The role of intonation and facial gestures in conveying interrogativity

Joan Manel Borràs Comes

TESI DOCTORAL UPF / 2012

DIRECTORA DE LA TESI

Dra. Pilar Prieto i Vives

DEPARTAMENT DE TRADUCCIÓ I CIÈNCIES DEL LLENGUATGE

UNIVERSITAT POMPEU FABRA

Als meus

I a prop la mar, la mar que tant estimo.
Aquí he viscut i això conec.
GERARD VERGÉS

Acknowledgments

First and foremost, I would like to thank the director of this thesis, Dr. Pilar Prieto, for her trust and support and the long hours she has devoted to guiding my progress as a researcher. In her I have seen scientific research at its best. And thanks to her, I was accepted, while still a complete novice, into a team of international prominence and well-deserved influence in this field. It is my experience in that context, under the leadership of Dr. Prieto, that has brought me to this moment of great personal achievement and satisfaction. For this I am profoundly grateful.

I give thanks also to the funding institutions and research projects that have made this work possible, in particular the Ministry of Education and Science project “Bilingualism and Cognitive Neuroscience” (CONSOLIDER-INGENIO 2010 Programme CSD2007-00012), under the direction of Núria Sebastián-Gallés and Itziar Laka, who very generously deposited their full trust in my abilities. I am also indebted to the projects “Estructura prosòdica i adquisició de la prosòdia en català i espanyol” (HUM2006-01758) and “The role of tonal scaling and tonal alignment in distinguishing intonational categories in Catalan and Spanish” (FFI2009-07648/FILO), funded respectively by the Ministry of Science and Innovation and the Ministry of Education and Science. I also owe a great deal to the *Grup de Recerca Emergent* project “Grup d’Estudis de Prosòdia” (2009 SGR-701), funded by the Generalitat de Catalunya through the *Pla de Recerca de Catalunya*. Finally, to I am grateful to the Catalan government’s *Agència de Gestió d’Ajuts Universitaris i de Recerca* for having endowed me with a Fellowship for Research Abroad (BE-DGR 2010, 2010 BE1 00207), which allows me to obtain official European Union accreditation for this thesis.

I give my thanks to Dr. Carles Escera and Dr. Jordi Costa-Faidella at the Universitat de Barcelona, without whose expertise and dedication the electrophysiological study presented in this thesis — and much less its successful publication — would not have been possible. Thanks also to Dr. Marc Swerts and to the soon-to-be Dr. Constantijn Kaland, for their kind welcome at the University of Tilburg and their indispensable assistance in the development of the study that concludes this thesis.

Thanks to the Universitat Pompeu Fabra, and also to the Universitat Autònoma de Barcelona, for the use of their resources. In particular, I would like to thank the Departament de Traducció i Ciències del Llenguatge of the Universitat Pompeu Fabra for its academic goodwill and administrative effectiveness.

Thanks to all the teachers that I have had, each of whom gave me their own special grain of courage in the course of my academic training. Thank you, Dr. Miquel Àngel Pradilla and Dr. Josep Maria Pujol, for the enthusiasm and knowledge that I have received. And thanks to those all who goodheartedly agreed to participate in the several experiments that I have carried out.

Thanks to the various members of academia who as conference participants or reviewers for scientific journals have shown me how to fine-tune my research (L. Aguilar, A. Chen, C. De-la-Mota, M. D'Imperio, L. Dilley, G. Elordieta, E. Estebas, S. Frota, J. Hualde, E. Krahmer, I. Mascaró, L. McNally, B. Post, J. Quer, Y. Rose, S. Soto, M. Vigário, etc.), and most especially the people associated with the "Grup d'Estudis de Prosòdia".

Thanks, *grepets* (M. Armstrong, R. Craviotto, V. Crespo, N. Esteve, S. González, P. Roseano, R. Sichel, M. Vanrell, etc.), for the moments, conversations, help and strength that you have given me.

Infinite thanks to my family, my parents and my sister Carme, who have seen precious little of me but who never for an instant let me feel alone. I hope my friends will also forgive me the neglect they have endured. And last but for from least, thank you, Mon, for all you have given me in the six years that we have been together, despite all the time stolen from you by these pages.

Abstract

This thesis investigates the role that different aspects of audiovisual prosody play in the production and perception of interrogativity. To this end, two types of statements and two types of questions are analyzed: information and contrastive focus statements (IFS, CFS), and information-seeking and counter-expectational questions (ISQ, CEQ). A multimodal approach is thus followed for the study of interrogativity, by means of a variety of production and perception experiments, from games specifically designed to elicit spontaneous productions of specific discourse categories to the analysis of event-related potentials. The first study reveals that pitch range differences are the main intonational cue used by Central Catalan speakers in order to distinguish between IFS and CEQ. The second study shows that such intonational contrasts are encoded automatically in the auditory cortex. Both studies strengthen the argument that pitch range features need to be represented descriptively at the phonological level. The third study shows that facial gestures are the most influential elements that Catalan listeners rely on to decide between CFS and CEQ interpretations, though bimodal integration with acoustic cues is necessary in order for perceptual processing to be accurate and fast. The fourth study reveals that Catalan and Dutch speakers mainly rely on language-specific auditory differences in order to detect IFS and ISQ, but also that the presence of gaze increases the identification of an utterance as a question. Finally, this study demonstrates that a concentration of several response-mobilizing cues in a sentence is positively correlated with the perceivers' ratings of these utterances as interrogatives.

Resum

Aquesta tesi investiga el rol que exercixen diversos aspectes de la prosòdia audiovisual en la producció i la percepció de la interrogativitat. A tal efecte, s'analitzen dos tipus d'oracions declaratives (de focus informatiu i de focus contrastiu; IFS i CFS) i dos tipus d'oracions interrogatives (de cerca d'informació i d'antiexpectació; ISQ i CEQ). Així, la tesi estudia la interrogativitat des d'una perspectiva multimodal, amb diferents experiments de producció i de percepció que van des de jocs especialment dissenyats per elicitar produccions espontànies de determinades categories discursives fins a l'anàlisi de potencials evocats cerebrals. El primer estudi revela que els parlants de català central empren principalment el camp tonal per distingir entre IFS i CEQ. El segon, que el còrtex auditiu codifica automàticament tal contrast entonatiu. Ambdós estudis conclouen que cal explicitar les propietats del camp tonal quan es descriu fonològicament l'entonació de la llengua. El tercer estudi mostra la major influència dels gestos facials a l'hora de distingir CFS i CEQ en català, així com la necessitat d'integrar perceptivament les variables visuals i les acústiques perquè la idenficació siga acurada i ràpida. El quart estudi revela com els parlants de català i de neerlandès es basen principalment en les diferències auditives de les seues respectives llengües a l'hora de distingir IFS i ISQ, però també com el fet que el parlant mire el seu interlocutor incrementa la interpretació interrogativa d'una oració. Finalment, l'estudi demostra que la presència de diversos indicis mobilitzadors de resposta en una oració està positivament correlacionada amb les interpretacions interrogatives que els oients en fan.

List of original publications

CHAPTER 2

Borràs-Comes, J., Vanrell, M. M., & Prieto, P. (*accepted pending minor revisions*). The role of pitch range in establishing intonational contrasts. *Journal of the International Phonetics Association*.

CHAPTER 3

Borràs-Comes, J., Costa-Faidella, J., Prieto, P., and Escera, C. (2012). Specific neural traces for intonational discourse categories as revealed by human-evoked potentials. *Journal of Cognitive Neuroscience*, 24(4), pp. 843-853.

CHAPTER 4

Borràs-Comes, J., & Prieto, P. (2011). 'Seeing tunes'. The role of visual gestures in tune interpretation. *Journal of Laboratory Phonology*, 2(2), pp. 355-380.

CHAPTER 5

Borràs-Comes, J., Kaland, C., Prieto, P., & Swerts, M. (*submitted*). Audiovisual correlates of interrogativity: a crosslinguistic study. *Journal of Nonverbal Behavior*.

