eds.), *Phonetica. Trabajos de Fonética Experimental* [Phonetica. Works on Experimental Phonetics] (pp. 173-194). Bilbao: Universidad de Deusto.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*(2), 181-218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8-11. doi:10.1016/j.tics.2003.10.016

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel / hierarchical models*. New York: Cambridge University Press.

Gentilucci, M., & Bernardis, P. (2007). Imitation during phoneme production. *Neuropsychologia, 45*(3), 608-615. doi:10.1016/j.neuropsychologia.2006.04.004

Gervain, J., & Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annual Review of Psychology*, *61*(1), 191-218. doi:10.1146/annurev.psych.093008.100408

Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., & Steiner, I. (2018). Convergence of pitch accents in a shadowing task. In *9th International Conference on Speech Prosody 2018*, Poznań, PL, 225-229 doi:10.21437/speechprosody.2018-46

Gibbon, D. (2015). Speech rhythms – modeling the groove. In R. Vogel & R. van de Vijver (Eds.), *Rhythm in Cognition and Grammar: A Germanic Perspective* (pp. 108-161). Berlin: De Gruyter Mouton.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, *5*(8), 509-524. doi:10.1111/j.1749-818x.2011.00295.x

Gil, J., & Llisterri, J. (2004). Fonética y fonología del español en España (1978-2003) [Phonetics and phonology of Spanish in Spain]. *Lingüística Española Actual*, *26*(2), 5-44.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. *Contexts of Accommodation: Developments in Applied Sociolinguistics*, *1*, 1-68.

Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access, *Psychological Review, 105*(2), 251-279.

Goldinger, S., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review, 11*(4), 716-722. doi:10.3758/bf03196625

Goswami, U., & Leong, V. (2013). Speech rhythm and temporal structure: Converging perspectives? *Laboratory Phonology, 4*(1), 67-92. doi:10.1515/lp-2013-0004

Grau, A. (2013). Reconsidering syllabic minimality in Spanish truncation. *ELUA - Estudios de Lingüística Universidad de Alicante*, *27*, 121-143. doi:10.14198/elua2013.27.05

Gregory, S., & Hoyt, B. (1982). Conversation partner mutual adaptation as demonstrated by Fourier series analysis. *Journal of Psycholinguistic Research*, *11*(1), 35-46.

Gregory, S., & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology, 70*(6), 1231-1240.

Gregory, S., Webster, S., & Huang, G. (1993). Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. *Language & Communication, 13*(3), 195-217. doi:10.1016/0271-5309(93)90026-j

Guardiola, M., & Bertrand, R. (2013). Interactional convergence in conversational storytelling: When reported speech is a cue of alignment and/or affiliation. *Frontiers in Psychology*, *4:705*, 1-17. doi:10.3389/fpsyg.2013.00705

Guasti, M. (2002). *Language acquisition: The growth of grammar*. Cambridge, MA: MIT Press.

Harris, M. (2015). *Quantifying speech rhythms: Perception and production data in the case of Spanish, Portuguese, and English* (Doctoral dissertation). Retrieved from http://search.proquest.com/docview/1679467612

Hay, J., Jannedy, S., & Mendoza, N. (1999). Oprah and /ay/: Lexical frequency, referee design and style. In *Proceedings of the 14th International Congress of Phonetic Sciences*, Berkeley, US, 1389-1392.

Heath, J. (2014). Accommodation can lead to innovated variation. *University of California, Berkeley. Phonology Lab Annual Report,* 119-145.

Heath, J. (2015). Convergence through divergence: Compensatory changes in phonetic accommodation. In *LSA 2015 Annual Meeting Extended Abstracts*, Portland, US, 1-4.

Heldner, M., Edlund, J., & Hirschberg, J. (2010). Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech 2010*, Makuhari, JP, 3054-3057.

Hess, U., & Blairy, S. (2001). Facial mimicry and emotional contagion to dynamic emotional facial expressions and their influence on decoding accuracy. *International Journal of Psychophysiology, 40*, 129-141. doi:10.1016/s0167-8760(00)00161-6

Himberg, T., Hirvenkari, L., Mandel, A., & Hari, R. (2015). Word-by-word entrainment of speech rhythm during joint story building. *Frontiers in Psychology, 6:797*, 1-5. doi:10.3389/fpsyg.2015.00797

Horii, Y. (1982). Some voice fundamental frequency characteristics of oral reading and spontaneous speech by hard-of-hearing young women. *Journal of Speech and Hearing Research, 25*, 608-610. doi:10.1044/jshr.2504.608

Hualde, J. (2007). Stress removal and stress addition in Spanish. *Journal of Portuguese Linguistics, 5/6*, 59-89. doi:10.5334/jpl.145

Hualde, J. (2009). Unstressed words in Spanish. *Language Sciences*, *31*(2-3), 199-212. doi:10.1016/j.langsci.2008.12.003

Hualde, J. (2010). Secondary stress and stress clash in Spanish. In M. Ortega (Ed.), *Selected Proceedings of the Fourth Conference on Laboratory Approaches to Spanish Phonology* (pp. 11-19). Somerville, MA: Cascadilla Proceedings Project.