Table of contents

List of tables

List of figures

List of abbreviations

AEP	Auditory Evoked Potentials
AM	Autosegmental-Metrical
AO	Auditory-Only
AU	Action Unit
AV	Audiovisual
CEQ	Counter-expectational Question
CFS	Contrastive Focus Statement
CI	Categorization Index
DEV	Deviant
EEG	Electroencephalography
ERP	Event-Related Potentials
FACS	Facial Action Coding System
GLMM	Generalized Linear Mixed Model
IFS	Information Focus Statement
IP	Intonational Phrase
ISQ	Information-Seeking Question
MMN	Mismatch Negativity
PSOLA	Pitch Synchronous Overlap and Add
RT	Reaction Times
SOA	Stimulus-Onset Asynchrony
STD	Standard
ToBI	Tones and Break Indices
VO	Visual-Only

CHAPTER 1
Introduction

The main aim of this thesis is to deepen our knowledge about interrogativity, specifically about how speakers mark it and, especially, how they detect it. This is to say, it seeks to determine the elements that allow us to differentiate an interrogative sentence from a declarative sentence both in speech production and in speech perception. The motivation behind this thesis is thus the desire to better understand one of the core aspects of human communication, namely the mechanism by which we comprehend whether information is being given or is being asked for.

It is well known that different intonation contours serve as interrogative markers in a number of languages. However, though one of the main functions of intonation is to convey the pragmatic meaning of a sentence, many intonation studies have described the intonational phonology of a language without taking explicitly into account those pragmatic contexts. In this regard, previous intonation studies are based on read speech and also tend to ignore other linguistic correlates, like gestures, which accompany intonation patterns in normal face-to-face communication. This thesis deals with two types of statements and two types of *yes-no* questions, which can be classified as neutral (i.e., nonbiased) and biased depending on the way in which they convey their semantic content.

In the case of statements, we distinguish between *information focus statement*s (IFS) and *contrastive focus statements* (CFS). By IFS, we refer to a neutral statement, i.e., a statement which carries new information in which there is a particular constituent that is focalized with respect to the background. On the other hand, a CFS refers to the marking of a constituent as "a direct rejection of an alternative" (Gussenhoven 2007). A CFS typically corrects "the value of the alternative assigning a different value" (Cruschina 2011). Therefore, the main difference between the two focus types is that while a CFS is dependent on a preceding assertion, which is denied/corrected by the new focalized item, an IFS is not. This denial or correction is often made explicit in the intonation and gestural planes of most intonational languages.

In the case of questions, we distinguish between *information-seeking questions* (ISQ) and *counter-expectational questions* (CEQ). By ISQ, we refer to the sort of question specifically designed to gather information from a recipient, with no special intuitions required on its response on the part of the respondent. On the other hand, CEQs are related to echo questions. Echo questions are those in which the listener repeats information that s/he has just heard, generally either because s/he has not properly heard or understood what was said or because the implications of that information are in conflict with his/her previous expectations. CEQs represent the latter type, and they are sometimes characterized by a nuance of surprise or incredulity. As Cohen (2007: 133) states, "an incredulity question expresses the claim that in none of the speaker's belief (or normative) worlds is the echoed statement true — hence the incredulity (or indignation) expressed toward that statement" (see Cohen 2007 on the further distinction between *echo* and *incredulity questions*). As in the case of CFS, the nuance of unexpectedness, surprise or incredulity of a CEQ is often marked by intonation and specific gesture patterns in many intonational languages.

In order to analyze intonational patterns, we use the Tone and Break Indices (ToBI) transcription system, which based on the Autosegmental and Metrical (AM) theory of intonation. Briefly, this approach describes the intonation of a sentence by distinguishing those tones associated with stressed syllables (*pitch accents*) from those aligned at the right edge of a phrase (*boundary tones* and *phrase accents*). The two basic units that make up pitch accents and edge tones are H[igh] and L[ow] tones, respectively interpreted as an increase or decrease of pitch within an utterance's tune. In most languages, pitch accents are generally composed of one or two tones, the most prominent of which is marked with an asterisk (T*). Edge tones are generally perceived as falling or rising melodic patterns, or a combination thereof, and are generally transcribed as a percentage symbol (T%) or dash (T–). Because it is a phonological transcription system, ToBI requires expert human knowledge for the characterization of the prosodic events specific to each language, and many language-specific ToBI transcription systems have been developed since the appearance of Pierrehumbert's (1980) dissertation on the English intonational system (see Ohio State University Department of Linguistics 1999).

This thesis is organized in four main studies, which are presented in Chapters 2 to 5. First, I analyze the role that a specific intonational feature plays in the distinction between statements and counter-expectational questions in Catalan. This intonational feature is pitch range, namely the distance or span between the lowest and the highest f0 values observed in utterance pitch accent (i.e., a valley and a peak; see Gussenhoven 2004). The reason behind choosing Catalan as a test language is that in this language, as in some other Romance languages, a rising-falling nuclear pitch contour — i.e., a rising pitch accent associated with the utterance-final stressed syllable followed by a low boundary tone — may be used to convey either IFS, CFS, or CEQ, depending on the utterance's pitch range properties. This intonation-based contrast will be analyzed in Chapters 2 and 3. Given that these contrasts can also be cued by means of specific facial gestures, the interaction between auditory and gestural cues in the perception of statements and questions will be analyzed in Chapter 4. As the experiments discussed up to Chapter 4 compare statements with a biased type of question, Chapter 5 analyzes how neutral questions (ISQ) are detected when compared with neutral statements (IFS).

Table 1 shows a summary of the types of declaratives and interrogatives that are analyzed in this thesis.

Table 1. Sentence meanings analyzed in this thesis.

statements	neutral	Information Focus Statement (IFS)
	biased	Contrastive Focus Statement (CFS)
questions	neutral	Information-Seeking Question (ISQ)
	biased	Counter-Expectational Question (CEQ)

The aim of the first study (**Chapter 2**) is to investigate how IFS, CFS, and CEQ are distributed across the pitch range continuum and whether Catalan listeners use these pitch range distinctions to identify such meanings. It is well known that different intonation contours serve as interrogative markers in a number of languages, but whether pitch accent range differences are used by languages to express such a discrete linguistic distinction is still an unresolved issue in the field of intonational phonology. To this end, we performed two tasks especially appropriate for this purpose. First, we used an identification task with three possible response options, thus allowing for the simultaneous comparison of the three categories (IFS, CFS, and CEQ). Second, we used a congruity task, which makes it possible to investigate the degree to which listeners are aware of the semantic appropriateness of a particular intonation contour to a given discourse context and whether they are able to detect an incongruous use of this contour. In the two tasks, the identification responses are complemented with the analysis of the reaction time measures, as these measures have been found to be useful to investigate the discreteness of different intonational contours. Whereas the perceived difference between the two types of statements cannot be exclusively explained by pitch range differences, the results of the first study show a clear contrast between IFS and CEQ.

Given the results in Chapter 2, **Chapter 3** tests for the perception of this contrast using an electrophysiological brain exploration. A series of studies have indicated that segmental and tonal phonological distinctions can be represented in pre-attentive auditory sensory memory using the auditory mismatch negativity (MMN) event-related brain potential (ERP). In this study we tested whether within-category and across-category intonational contrasts between IFS and CEQ in an intonation language will also elicit distinct neurophysiological patterns of activity, which would then support a distinct neurophysiological pattern for IFS and CEQ and the automatic encoding of intonational contrasts in the auditory cortex. Moreover, this finding would represent evidence that the processing of intonational contrasts by the human brain is done in a similar fashion to that of segmental contrasts.

As statements and questions are produced in normal face-to-face communication, they are associated with certain specific facial gestures, such as head and eyebrow movements. In our third study (**Chapter 4**) we analyze another unresolved question in the field of audiovisual prosody, namely how acoustic and visual cues interact in the native perception of such a pragmatic difference. Though the majority of studies on audiovisual prosody have found a complementary mode of processing whereby sight provides relatively weak and redundant information in comparison with strong auditory cues, other work has found that sight provides information more efficiently than hearing. In this chapter we take into account the roles of both pitch range and facial gestures in the distinction between CFS and CEQ. After we had synthesized the auditory and gestural signals that are characteristic of these particular pragmatic meanings using recordings and a digital image-morphing technique, subjects participated in two multimodal identification tasks in which they were presented with congruent and incongruent combinations of such audiovisual cues in order to analyze their perceived degree of interrogativity.

In our last study (**Chapter 5**) we further analyze the audiovisual perception of interrogativity, but this time confronting the contrast between information focus statements (IFS) and information-seeking questions (ISQ), which each represent the most neutral types of the two pragmatic meanings. We used a natural setting in order to elicit a series of statements and questions. Then, on the basis of these elicited materials, we had subjects participate in unimodal and multimodal identification tasks (this time using only congruent audiovisual combinations). This methodology allowed us to investigate the core mechanisms involved in conveying interrogativity in both speech production and perception. This investigation compared the respective strategies used by Catalan and Dutch speakers. While both languages are known to use intonation for marking interrogativity, Dutch also exploits syntactic inversion for this purpose, which is the reason for comparing Dutch and Catalan in this study. This task would allow us to assess whether participants speakers of the respective languages differentiate neutral statements from questions unimodally and/or multimodally. It would also tell us which auditory and gestural features — i.e., syntactic inversion when available, rising intonation contours, gaze, eyebrow raising — were most frequently used in production and perception, and whether these strategies interacted in the participants' identification of an utterance as a question.

One feature of our methodology that should be highlighted is our multimodal approach to the study of interrogativity. Most traditional studies have neglected the nonverbal component of the declarative / interrogative distinction and have mainly focused on its syntactic, morphological, and intonational marking. There is also thus far only limited research that takes into account more than one strategy at a time and explains their potential interaction as response-mobilizing features (see Stivers & Rossano 2010).

The second feature that we regard as contributing particular value is the variety of methodologies that were applied in the several experiments analyzed in this thesis with the aim of improving the 'ecological validity' of our results. In our production experiments, for example, we collected data through both Discourse Completion Tests, broadly used in pragmatics research (Kasper & Dahl 1991, Cohen 1996, Billmyer & Varghese 2000, Golato 2006, Nurani 2009) and games, like the version of

Guess Who, specifically adapted to elicit spontaneous productions of specific discourse categories (Ahmad et al. 2011). As for perception experiments, we used different behavioral approaches, like congruency and identification tests (unimodal or multimodal, binomial or multinomial), from which we have analyzed both responses and reaction times, and an electrophysiological exploration using event-related potentials with the use of a mismatch paradigm (Näätänen 2001).

CHAPTER 2

The role of pitch range in establishing discourse categories

2.1. Introduction

As is well known, intonational languages use pitch variation to express differences in pragmatic and discourse meanings. Though early approaches distinguished among four (Trager & Smith 1951) or three level tones (Stockwell et al. 1956), the Autosegmental-Metrical (AM) model takes as a central assumption that only two tones, Low and High, are necessary to distinguish pitch accent and boundary tone categories in English. This means that all remaining pitch range variation exclusively expresses differences in emphasis or prominence (Pierrehumbert 1980, Beckman & Pierrehumbert 1986, Bolinger 1986, Dilley 2010, and others). This assumption relies on a version of the so-called Free Gradient Hypothesis (Ladd 1994, Ladd 1996, Gussenhoven 1999), which holds that one of the most common effects of gradually expanding the pitch range of a given pitch accent is the pragmatic reinforcement of the utterance (namely an increase in the degree of the speaker's involvement in the speech act). In line with this, Liberman and Pierrehumbert (1984) demonstrated in their study of English pitch range that a gradual increase in emphasis was correlated with an increase in pitch range of the pitch accent.

Notwithstanding, work on English and other languages has revealed that pitch range variation can express categorical differences in meaning even within the AM framework (Ward & Hirschberg 1985, Hirschberg & Ward 1992, Ladd 1994, Ladd 1996, Ladd & Morton 1997, Chen 2003, Braun 2006, Vanrell 2006, Savino & Grice 2011, Vanrell 2011). It is generally accepted that tones in tonal languages behave as phonemic units and. In the last decades, work within the intonational phonology field has shown that intonational contrasts apply to intonational languages, the latter conveying “meanings that apply to phrases or utterances as a whole, such as sentence type or speech act, or focus and information structure” (Ladd 1996: 7). For example, Ladd and Morton (1997) investigated the contrast between normal vs. emphatic rising pitch accents in English. Though an abrupt shift in identification from normal to emphatic interpretations was found as pitch range increased, little evidence was provided of an associated peak in discriminability between stimulus pairs. Chen’s (2003) replication of the experiment claimed that taking the identification results together with an analysis of reaction time (RT) data revealed that the perceived distinction between a normal high accent and an emphatic high accent is of a discrete nature. Hirschberg and Ward (1992) showed that a larger pitch range of the English rise-fall-rise tune can change the interpretation of an utterance from one of uncertainty to one of incredulity. Finally, Calhoun (2004) found that themes and rhemes are marked by distinctive pitch accents and that the most reliable cue to the theme and rheme accents is pitch height.

Some recent work on Romance languages has found that pitch range variation can also convey discrete intonational contrasts. Savino and Grice (2011) demonstrated that the pitch range of a rising pitch accent was responsible for the difference between information-seeking and counter-expectational questions in Bari Italian (where the latter are produced with an expanded pitch range). The listeners’ responses and reaction times obtained by means of a semantically motivated identification task provided clear evidence for the categorical use of pitch range variation in Bari Italian question interpretation. Similarly, by using the results of a gating experiment, Face (2005, 2007, 2011) claimed for Spanish that the height of the initial f0 peak of an utterance allows listeners to distinguish between declaratives and *yes-no* questions, thus arguing for the phonologization of pitch range. This was consistent with Prieto (2004), who found that the height of the initial f0 peak varies depending on sentence type; specifically, *yes-no* questions, *wh-* questions, exclamatives, and imperatives all have significantly higher initial f0 peaks than declaratives. Moreover, Vanrell (2011) showed for falling nuclear pitch accents (H+L* L%) that the pitch height of the high leading tone is the main cue used by Majorcan Catalan listeners to distinguish between a *wh-* question and two types of *yes-no* questions. That is, an upstepped leading high tone signals a *yes-no* question in which the speaker has no previous knowledge about the answer, whereas a non-upstepped leading tone signals that the speaker is asking a *yes-no* question about mutually shared information; in addition, a downstepped leading tone signals a *wh-* question.[1]

In general, these investigations demonstrate that pitch range variation can be perceived in a discrete fashion in some languages and thus strengthen the arguments in favor of treating pitch range differences in phonological terms in these languages. The idea of enriching the traditional High-Low dichotomy with a finer differentiation of pitch range was already advocated by researchers such as Ladd (1994:60), who pointed out that "the Bruce-Pierrehumbert approach to intonational phonology must be enriched with a notion of categorical distinctions of pitch range. We need to get rid of the idea that any distinction that is orthogonal to the basic opposition between High and Low tones is ipso facto gradient: both gradient factors and categorical ones play a role in the vertical scale of any given tone".

In this chapter, we investigate more extensively the role of pitch accent range variation in conveying intonational contrasts in Catalan. In our previous descriptive studies based on the analysis of Catalan dialectal data from the *Interactive Atlas of*

[1] Similar conclusions have been drawn when examining boundary tones. Crosslinguistic studies have reported active mid-level boundary tones contrasting with high-level tones in the phonological domain of English (Beckman & Ayers Elam 1997), Greek (Arvaniti & Baltazani 2004), German (Grice et al. 2005), Spanish (Beckman et al. 2002), Korean (Lee 2004), and Catalan (Vanrell 2011).