Hualde, J. (2012). Stress and rhythm. In J. Hualde, A. Olarrea & E. O'Rourke (Eds.), *The Handbook of Hispanic Linguistics* (pp. 153-171). Hoboken, NJ: Wiley-Blackwell.

Hualde, J. (2014). *Los sonidos del español* [The sounds of Spanish]. UK: Cambridge University Press.

Hualde, J., & Nadeu, M. (2014). Rhetorical stress in Spanish. In H. van der Hulst (Ed.), *Word Stress: Theoretical and Typological Issues* (pp. 228-252). Cambridge: Cambridge University Press.

Hudson, A., & Holbrook, A. (1982). Fundamental frequency characteristics of young black adults: Spontaneous speaking and oral reading. *Journal of Speech and Hearing Research, 25*, 25-28. doi:10.1044/jshr.2501.25

Hurford, J. (2012). *The origins of grammar: Language in the light of evolution II*. New York, NY: Oxford University Press.

Inui, N. (2018). *Interpersonal coordination.* Singapore, SG: Springer Nature.

Jones, M, Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science*, *13*(4), 313-319. doi:10.1111/j.0956-7976.2002.00458.x

Kappes, J., Baumgaertner, A., Peschke, C., & Ziegler, W. (2009). Unintended imitation in nonword repetition. *Brain and Language, 111*(3), 140-151. doi:10.1016/j.bandl.2009.08.008

Kawasaki, M., Yamada, Y., Ushiku, Y., Miyauchi, E., & Yamaguchi, Y. (2013). Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Scientific Reports*, *3*(1692), 1-8. doi:10.1038/srep01692

Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta Psychologica, 32*, 100-125. doi:10.1016/0001-6918(70)90094-6

Kim, M., Horton, W., & Bradlow, A. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology, 2*(1), 1-30. doi:10.1515/labphon.2011.004

King, S., Harley, H., & Janik, V. (2014). The role of signature whistle matching in bottlenose dolphins, *Tursiops truncatus*. *Animal Behaviour*, 96, 79-86. doi:10.1016/j.anbehav.2014.07.019

Kisilevsky, B., Hains, S., Brown, C., Lee, C., Cowperthwaite, B., Stutzman, S., Swansburg, M., Lee, K., Xie, X., Huang, H., Ye, H., Zhang, K., & Wang, Z. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development*, *32*, 59-71. doi:10.1016/j.infbeh.2008.10.002

Kizach, J. (2014). *Analyzing Likert-scale data with mixed-effects linear models: A simulation study*. Poster session presented at Linguistic Evidence 2014, Tübingen, DE.

Koban, L., Ramamoorthy, A., & Konvalinka, I. (2017). Why do we fall into sync with others? Interpersonal synchronization and the brain's optimization principle. *Social Neuroscience*, 1-9. doi:10.1080/17470919.2017.1400463

Kohler, K. (2009). Rhythm in speech and language: A new research paradigm. *Phonetica, 66*, 29-45. doi:10.1159/000208929

Konvalinka, I., Bauer, M., Stahlhut, C., Hansen, L, Roepstorff, A., & Frith, C. (2014). Frontal alpha oscillations distinguish leaders from followers: Multivariate decoding of mutually interacting brains. *NeuroImage*, *94*, 79-88. doi:10.1016/j.neuroimage.2014.03.003

Kousidis, S., Dorran, D., McDonnell, C., & Coyle, E., (2009). Times series analysis of acoustic feature convergence in human dialogues. In *Proceedings of SPECOM 2009*, St. Petersburg, RU, 1-6.

Kousidis, S., Dorran, D., Wang, Y., Vaughan, B., Cullen, C., Campbell, D., McDonnell, C., & Coyle, E. (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. In *Proceedings of Interspeech 2008*, Brisbane, AU, 22-26.