Catalan Intonation (Prieto & Cabré 2007-2012, see also Prieto 2002) using Cat_ToBI (Prieto et al. 2009, Prieto *in press*, Aguilar et al. 2009) we observed that the rising pitch accent of information focus statements (IFS) was produced with a narrow pitch range, while that of contrastive focus statements (CFS) and counter-expectational questions (CEQ) was produced with a wider pitch range. In these three types of utterance, the alignment properties of the tones are found to be the same, i.e., a low tone is aligned with the beginning of the accented syllable, the rising tone occurs within this accented syllable, and the peak of this rise is always aligned with the end of the accented syllable.[2] Similar observations have been made for other Romance languages such as Friulian (Roseano et al. 2011) and Castilian Spanish (Estebas-Vilaplana & Prieto 2010). Examples of linguistic contexts eliciting these three types of pragmatic meanings are shown in (1).[3]

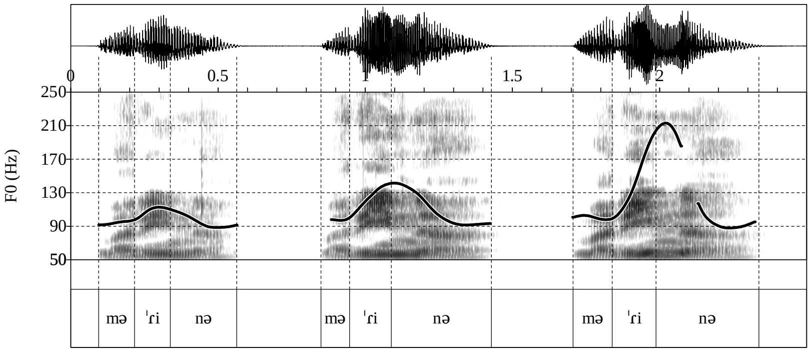
[2] The AM representation adopted for this rising accent is L+H* L%, as stated in Prieto (in press) and Prieto et al. (2009). These publications report that no differences in the peak alignment are found between the three contours (see also Prieto 2005).

[3] Even though Romance languages such as Catalan, Italian, and Spanish have been said to mark CFS through syntactic mechanisms (Vallduví 1991, Ladd 1996), this does not exclude an active role for intonation, especially in those cases in which word order remains the same (Estebas-Vilaplana 2009 for Catalan; Face & D'Imperio 2005 for Italian and Spanish). According to previous research on this issue (Solà 1990, Vallduví 1991), since prominence shift is a less-used strategy in Catalan to make the focused constituent fall under prominence, other syntactic mechanisms such as dislocation (*No les tinc, les claus*, lit. 'Not them I.have, the keys', 'I do not have the keys') or elision (*No les tinc*, lit. 'Not them I.have', 'I do not have them') of the nonfocal material of a sentence (Solà 1990, Vallduví 1991, Prieto & Cabré 2007-2011), focus fronting (*Negres, són, i no blanques*, lit. 'Blacks, they.are, and not whites', 'They are black, not white') or clefting (*És el Marçal (que/el que/qui/el qui) no suporto*, lit. 'Is the Marçal who not I.stand', 'It is Marçal who I cannot stand') (Solà 1990, Vallduví 1991) are proposed. Such sentence types are characterized by a similar intonation pattern L+H* L%, either produced in isolation or accompanied by the nonfocal material, which tends to undergo tonal compression.

(1)	a.	(IFS)	Com es diu, la seva filla?	*What's their daughter's name?*
			Marina.	*Marina.*
	b.	(CFS)	Es diu Júlia, ella, no?	*Her name's Júlia, isn't it?*
			Marina!	*[No! It's] Marina!*
	c.	(CEQ)	Li posaran Marina.	*They'll call her Marina.*
			Marina?	*Marina? [Really?]*

Figure 1 shows the waveforms and f0 contours of the proper noun *Marina* ([mə'ɾinə]) obtained as responses to the contexts in (1).

Figure 1. Waveforms and f0 contours of the proper name *Marina* produced with an IFS meaning (left), a CFS meaning (central position), and a CEQ meaning (right).



With the aim of investigating the role of pitch range in the interpretation of rising pitch accents in Catalan, we initially carried out two identification tasks with twenty native speakers of Catalan, the results of which are reported in Borràs-Comes et al. (2010). These tasks were identification tasks with binomial identification responses (two-way identification tasks), the first dealing with the contrast between IFS and CEQ and the second with the contrast between IFS and CFS. The identification results showed an S-shaped function for both comparisons, thus suggesting a discrete perception for three types of pragmatic meanings. However, an analysis of the reaction times revealed a significant reaction time peak only when IFS was compared with CEQ. As Chen (2003: 98) pointed out, "if the identification categories emerging from the response frequencies are not task-induced but linguistically real, we will expect that the within-category stimuli are comparable in terms of cognitive load and therefore will trigger similar mean RTs for identification", and vice versa. This close correlation has also been found in many other experiments (e.g., Falé & Hub Faria 2005 for European Portuguese, or Vanrell 2006 and Vanrell 2011 for Catalan). The fact that we found no peaks in RTs in the IFS vs. CFS comparison (Borràs-Comes et al. 2010) was interpreted as providing initial evidence for both a categorical effect in pitch range (i.e., the phonological difference between an IFS and a CEQ) and a gradient effect (i.e., the difference in pitch range between an IFS and a CFS).

The goal of the present chapter is to investigate more deeply the role of pitch accent range in conveying the abovementioned pragmatic meaning distinctions in Catalan (IFS, CFS, and CEQ) by using two tasks that are especially appropriate for this purpose. First, we will use an identification task allowing for the simultaneous comparison of the three categories (Experiment 1) and then we will take linguistic context explicitly into account in order to test for the congruity of each target sentence occurring in a typical linguistic context for each pragmatic meaning (Experiment 2). These experiments are complemented with the results of reaction time measures, as these measures have been found to be significantly useful to investigate the discreteness of different intonational contours. Following our initial findings showing that the comparisons between IFS/CFS and IFS/CEQ do not behave alike, we initially hypothesized that the three categories would not be distributed in three well-differentiated areas of the pitch height continuum depending on the height of the H tone, but rather in only two such areas.[4]

[4] Note that a three-way distinction in pitch height does not represent a very marked situation crosslinguistically if we consider the tonal height distinctions reported for tonal languages. For example, in some African

Another goal of the chapter is to assess the utility of these tasks for the investigation of the role of intonational differences in conveying pragmatic meaning distinctions. A triple identification task and a congruity task were thus conducted to test for the presence and hierarchy of this potential three-way distinction between rising pitch accents in Catalan. This would give us more information about the suitability of binomial identification tasks for the investigation of pragmatic meanings. In other words, we want to know if such a three-way contrast in identification will lead to similar results as a two-way contrast and whether the results of such a study can be corroborated by using a congruity task. Experiment 1 consisted of a semantically motivated identification test in which participants had to identify each of the three meanings (IFS, CFS, and CEQ) for a set of isolated stimuli, allowing for a triple response. To our knowledge, no similar triple identification tasks have been previously applied to intonation, and so this is the first study approaching the analysis of intonational contrasts that allows for more than two responses at a time. Experiment 2 consisted of a congruity test which tested participants' acceptance of each stimulus occurring within a typical communicative context. This type of task allows us to investigate whether listeners are aware of the semantic appropriateness of a particular intonation contour to a given

languages there is a distinction between lexical tones that are High and Overhigh (McHugh 1990 for Chaga). Likewise, Francis et al. (2003) report a three-way distinction between lexical tones in Cantonese. In this tonal language, the same syllable /ji/ means 'doctor' when produced with a high-level tone, 'two' when produced with a low-level tone, and 'spaghetti' when produced with a mid-level tone. The results of two identification experiments showed that the perception of Cantonese level tones is qualitatively similar to that presented by Abramson (1979) for Thai level tones. The listeners showed evidence of the presence of category boundaries in an identification task, but no corresponding peaks in discrimination accuracy. Just as there are tonal languages with two or three distinct level tones, it would not be surprising if some intonation languages can make use of more than two level tones to express a variety of pragmatic meanings.

communicative context and can detect an incongruous use of this contour. This methodology has been used successfully by other researchers investigating intonation contrasts (see Rathcke & Harrington 2010, Vanrell 2011, Crespo-Sendra 2011). A set of twenty native speakers participated in the two experiments. Methodologically, we believe that a combination of congruity and identification tasks with three possible responses (along with reaction time measures) can be profitably used to investigate more than two intonational categories in context.

2.2. Experiment 1

2.2.1. Methodology

This experiment consisted of an identification task with three possible response options. In other words, participants had to classify each of the auditory stimuli as conveying one of the three pragmatic meanings of interest in our study, namely IFS, CFS, and CEQ. As noted above, as far as we know no similar triple identification task has thus far been used to investigate potential differences in intonational pitch perception. We initially hypothesized that the triple response procedure would be able to test whether Catalan listeners would be capable of distributing the acoustic pitch range continuum into three or two discrete categories.

Participants

A set of twenty native speakers of Central Catalan participated in the experiment. All subjects were undergraduates studying journalism or translation at the Campus de la Comunicació of the Universitat Pompeu Fabra in Barcelona and were paid for their participation. They were 7 men and 13 women. All were right-handed and none of them had previous experience with linguistic perception tasks. The age of the participants was between 19 and 37 (average = 21.6, standard deviation = 4.07). The average Catalan dominance of the participants (taken from a report on the daily interactions per day in Catalan provided by the participants themselves) was 86% (standard deviation = 12.83%).

Materials

We first recorded the three short dialogs shown in (2) in order to produce an appropriate context for an IFS (2a), a CFS (2b), and a CEQ (2c). A male Catalan native speaker was recorded using a Marantz PMD-660 digital recorder in a quiet room at the Universitat Pompeu Fabra. The productions were elicited using the discourse completion test method (Nurani 2009).

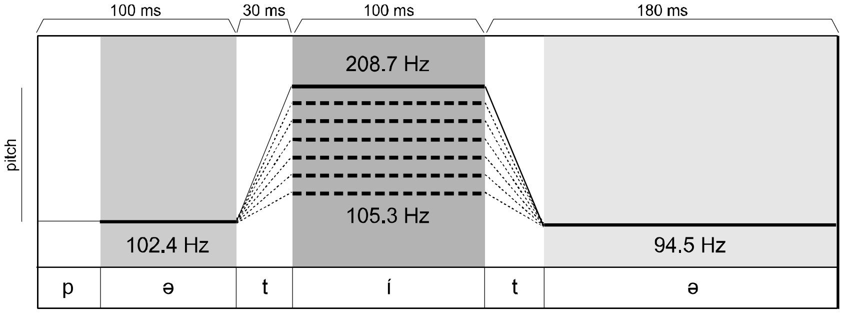
(2)	a.	(IFS)	Com la vols, la cullera?	*What type of spoon do you want?*
			Petita, [sisplau].	*[I want a] little [spoon, please].*
	b.	(CFS)	Volies una cullera gran, no?	*You want a big spoon, don't you?*
			Petita, [la vull, i no gran].	*[I want a] little [one, not a big one].*
	c.	(CEQ)	Jo la vull petita, la cullera	*I want a little spoon.*
			Petita? [N'estàs segur?]	*[A] little [one]? [Are you sure?]*

We then created a synthesized continuum for the noun phrase *petita* [pə.'ti.tə] ('little'-fem) by modifying the f0 peak height in 11 steps (distance between each one = 1.2 semitones).[5] A single item was used so that listeners could easily keep in mind the three linguistic contexts provided at the beginning of the task. The speech manipulation was performed on a single [pə.'ti.tə] recording by means of the Pitch Synchronous Overlap and Add (PSOLA) resynthesis routine available in the Praat speech analysis and resynthesis software (Boersma & Weenink 2008), which keeps the segmental information invariable, thus making it possible to

[5] This target word (which contains voiceless plosives) was selected so that we would be able to use the same target materials as in the electrophysiological experiment that is presented in Chapter 3, which required to have a voiceless segment in order to adequately control for the specific point in time in which the auditory mismatch occurs.

test for only the changes in pitch height. Figure 2 shows an idealized schema of the pitch manipulation in the target noun phrase. As shown in the figure, pitch movements were realized with a rising tonal movement starting at onset of the accented syllable /'ti/, which was preceded by a low plateau for the syllable [pə] (102.4 Hz, 100 ms). The posttonic syllable [tə] was realized with a falling tonal movement (94.5 Hz, 180 ms). The peak height continuum ranged from 105.3 Hz to 208.7 Hz, and the total duration of each stimulus was 410 ms.

Figure 2: Idealized schema of the pitch manipulation in the noun phrase *petita* [pə.'ti.tə] ('little'-fem.). Duration of the segments is shown at the top, and the correspondence with each segment is shown at the bottom. The Hz values at the center of the image represent the final frequencies of the extreme stimuli (steps 1 and 11).



Procedure

Participants were instructed to pay attention to the intonation of the stimuli and indicate which interpretation was more likely for each stimulus by pressing the corresponding computer key, namely "A" for *Afirmació* ('Statement', i.e., IFS), "C" for *Correcció* ('Correction', i.e., CFS) and "P" for *Pregunta* ('Question', i.e., CEQ). These three labels were chosen because they would suggest intuitive response labels to participants with no previous experience with linguistic perception tasks. Prior to the experiment, subjects gave verbal confirmation to the experimenter of their understanding of the three different linguistic contexts.

The task consisted of 6 blocks in which all stimuli in the continuum were presented to the subjects in a randomized order, i.e., the order of the stimuli inside each trial list was different for each block (with no order constraints) and for each subject. An interval of 15 seconds of silence was inserted between each block. The interstimulus interval was set at 1s. We obtained a total of 1,320 responses for this experiment (11 steps × 6 blocks × 20 listeners). The experiment lasted approximately 8 minutes. This includes a brief training session intended to get subjects used to the stimuli and the task, which consisted of the same procedure as the experimental task with the difference that subjects were asked only to identify isolated instances of extreme and central stimuli (specifically, stimuli 1, 2, 5, 7, 10, and 11). No feedback was provided. No counterbalancing was used between Experiments 1 and 2 (see description below), and subjects performed a distractor behavioral task between the two experimental segments which consisted of identifying which one was the stressed syllable of a set invented words produced with seven different intonational contours.

The experiment was set up by means of the psychology software E-prime version 2.0 (Psychology Software Tools Inc. 2009), and identification responses and RTs were automatically recorded using this software. Subjects were instructed to press the button as quickly as they could. The experiment was set up in such a way that the next stimulus was presented only after a response had been given.

2.2.2. Results

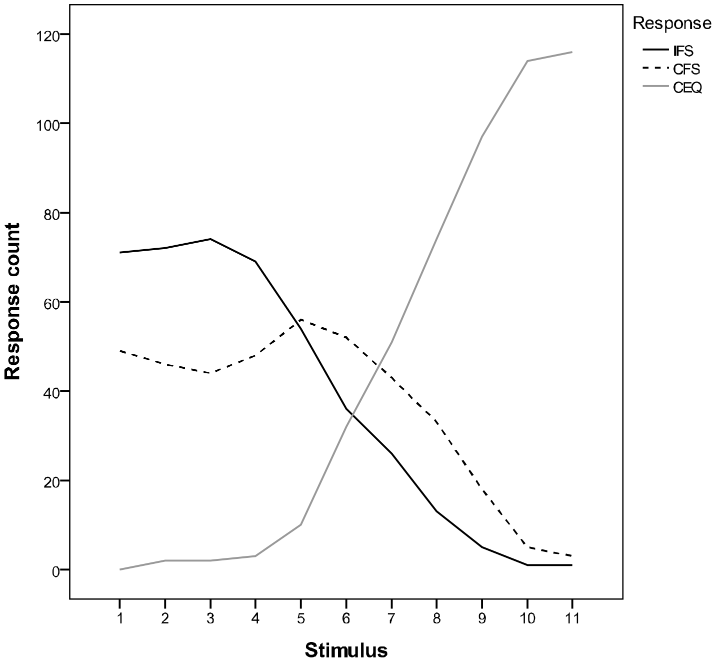
Identification results

Figure 3 shows the results of Experiment 1. The y-axis represents the absolute number of responses given to each stimulus. The x-axis represents the steps of the acoustic continuum. Different line types represent the different identification responses given (IFS: solid black line, CFS: dashed black line, CEQ: solid grey line). The graph actually presents a summary of how the participants categorized the acoustic space into three parts. On the one hand, it shows that the distribution of IFS and CFS responses are closer and more frequent for the lower stimuli, roughly differentiable between stimuli 1 and 4. On the other hand, the distribution of CEQ responses is clearly different from that of statements and shows a great frequency between stimuli 8 and 11. Thus, the graph shows that responses present an unsettled distribution between stimuli 5 and 7.

A Generalized Linear Mixed Model (GLMM) analysis (multinomial distribution) was performed with identification of the three possible categories as the dependent variable.[6] Stimulus was set as the fixed factor, and subject × block were set as crossed random factors (thus avoiding at the same time inter-subject variation and possible effects of fatigue, boredom, and practice). Results showed a significant effect of stimulus over the response given ($F_{20,\,1299} = 19.014$, $p < .001$).