Kuznetsova, A. Brockhoff, P., & Christensen, R. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26. doi:10.18637/jss.v082.i13

Lass, N., & Sandusky, J. (1971). A study of the relationship of diadochokinetig rate, speaking rate and reading rate. *Today's Speech, 19*(3), 49-54. doi:10.1080/01463377109368992

Lee, C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P., & Narayanan, S. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Proceedings of Interspeech 2010*, Makuhari, JP, 793-796.

Lee, Y., Gordon, S., Parrell, B., Lee, S., Goldstein, L., & Byrd, D. (2018). Articulatory, acoustic, and prosodic accommodation in a cooperative maze navigation task. *PLoS ONE*, *13*(8), 1-26. doi:10.1371/journal.pone.0201444

Lelong, A., & Bailly, G. (2011). Study of the phenomenon of phonetic convergence thanks to speech dominoes. In A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud & A. Nihjolt (Eds.), *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue* (pp. 273-285). New York: Springer.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech 2011*, Florence, IT, 3081-3084.

Levitan, R., Gravano, A., & Hirschberg, J. (2011). Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics,* Portland, US, 113-117.

Levitan, R., Gravano, A., Wilson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2012*, Montreal, CA, 11-19.

Lewandowski, E., & Nygaard, L. (2018). Vocal alignment to native and non-native speakers of English. *The Journal of the Acoustical Society of America, 144*(2), 620-633. doi:10.1121/1.5038567

Lieberman, P., & Blumstein, S. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge: Cambridge University Press.

Llisterri, J., Machuca, M., de la Mota, C., Riera, M., & Ríos, A. (2003). The perception of lexical stress in Spanish. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, SP, 2023-2026.

Louwerse, M., Dale, R., Bard, E., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404-1426. doi:10.1111/j.1551-6709.2012.01269.x

Lubold, N., & Pon-Barry, H. (2014). A comparison of acoustic-prosodic entrainment in face-to-face and remote collaborative learning dialogues. In *Proceedings of the IEEE Spoken Language Technology Workshop 2014*, South Lake Tahoe, US, 288-293.

Lykartsis, A., Lerch, A., & Weinzierl, S. (2015). Analysis of speech rhythm for language identification based on beat histograms. *DAGA 2015,* Nuremberg, DE, 1-4.

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*(4), 499-511. doi:10.1017/s0140525x98001265

Mampe, B., Friederici, A., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology, 19,* 1994-1997. doi:10.1016/j.cub.2009.09.064

Manson, J., Bryant, G., Gervais, M., & Kline, M. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior, 34*(6), 419-426. doi:10.1016/j.evolhumbehav.2013.08.001

Marslen, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*, 522-523.

Marslen, W. (1985). Speech shadowing and speech comprehension. *Speech communication*, *4*, 55-73.

McGarva, A., & Warner, R. (2003). Attraction and social coordination: Mutual entrainment of vocal activity rhythms. *Journal of Psycholinguistic Research*, *32*(3), 335-354.

Merker, B., Madison, G., & Eckerdal, P. (2009). On the role and origin of isochrony in human rhythmic entrainment. *Cortex*, *45*(1), 4-17. doi:10.1016/j.cortex.2008.06.011

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, *109*, 168-173. doi:10.1016/j.cognition.2008.08.002

Mooney, S., & Sullivan, G. (2015). Investigating an acoustic measure of perceived isochrony in conversation: Preliminary notes on the role of rhythm in turn transitions. *University of Pennsylvania Working Papers in Linguistics, 21*(2), 128-135.

Mora, E., Villamizar, T., Blondet, M., & López, Y. (1999). Hacia una caracterización rítmica del español hablado en Venezuela [To a rhythmic characterization of the Spanish spoken in Venezuela]. *Boletín Antropológico*, *47*, 75-87.

Muir, K., Joinson, A., Cotterill, R., & Dewdney, N. (2016). Characterizing the linguistic chameleon: Personal and social correlates of linguistic style accommodation. *Human Communication Research*, *42*(3), 462-484. doi:10.1111/hcre.12083

Muir, K., Joinson, A., Cotterill, R., & Dewdney, N. (2017). Linguistic style accommodation shapes impression formation and rapport in computer-mediated communication.

*Journal of Language and Social Psychology*, *36*(5), 1-24. doi:10.1177/0261927x17701327

Mysak, E. (1959). Pitch and duration characteristics of older males. *Journal of Speech and Hearing Research*, *2*(1), 46-54. doi:10.1044/jshr.0201.46

Namy, L., Nygaard, L., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology, 21*(4), 422-432. doi:10.1177/026192702237958

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology, 32*(5), 790-804.