[6] All responses and RT were analyzed through a Generalized Linear Mixed Model (GLMM) using IBM SPSS Statistics 19.0 (IBM Corp. 2010). As Baayen et al. (2008) and Quené and van den Bergh (2008) point out, mixed-effects modeling offers considerable advantages over repeated measures ANOVA. Specifically for our data, they are suitable to analyze noncontinuous dependent variables, such as binomial and multinomial responses. On the other hand, we can control for both fixed and random factors (in our case, SUBJECT and BLOCK) at the same time.

Figure 3: Absolute number of given responses for each stimulus, for Experiment 1. IFS = solid black line; CFS = dashed line; CEQ = solid grey line.



Because a multinomial distribution of the dependent variable does not allow the extraction of estimated means, new GLMM analyses were conducted for each possible pair of responses (namely IFS vs. CFS, IFS vs. CEQ, and CFS vs. CEQ) in order to determine whether identification responses, when compared one to another, would show a significant distribution among the stimuli in the continuum. The overall test results showed a lower result of the Fisher's F test when applied to the comparison between the two types of statements: IFS vs. CFS ($F_{10,\ 808}$ = 4.706, p < .001), IFS vs. CEQ ($F_{10,\ 913}$ = 24.878, p < .001), and CFS vs. CEQ ($F_{10,\ 888}$ = 25.891, p < .001). This means that the different distribution of IFS and CFS among the stimuli is less clear than when each of these given responses are compared with the distribution of CEQ responses.

Table 2 shows the results of the Bonferroni deviation contrasts within each stimulus in the continuum. These results provide important information for detecting that each pair of categories has a significantly different distribution along the acoustic continuum, i.e., the distribution of responses is significantly different between the three categories for each step in the continuum. We have two exceptions to this generalization, namely, that (a) as expected, at stimulus number 6, the comparisons between IFS vs. CFS, IFS vs. CEQ, and CFS vs. CEQ are not shown to be significant; and (b) at stimulus numbers 7-11, there is no significant difference between IFS and CFS, revealing that the distributions of their responses are similar.

Table 2. Results of the Bonferroni deviance contrasts (over each possible pair of responses) within each stimulus of Experiment 1.

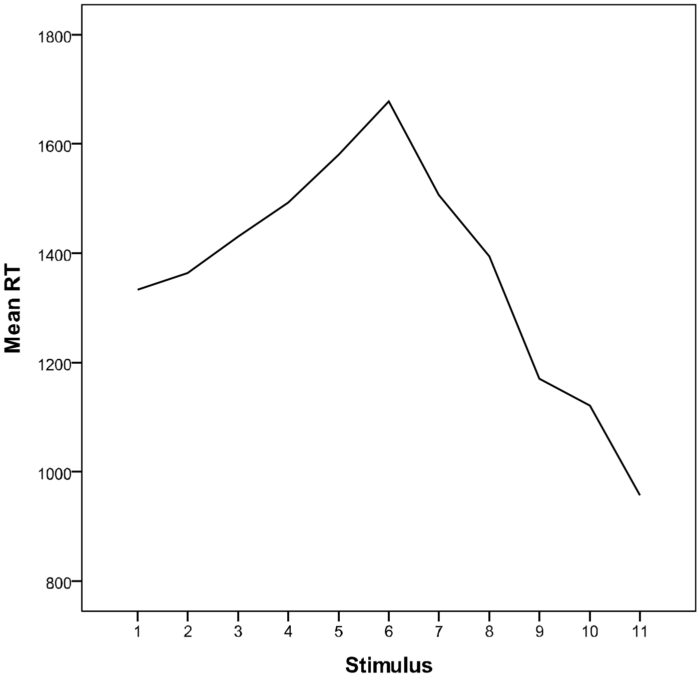
	IFS vs. CFS		IFS vs. CEQ		CFS vs. CEQ	
stimulus	t	Sig.	t	Sig.	T	Sig.
1	-3.332	**.008**	-4.761	**<.001**	-3.822	**.001**
2	-3.630	**.003**	-5.508	**<.001**	-4.114	**<.001**
3	-3.918	**.001**	-5.260	**<.001**	-4.110	**<.001**
4	-3.281	**.009**	-5.501	**<.001**	-4.371	**<.001**
5	-1.546	.613	-4.549	**<.001**	-3.564	**.001**
6	-0.092	1.000	-0.118	.906	0.001	.999
7	0.420	1.000	2.759	**.012**	2.905	**.008**
8	1.703	.533	5.814	**<.001**	5.748	**<.001**
9	1.973	.342	7.099	**<.001**	8.251	**<.001**
10	1.324	.743	5.600	**<.001**	8.670	**<.001**
11	0.699	1.000	5.669	**<.001**	7.954	**<.001**

In sum, the results of the triple response identification task indicate that Catalan listeners clearly associate the higher end of the pitch range continuum with a CEQ interpretation, and that they perceive a greater degree of ambiguity when processing the lower end of the pitch range continuum, with a very similar distribution between IFS and CFS interpretations.

Reaction times

Figure 4 shows the averaged RTs for each pair of responses obtained in Experiment 1. The y-axis represents the mean RT, and the x-axis represents the steps in the pitch continuum. The graph shows a clear RT peak at stimulus 6, with a more pronounced slope towards the high end of the continuum than towards the low end.

Figure 4: Averaged reaction time (RT) measures (in ms) for Experiment 1.



A GLMM was applied with the RT measures as the dependent variable, stimulus, response given and their interaction as fixed factors, and subject × block as crossed random factors. There were significant effects for stimulus ($F_{10,\ 1211}$ = 2.732, p = .003), response given ($F_{2,\ 1254}$ = 5.402, p = .005), and their interaction ($F_{20,\ 1227}$ = 2.379, *p* = .001). In order to determine whether stimulus had a significant variation within each response given, deviation contrasts were extracted. The overall test results showed an effect of stimulus for IFS ($F_{10, 1288}$ = 5.917, *p* < .001) and CEQ ($F_{10, 1288}$ = 2.318, *p* = .011), but not for CFS ($F_{10, 1288}$ = 1.766, *p* = .062). This means that we can only argue for a RT peak when IFS and CEQ responses were analyzed.

In sum, the results of the triple response identification task indicate that Catalan listeners associate the higher end of the pitch range continuum with a CEQ interpretation, and that they display more perceptual confusion in the lower end of the pitch range continuum, which is distributed between the IFS and CFS responses. Taking into account the RT measures, this suggests a fairly close association of the lower end of the continuum with IFS responses, but no clear conclusions about the role of pitch range in determining a CFS interpretation.

2.3. Experiment 2

2.3.1. Methodology

This experiment consisted of a congruity task which had the goal of assessing participants' preference for a particular stimulus as more acceptable in a given communicative context. As noted above, this task makes it possible to investigate whether listeners are aware of the semantic degree of appropriateness of a particular intonation contour to a given discourse context and whether they are able to detect an incongruous use of this contour.

Participants

For this experiment, the same set of participants was presented with the three types of linguistic contexts shown in (2), each time followed by the target utterance *Petita* ('little'-fem). The same set of subjects participated in both experiments because this would increase the comparability between the results of the two tasks.

Materials

The context recordings were of a female native speaker of Central Catalan. Each context was systematically combined with all the target utterances. Their duration was approximately 1,450 ms and their pitch range was between approximately 176.27 Hz and 299.17 Hz. An interval of 300 ms of silence was inserted between the context and the target utterance.

In this experiment we used 6 stimuli only, specifically, stimuli 1-3-5-7-9-11 from the continuum used in Experiment 1. Thus, the distance between each step in the continuum in this case was 2.4 semitones rather than 1.2.

Procedure

Subjects were asked to rate the target word as being semantically ‘appropriate’ or ‘inappropriate’ within that specific linguistic context by pressing the corresponding computer key, namely “A” for *adequat* (‘appropriate’) and “I” for *inadequat* (‘inappropriate’). Thus, we obtained information about the perceived congruity of each combination of linguistic context + target stimulus. A brief training session was conducted prior to the task, consisting of rating the acceptability of stimuli 3 and 9 within each of the three communicative contexts. As in Experiment 1, the aim of the training session was merely to get participants used to the task and they received no feedback. (Stimuli 2 and 5 were chosen because they were neither extreme nor central in the auditory continuum and were equidistant from the midpoint.)