Navarro-Tomás, T. (2004/1918). *Manual de pronunciación española* [Handbook of Spanish pronunciation] (28th ed.). Madrid: Consejo Superior de Investigaciones Científicas.

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,* 756-766. doi:10.1037/0096-1523.24.3.756

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, Columbus, US, 169-172.

Neumann, R., & Strack, F. (2000). "Mood contagion": The automatic transfer of mood between persons. *Journal of Personality and Social Psychology, 79*(2), 211-223.

Nguyen, N., & Delvaux, V. (2015). Role of imitation in the emergence of phonological systems. *Journal of Phonetics*, *53*, 46-54. doi:10.1016/j.wocn.2015.08.004

Niederhoffer, K., & Pennebaker, J. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21*(4), 337-360. doi:10.1177/026192702237953

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*, 132-142. doi:10.1016/j.wocn.2010.12.007

Nieuwenhuis, R., Grotenhuis, M., & Pelzer, B. (2012). influence.ME: Tools for detecting influential data in mixed effects models. *R Journal*, *4*(2), 38-47.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625-632. doi:10.1007/s10459-010-9222-y

Oben, B., & Brône, G. (2015). What you see is what you do: on the relationship between gaze and gesture in multimodal alignment. *Language and Cognition*, *7*(4), 546-562. doi:10.1017/langcog.2015.22

Ohannesian, M. (2004). *La asignación del acento en castellano* [Stress assignment in Castilian] (Doctoral dissertation). Retrieved from https://dialnet.unirioja.es/servlet/tesis?codigo=5099

Ortega, M., & Prieto, P. (2007). Disentangling stress from accent in Spanish: Production patterns of the stress contrast in deaccented syllables. *Current Issues in Linguistic Theory*, 155-176. doi:10.1075/cilt.282.11ort

Pamies, A. (1999). Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages. *Language Design: Journal of Theoretical and Experimental Linguistics*, *2*, 103-130.

Pardo, J. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382-2393. doi:10.1121/1.2178720

Pardo, J. (2013a). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4:559*, 1-5. doi:10.3389/fpsyg.2013.00559

Pardo, J. (2013b). Reconciling diverse findings in studies of phonetic convergence. In *Proceedings of Meetings on Acoustics 2013,* Montreal, CA, 1-4.

Pardo, J., Gibbons, R., Suppes, A., & Krauss, R. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*, 190-197. doi:10.1016/j.wocn.2011.10.001

Pardo, J., Jay, I., & Krauss, R. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics, 72*(8), 2254–2264. doi:10.3758/app.72.8.2254

Pardo, J., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language, 69*(3), 183-195. doi:10.1016/j.jml.2013.06.002

Patterson, D., & Ladd, D. (1999). Pitch range modelling: Linguistic dimensions of variation. In *4th International Congress of Phonetic Sciences*, San Francisco, US, 1169-1172.

Pépiot, E. (2014). Male and female speech: A study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. In *Speech Prosody*, Dublin, IE, 305-309.

Pickering, M., & Ferreira, V. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427-459.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169-190.

Piña, R., & Díaz-Campos, M. (2005). *Revising secondary stress: A laboratory approach study of Spanish prosody*. Poster session presented at the Phonetics and Phonology in Iberia 2005 Conference, Barcelona, ES.

Planas, S. (2013). El grupo rítmico y el grupo fónico en la clase de ELE [Rhythm and phonic groups in Spanish as a foreign language class]. *Revista Internacional de Lenguas Extranjeras*, *2*, 67-80. doi:10.17345/rile201367-80

Pointon, G. (1980). Is Spanish really syllable-timed? *Journal of Phonetics*, *8*(3), 293-304.

Prat, Y., Taub, M., & Yovel, Y. (2015). Vocal learning in a social mammal: Demonstrated by isolation and playback experiments in bats. *Science Advances*, *1*, e1500019. doi:10.1126/sciadv.1500019

Prieto, P., & van Santen, J. (1996). Secondary stress in Spanish: Some experimental evidence. In C. Parodi, C. Quicoli, M. Saltarelli & M. Zubizarreta (Eds.), *Aspects of Romance Linguistics* (pp. 336-356). Washington: Georgetown University Press.

Prieto, P., Vanrell, M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*, *54*, 681-702. doi:10.1016/j.specom.2011.12.001

Putman, W., & Street, R. (1984). The conception and perception of noncontent speech performance: Implications for speech-accommodation theory. *International Journal of the Sociology of Language*, *46*, 97-114.

Quezada, C., Robledo, J., Román, D., & Cornejo, C. (2012). Empatía y convergencia del tono fundamental [Empathy and pitch convergence]. *Revista de Lingüística Teórica y Aplicada*, *50*(2), 145-165. doi:10.4067/s0718-48832012000200007

Quilis, A. (1981). *Fonética acústica de la lengua española* [Acoustic phonetics of Spanish language]. Madrid: Gredos.

Quilis, A. (1993). *Tratado de fonología y fonética españolas* [Treatise on Spanish phonology and phonetics]. Madrid: Gredos.

Rahimi, Z., Litman, D., & Paletz, S. (2019). Acoustic-prosodic entrainment in multi-party spoken dialogues: Does simple averaging extend existing pair measures properly?. In M. Eskenazi, L. Devillers & J. Mariani (Eds.), *Advanced Social Interaction with Agents*. *Lecture Notes in Electrical Engineering* (pp. 169-177). Cham: Springer.

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, *105*(1), 512-521. doi:10.1121/1.424522

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265-292. doi:10.1016/s0010-0277(99)00058-x

Ramus, F., Dupoux, E., & Mehler, J. (2003). *The psychological reality of rhythm classes: Perceptual studies.* Paper presented at the 15th International Congress of Phonetic Sciences, Barcelona, Spain.

Rao, G., & Smiljanic, R. (2011). Effects of language, speaking style and age on prosodic rhythm. In *Proceedings of the XVIIth international Congress of Phonetic Sciences*, Hong Kong, CN, 1662-1665.

Rao, G., Smiljanic, R., & Diehl, R. (2013). Individual variability in phonetic convergence of vowels and rhythm. In *Proceedings of Meetings on Acoustics ICA 2013*, Montreal, CA, 1-3.

Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, *80*(1), 4-9. doi:10.1016/j.biopsycho.2008.05.002

R Core Team. (2017). *R: A language and environment for statistical computing* [Computer program]. Retrieved from http://www.R-project.org/

Real Academia Española. (2008). *Corpus de referencia del español actual (CREA)* [Reference corpus of current Spanish (CREA)]. Available online at http://corpus.rae.es/creanet.html

Real Academia Española. (2014). *Diccionario de la lengua española* [Dictionary of the Spanish Language] (23rd ed.). Madrid: Espasa Libros.

Reitter, D., Keller, F., & Moore, J. (2006). Computational modeling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, New York, US, 121-124.

Reitter, D., & Moore, J. (2007). Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, CZ, 808-815.

Rizzolatti, G. (1998). What happened to Homo habilis? (Language and mirror neurons). *Behavioral and Brain Sciences*, *21*(4), 527-528. doi:10.1017/s0140525x98431261

Roach, P. (1982). On the distinction between "stress-timed" and "syllable-timed" languages. In D. Crystal (Ed.), *Linguistic Controversies* (pp. 73-79). London: HodderArnold.

Robb, M., Maclagan, M., & Chen, Y. (2004). Speaking rates of American and New Zealand varieties of English. *Clinical Linguistics & Phonetics*, *18*(1), 1-15. doi:10.1080/0269920031000105336

Roca, I. (1986). Secondary stress and metrical rhythm. *Phonology Yearbook*, *3*, 341-370.

Rolke, B., & Hofmann, P. (2007). Temporal uncertainty degrades perceptual processing. *Psychonomic Bulletin & Review*, *14*(3), 522-526. doi:10.3758/bf03194101

Ruch, H., Zürcher, Y., & Burkart, J. (2017, November). The function and mechanism of vocal accommodation in humans and other primates. *Biological Reviews*, 1-18. doi:10.1111/brv.12382

Sanchez, K., Miller, R., & Rosenblum, L. (2010). Visual influences on alignment to voice onset time. *Journal of Speech, Language, and Hearing Research*, *53*, 262-272. doi:10.1044/1092-4388(2009/08-0247)

Sancier, M., & Fowler, C. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics, 25*(4), 421-436. doi:10.1006/jpho.1997.0051

Scharf, G., Hertrich, I., Roca, I., & Dogil, G. (1995). Articulatory correlates of secondary stress in Polish and Spanish. In *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, SE, 634-637.

Schoot, L., Heyselaar, E., Hagoort, P., Segaert, K. (2016). Does syntactic alignment effectively influence how speakers are perceived by their conversation partner? *PLoS ONE*, *11*(4), 1-22. doi:10.1371/journal.pone.0153521

Schultz, B., O'Brien, I., Phillips, N., McFarland, D., Titone, D., & Palmer, C. (2015). Speech rates converge in scripted turn-taking conversations. *Applied Psycholinguistics*, *37*(5), 1-20. doi:10.1017/s0142716415000545

Schweitzer, A., & Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *Proceedings of Interspeech 2013*, Lyon, FR, 525-529.