The task consisted of 5 blocks in which all stimuli in the continuum were presented twice within each of the three linguistic contexts in a randomized order. We thus obtained a total of 3,600 responses for this experiment (6 steps × 3 linguistic contexts × 5 blocks × 2 repetitions × 20 listeners). The experiment lasted approximately 22 minutes.

2.3.2. Results

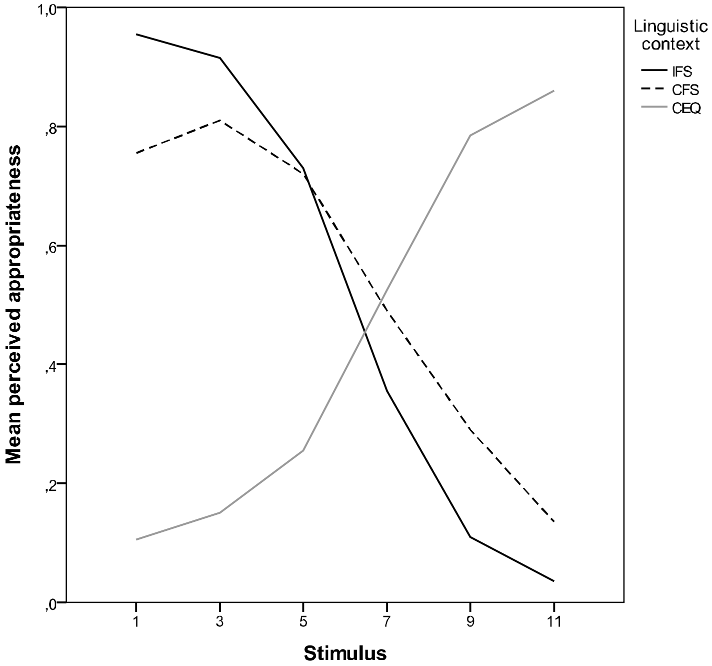
Congruity results

Figure 5 shows the semantic congruity results of our experiment. The y-axis represents the mean perceived appropriateness between the linguistic context and the target stimulus (x-axis). Different line types represent the linguistic contexts heard (IFS: solid black line, CFS: dashed black line, CEQ: solid grey line). For instance, stimulus 1 was accepted at a rate of .97 (i.e., 97% of the time) when occurring in an IFS linguistic context, .77 when occurring in the CFS context, and only .09 when occurring in the CEQ context. And the opposite pattern of results was obtained for stimulus 11. Interestingly, the results reveal that stimuli 1-5 are generally rated as appropriate for both IFS and CFS contexts: while the IFS and CFS functions are similar, they sharply contrast with the function found for the CEQ linguistic context. Subjects seem to divide the six-point continuum into two general categories, i.e., ‘statement’ and ‘question’, with the boundary located at stimulus 7 (which corresponds to 158.5 Hz), thus assigning both IFS and CFS to stimuli 1-5 and CEQ to 9-11.

A GLMM analysis (binomial distribution) was conducted with appropriateness as the dependent variable, linguistic context, stimulus, and their interaction as fixed factors, and subject × block as crossed random factors. Main effects of linguistic context ($F_{2,\ 3582} = 8.810$, $p < .001$) and stimulus ($F_{5,\ 3582} = 29.284$, $p < .001$) were found and, crucially, an interaction between linguistic context and stimulus ($F_{10,\ 3582} = 92.269$, $p < .001$) was also detected.

In order to know how the three meanings are distributed in the pitch range continuum, we must analyze which part of the continuum contains a significant number of ‘appropriate’ and ‘inappropriate’ responses for each discourse context separately. To this end, Bonferroni deviation contrasts were extracted (over the two available responses, i.e., ‘appropriate’ and ‘inappropriate’) within each stimulus. The results of the deviation contrasts are

Figure 5. Mean rate of appropriateness for each type of communicative situation (IFS context: solid black line, CFS context: dashed line, CEQ context: solid grey line).



presented in Table 3. The first column for each meaning contains the results of the *t* tests (where a positive value indicates a preference for 'appropriate' responses), and the second column contains the significance of this preference (all < .001 except when stimulus 7 is presented with a CFS context). More specifically, it is shown that stimuli 1-5 were significantly categorized as 'appropriate' for an IFS context, and 7-11 were considered 'inappropriate'. For CFS, stimuli 1-5 were considered 'appropriate', stimulus 7 was not associated with any response, and stimuli 9-11 were considered 'inappropriate'. For CEQ, stimuli 1-5 were considered 'inappropriate', and stimuli 7-11 were considered 'appropriate'. The roughly parallel results for IFS and CFS indicate that both meanings share the lower part (stimuli 1-5) as appropriate pitch range values, whereas CEQ occupy the higher part (stimuli 7-11). This means that (a) the location of the boundary between statements (IFS and CFS) and questions (CEQ) falls immediately before stimulus 7, and (b) IFS and CFS contexts share the same perceptual behavior, both contrasting with the distribution of CEQ responses.

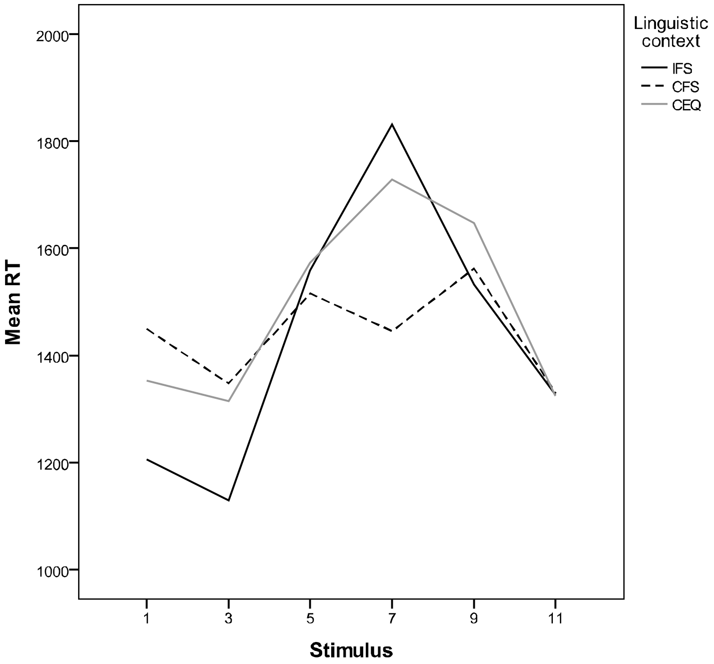
Table 3. Results of the Bonferroni deviance contrasts (applied to 'appropriate' and 'inappropriate' responses) within each stimulus, for the three linguistic contexts.

	IFS context		CFS context		CEQ context	
stimulus	t	Sig.	t	Sig.	t	Sig.
1	**10.094**	<.001	**6.735**	<.001	-9.227	<.001
3	**9.981**	<.001	**8.272**	<.001	-8.178	<.001
5	**5.538**	<.001	**5.712**	<.001	-5.181	<.001
7	-4.153	<.001	-1.207	.227	**2.849**	<.001
9	-10.210	<.001	-7.104	<.001	**10.030**	<.001
11	-10.304	<.001	-10.972	<.001	**11.637**	<.001

Reaction times

Figure 6 shows the averaged RTs obtained for each linguistic context in our congruity test (IFS: solid black line, CFS: dashed line, CEQ: solid grey line). The y-axis represents the mean RT, and the x-axis represents the steps in the acoustic continuum. Specifically, the analysis of RT measures in a congruity test can shed light on the potential perceptual confusion of associating a given pitch range with a specific linguistic context (i.e., a RT peak for a specific meaning can be interpreted as indicating that that meaning has a specific pitch range for its production). The graph indicates a clear increase in RTs observed near stimulus 4 for both IFS and CEQ contexts, but not for CFS. This coincides with our analysis of the RT of the identification task.

Figure 6. Averaged reaction time (RT) measures (in ms), according to linguistic contexts (IFS: solid black line, CFS: dashed line, CEQ: solid grey line).