Searle, J. (1972). Chomsky's revolution in linguistics. *The New York Review of Books*, *18*(12). Retrieved from http://www.nybooks.com/articles/10142

Shih, S., Grafmiller, J. Futrell, R., & Bresnan, J. (2015). Rhythm's role in genitive construction choice in spoken English. In R. Vogel & R. van de Vijver (Eds.), *Rhythm in Cognition and Grammar: A Germanic Perspective* (pp. 401-452). Berlin: De Gruyter Mouton.

Shockley, K. (2005). Cross recurrence quantification of interpersonal postural activity. In M. Riley and G. Van Orden (Eds.), *Tutorials in Contemporary Nonlinear Methods for the Behavioral Science*s (142-177). Retrieved from http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp

Shockley, K., Sabadini, L., & Fowler, C. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422-429. doi:10.3758/bf03194890

Shockley, K., Santana, M., & Fowler, C. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 326-332. doi:10.1037/0096-1523.29.2.326

Simpson, A. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass, 3*(2), 621-640. doi:10.1111/j.1749-818x.2009.00125.x

Smith, C. (2007). Prosodic accommodation by French speakers to a non-native interlocutor. In *Proceedings of the XVIth International Congress of Phonetic Sciences*, Saarbrücken, DE, 1081-1084.

Snidecor, J. (1943). A comparative study of the pitch and duration characteristics of impromptu speaking and oral reading. *Speech Monographs*, *10*(1), 50-56. doi:10.1080/03637754309390077

Solé, M. (1984). Experimentos sobre la percepción del acento [Experiments on stress perception]. *Estudios de Fonética Experimental*, *1*, 131-242.

Spang, K. (1983) *Ritmo y versificación. Teoría y práctica del análisis métrico y rítmico* [Rhythm and versification. Theory and practice of metrical and rhythmic analysis]. Murcia: Universidad de Murcia.

Späth, M., Aichert, I., Ceballos, A., Wagner, E., Miller, N., & Ziegler, W. (2016). Entraining with another person's speech rhythm: Evidence from healthy speakers and individuals with Parkinson's disease. *Clinical Linguistics and Phonetics*, *30*(1), 68-85. doi:10.3109/02699206.2015.1115129

Stockwell, R., Bowen, J., & Silva, I. (1956). Spanish juncture and intonation. *Language*, *32*(4), 641-665.

Street, R. (1984). Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research, 11*(2), 139-169.

Thomason, J., Nguyen, H., & Litman, D. (2013). Prosodic entrainment and tutoring dialogue success. In H. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 750-753). Heidelberg: Springer. doi:10.1007/978-3-642-39112-5_104

Thomson, R., Murachver, T., & Green, J. (2001). Where is the gender in gendered language?. *Psychological Science, 12*(2), 171-175.

Toledo, G. (1988). *El ritmo en el español. Estudio fonético con base computacional* [The rhythm of Spanish. A phonetic study with computational basis]. Madrid: Gredos.

Toledo, G. (1989). Alternancia y ritmo en el español [Alternation and rhythm in Spanish]. *Estudios Filológicos*, *24*, 19-30.

Toledo, G. (1994). Compresión rítmica en el español caribeño: Habla espontánea [Rhythmic compression in Caribbean Spanish: Spontaneous speech]. *Estudios de Fonética Experimental*, *6*, 187-217.

Toledo, G. (2009). Métricas rítmicas en tres dialectos Amper-Hispanoamérica [Rhythmic metrics in three Amper-Hispanoamérica dialects]. *Ianua. Revista Philologica Romanica*, *9*, 1-21.

Toledo, G. (2010a). Métricas rítmicas en microdiscursos [Rhythmic metrics in microdiscourses]. *Onomázein*, *1*(21), 71-95.

Toledo, G. (2010b). Métricas rítmicas en discursos peninsulares [Rhythmic metrics in peninsular discourses]. *Boletín de Lingüística*, *22*(33), 88-113.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*, 1-17. doi:10.1017/s0140525x05000129

Toro, J., Rodríguez, A., & Sebastián-Gallés, N. (2007). Stress placement and word segmentation by Spanish speakers. *Psicológica: International Journal of Methodology and Experimental Psychology*, *28*, 167-176.

Traunmüller, H., & Eriksson, A. (1994). *The frequency range of the voice fundamental in the speech of male and female adults* (Technical Report). University of Stockholm. Retrieved from: https://pdfs.semanticscholar.org/aa8b/acb5e7843740fbea24742c3046fbcc009a49.pdf

Trehub, S. (1990). The perception of musical patterns by the human infant. In I. Berkeley & W. Stebbins (Eds.), *Comparative Perception, Vol. 1: Mechanisms* (pp. 429-459). New York: Wiley.

Trevarthen, C. (1998). The concept and foundations of infant intersubjectivity. In S. Braten (Ed.), *Intersubjective Communication and Emotion in Early Ontogeny* (pp. 15-46). Cambridge: Cambridge University Press.

Ulloa, M. (2011). ¿Hablan a golpes?: El enlace [Do they speak in beats? The fusion]. *Foro de Profesores de E/LE*, *7*, 1-10.

Urrutia, H. (2007). La naturaleza del acento en español: Nuevos datos y perspectivas [The nature of stress in Spanish: New data and prospects]. *RLA. Revista de Lingüística Teórica y Aplicada, 45*(2), 135-142. doi:10.4067/s0718-48832007000200010

Van Puyvelde, M., Loots, G., Gillisjans, L., Pattyn, N., & Quintana, C. (2015). A cross-cultural comparison of tonal synchrony and pitch imitation in the vocal dialogs of Belgian Flemish-speaking and Mexican Spanish-speaking mother–infant dyads. *Infant Behavior and Development*, *40*, 41-53. doi:10.1016/j.infbeh.2015.03.001

Van Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., & Stokes, M. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, *84*(3), 917-928.

Vaughan, B. (2011). Prosodic synchrony in co-operative task-based dialogues: A measure of agreement and disagreement. In *Proceedings of Interspeech 2011*, Florence, IT, 1865-1868.

Venables, W., & Ripley, B. (2002). *Modern applied statistics with S*. New York: Springer.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209-232. doi:10.1016/j.specom.2013.09.008

Wallis, E. (1951). Intonational stress patterns of contemporary Spanish. *Hispania, 34*(2), 143-147.

Wang, M., Kong, L., Zhang, C., Wu, X., & Li, L. (2018). Speaking rhythmically improves speech recognition under "cocktail-party" conditions*. The Journal of the Acoustical Society of America, 143(*4), EL255-EL259. doi:10.1121/1.5030518

Wang, Y., Yen, J., & Reitter, D. (2015). Pragmatic alignment on social support type in health forum conversations. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, Denver, US, 9-18.

Ward, A., & Litman, D. (2007). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education 2007*, Farmington, US, 57-60.

Weirich, M., & Simpson, A. (2014). Differences in acoustic vowel space and the perception of speech tempo. *Journal of Phonetics, 43*, 1-10. doi:10.1016/j.wocn.2014.01.001

Weise, A., & Levitan, R. (2018). Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of NAACL-HLT 2018,* New Orleans, US, 297-302.

White, L., Mattys, S., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language 66*(4), 665-679. doi:10.1016/j.jml.2011.12.010

Winter, B., & Grawunder, S. (2012). The phonetic profile of Korean formal and informal speech registers. *Journal of Phonetics*, *40*(6), 808-815. doi:10.1016/j.wocn.2012.08.006

Wretling, P., & Eriksson, A. (1998). Is articulatory timing speaker specific? – Evidence from imitated voices. In *Proceedings of Fonetik 98*, Stockholm, SE, 48-51.

Wynn, C., Borrie, S., & Sellers, T. (2018). Speech rate entrainment in children and adults with and without autism spectrum disorder. *American Journal of Speech-Language Pathology, 27,* 1-10. doi:10.1044/2018_ajslp-17-0134

Xu, Y., & Reitter, D. (2016). Convergence of syntactic complexity in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, DE, 1-6.

Yu, A., Abrego, C., & Sonderegger, M. (2013). Phonetic Imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PLoS ONE, 8*(9), 1-13. doi:10.1371/journal.pone.0074746

Yu, L., Hattori, Y., Yamamoto, S., & Tomonaga, M. (2018). Understanding empathy from interactional synchrony in humans and non-human primates. In L. Di Paolo, F. Di Vincenzo & F. De Petrillo (Eds.), *Evolution of Primate Social Cognition. Interdisciplinary Evolution Research* (pp. 47-58). Cham: Springer.

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. In *Proceedings of Interspeech 2006,* Pittsburgh, US, 1-4.

Zhang, Z. (2016). Residuals and regression diagnostics: Focusing on logistic regression. *Annals of Translational Medicine*, *4*(10):195, 1-8. doi:10.21037/atm.2016.03.36

# Appendix 1: Complete list of sentences used in Experiment 1

*Regular groups*

| | |
|---|---|
| El padre te quiere con ganas | The father likes you a lot |
| El carro los trae de vuelta | The car brings them back |
| La lluvia se siente de cobre | The rain feels like cooper |
| La madre los llama sin cargo | The mother calls them for free |
| Las noches les pasan de largo | The nights passed them by |
| Las niñas les muestran las caras | The girls show their faces to him |
| Los casos le salen por miles | The cases are numerous for him |
| Los malos lo sufren sin rabia | Bad people suffer it without rage |
| Los niños nos tienen tu gato | The kids keep your cat for us |
| Mi prima le saca las fotos | My cousin takes his pictures |
| Mi novia se lleva tu nota | My girlfriend takes your note away |
| Su perro te llena de barro | His dog gets you muddy |
| Su casa se cae por partes | His house falls apart |
| Tu lora lo dijo con gracia | Your parrot said it gracefully |
| Tu primo nos paga las cuentas | Your cousin pays us the bills |
| Tus gastos nos dejan sin plata | Your expenses leave us without money |

*Irregular groups*

| | |
|---|---|
| Alba me le subió de precio | Alba raised the price for me |
| Ana me les cambió la vida | Ann changed their lives for me |
| Ángel te lo pasó de lado | Angel passed it edgewise to you |
| Carlos te la mostró con fotos | Charles showed it to you with photos |
| Carmen te las perdió por miedo | Carmen lost them for fear |
| Clara te los jugó sin ganas | Clara unwillingly played them to you |
| Ella nos los negó por plata | She denied them to us for money |
| Ellas me los tendrán si quieren | They will keep them for me if they want |
| Laura nos lo pidió sin gracia | Lora gracelessly asked us for it |
| Mario se nos quedó sin casa | Mario was left homeless |
| Marta se las dará si paga | Martha will give them to her if she pays |
| Pablo se le salió con rabia | Paul got away from him in anger |
| Paco nos la sacó del barro | Paco took it out of the mud for us |
| Pedro nos las verá si puede | Peter will see them for us if he can |
| Sara se les llevó la lora | Sarah took the parrot away from them |
| Sergio me lo creyó del todo | Serge totally believed me |

*Regular feet*

| | |
|---|---|
| Él me la trae sin crédito | He brings it to me without credit |
| Él te la cambia por máquinas | He trades it with you for machines |
| Flor se te sale del médico | Flor gets away from the doctor's office |
| John me le compra la fórmula | John buys the formula for me |
| Juan se la sabe sin cálculos | John knows it without calculations |
| Luis me los paga con dólares | Lewis pays them to me with dollars |
| Luz nos los tiene sin pérdidas | Luz keeps them for us without losses |
| Mar me lo cuenta los miércoles | Mar tells it to me on Wednesdays |
| Juan te los llama con música | John calls them for you with music |
| Paz nos la debe del sábado | Paz owes us for Saturday |
| Paul se les ríe tras cámaras | Paul laughs at them behind the cameras |
| Rey nos las guarda con código | Ray keeps them for us with a code |
| Sol te lo lleva del público | Sol brings it to you from the public |
| Tú nos lo cuentas con método | You tell it to us with a method |
| Yo se lo digo sin lágrimas | I tell it to him without tears |
| Yo te las cambio por pájaros | I trade them with you for birds |

*Irregular feet*

| | |
|---|---|
| Él me siente de su público | He considers me a part of his audience |
| Él se ríe de mi síndrome | He laughs at my syndrome |
| Flor le debe por los músicos | Flor owes him for the musicians |
| John le paga sin tu fórmula | John pays him without your formula |
| Juan lo quiere sin los títulos | John wants it without the titles |
| Luis la lleva con sus líderes | Lewis brings her with his leaders |
| Luz lo guarda tras la cámara | Luz keeps it behind the camera |
| Mar nos cuenta de tus pérdidas | Mar tells us about your losses |
| Juan nos dijo de la práctica | John told us about the practice |
| Paz lo tiene con su plástico | Paz keeps it with its plastic |
| Paul me nota por mis músculos | Paul notices me for my muscles |
| Rey los compra con tus dólares | Ray buys them with your dollars |
| Sol se cae tras las máquinas | Sol falls down behind the machines |
| Tú los llamas por sus términos | You refer to them by their terms |
| Yo las cuento sin la técnica | I count them without the technique |
| Yo te veo tras mis lágrimas | I see you behind my tears |