A GLMM analysis (binomial distribution) was conducted, with the RT as the dependent variable, linguistic context and stimulus as fixed factors, and subject × block as crossed random factors. A main effect of stimulus ($F_{5,\ 3383}$ = 11,024, $p < .001$) was found. There was no effect of linguistic context ($F_{2,\ 3383}$ = 0.796, $p = .451$) and only a near-significant interaction between linguistic context and stimulus ($F_{10,\ 3383}$ = 1.801, $p = .055$).

In order to analyze the patterns of RT obtained for each discourse context, deviation contrasts were extracted, with a sequential Bonferroni adjusted significance level at .05. The overall test results showed an effect of stimulus for IFS ($F_{5,\ 3582}$ = 9.081, $p < .001$) and CEQ ($F_{5,\ 3582}$ = 4.437, $p < .001$), but not for CFS ($F_{5,\ 3582}$ = 1.108, $p = .354$). The sequential Bonferroni deviation contrasts (over the RT) showed that there was a significant RT peak in stimulus 4 for IFS (t_{3383} = 5.078, p < .001) and CEQ (t_{3383} = 3.021, p = .015), but not for CFS (t_{3383} = 0.047, p = 1).

In sum, the robustness of the RT results of the congruity test and its coincidence with the results of Experiment 2 shows that this type of task is very informative and useful when trying to uncover the phonologically relevant contrasts in intonation.

2.4. Discussion

The main goal of this chapter was to investigate the role of pitch accent range in conveying intonational differences in a language with a potential three-way pitch range contrast. We have investigated the potential phonological distinction between information focus statements (IFS), contrastive focus statements (CFS), and counter-expectational questions (CEQ) in Central Catalan by performing two complementary experimental tasks.

Experiment 1 tested the participants' interpretation of each isolated stimulus using a triple response identification task. The results of this experiment showed how participants distributed the acoustic continuum across the three possible responses. They associated IFS and CEQ with the lower and higher ends of the continuum respectively, while CFS responses were less consistently associated and skewed towards the lower stimuli (see Figure 3). In order to corroborate the results from this triple identification task and also take explicitly into account the linguistic context in which these three meanings can occur, a semantic congruity test was also conducted (Experiment 2). The results showed that the lower stimuli (1-5) were judged significantly more appropriate for both IFS and CFS contexts, while the higher stimuli (7-11) were the most congruent within the CEQ context (see Figure 5). Thus these results confirm the results from Experiment 1, namely that Catalan listeners associate the lower end of the pitch range continuum with statements (i.e., IFS and CFS) and the higher end of the continuum with questions.

Concerning the analysis of RT measures, as expected, they were found to correlate with the identification results and to increase for the stimuli located in the acoustic frontier between phonological distinctions. Experiment 1 showed a significant peak located at stimulus 6 and a significant role of pitch range only for IFS and CEQ interpretations. The analysis of RT measures from Experiment 2 clarifies this result because only two RT peaks were found, again for IFS and CEQ contexts (IFS: peak at stimulus 7; CEQ: peak-plateau at stimuli 5-9). Interestingly, the analysis of the RT from both experiments shows no significant role for the pitch range when CFS is involved. Following Chen (2003), a mean RT peak at the identification boundary indicates that an intonational contrast is discrete, so for the results of our experiments we cannot claim that CFS can be categorically determined by the pitch range (especially when it is compared to IFS, taking into account the identification responses), and only a gradient effect for pitch range is suggested in the identification of CFS.

Borràs-Comes et al. (2010) tested the participants' interpretation of similar isolated stimuli in a binomial way by comparing the perception of IFS vs. CEQ and IFS vs. CFS. No differences were found between the two identification functions, which meant that, according to identification responses, CFS and CEQ would be associated with similar pitch range values. However, the results of our present study show that, when participants are allowed to give any of the three possible responses, IFS and CFS show a similar distribution in the pitch range continuum. In line with this, we suggest that we need to use binomial identification tasks with caution, as they might be unsuitable for investigating differences in intonational categories if no additional measures (e.g., RTs or congruity tasks) are taken into account (see Chen 2003). In our view, if listeners have only two responses available for responding they can easily train themselves to categorize the given acoustic space into the two categories available (Ladd, p.c.). We thus argue that the extra cognitive load that a triple-response identification task asks for significantly increases the reliability of participants' categorization responses. Notwithstanding, the results of the congruity task are slightly different from those of the triple-identification task, especially in the distribution of IFS and CFS responses among lower stimuli (different in Experiment 1, but similar in Experiment 2). In this case, the triple-identification task might still lead participants to over-categorize the stimuli among all available responses. By contrast, congruity tasks crucially take into account linguistic context, i.e., the stimuli are always evaluated for their congruity or incongruity with the preceding context.

Thus, concerning methodology we would like to highlight the usefulness of using triple-answer identification tasks together with semantic congruity tests to investigate the phonological status of intonational contrasts. First, the results of these tasks reveal that listeners are not simply dividing the acoustic space into three categories. Second, one of the main advantages of using a congruity test is that it takes pragmatic context into account, by evaluating the degree of linguistic appropriateness of different intonation patterns within different discourse contexts. We thus argue that the use of triple-identification tasks together with semantic congruity tests can be a very effective strategy for the investigation of intonational phonology across languages.

Taken together, the two experiments have crucially shown that variation in pitch range is the main cue that Catalan listeners use to discriminate between IFS and CEQ, i.e., there is a threshold along a continuum of pitch range beyond which a CEQ meaning is consistently understood.

In line with our results, it is important to note that the identification of CFS in Catalan does not crucially rely on pitch height differences. Recent production results reported by Vanrell et al. (2012) showed that pitch range is not a stable cue in distinguishing non-focal vs. contrastive focal accents in Catalan. The absence of a categorical difference between IFS and CFS with respect to pitch height might thus be related with the reported preference for Catalan to use changes in syntactic structure for contrastive focus marking (Vallduví 1991).[7] Moreover, contextual pragmatic inference can be important to detect CFS online. As stated in Levinson (2010), pragmatic inference works well enough to detect more than half of all the *yes-no* questions that appear in English spontaneous speech (see also Stivers 2010), so it is possible to classify as a CFS any contradictory utterance provided as simply the last word in a conversation (i.e., when someone contradicts the assumption of the interlocutor, then it is assumed that they know the information at issue).

Beyond the specifics of Catalan intonation, which needs to be able to signal a phonological distinction between counter-expectational questions [L+¡H*] and statements [L+H*] (Aguilar et al. 2009), there is a more general issue that should be considered within the AM system, which is the concept of upstep. By including a category L+¡H* in the Cat_ToBI phonological analysis (and in any other ToBI analysis) — i.e., the upstepped high tone, as represented with a '¡' initial exclamation mark —, the concept of upstep becomes more ambiguous. This concept originally represented the raising of a H tone caused by the presence of a preceding H tone in the same prosodic phrase. Yet the inclusion of a tone like [L+¡H*] means that upstep is being used to expand the inventory of available pitch-accent phonological contrasts. This is also the case with the now common use of !H (especially in the field of boundary tones) to indicate a contrastive use of another level of pitch height. This has been also noted by Face (2011) for Castilian Spanish, who argues for an AM transcription system which takes pitch range into account without altering the dichotomy between L and H targets that exists in the ToBI system. He proposes that an intonational domain (which can range from a pitch accent to an intonational phrase) can be specified by a

[7] Taking the example *Vull TARONGES* 'I want ORANGES' (extracted from Prieto & Cabré 2007-2012), if the speaker wants to focalize the constituent *TARONGES* 'ORANGES' (i.e., s/he wants ORANGES and not some other fruit), s/he will resort most often to clause-external detachment (*TARONGES, vull* 'ORANGES, I want').

‘frame’ that sets “the space for the realization of the f0 rises and falls” (Face 2011: 89). Following Face, the Catalan IFS contour might be labeled [L+H*], while the Catalan CEQ contour might be labeled $_{\mathrm{H+}}$[L+H*], which would indicate that the high end of the continuum would be extended. This is an alternative transcription strategy which should be evaluated with rigor but which is beyond the scope of this investigation.

All in all, the results presented here represent new empirical evidence that pitch accent range variation can express categorical differences in meaning (Hirschberg & Ward 1992, Ward & Hirschberg 1985, Ladd 1994, Ladd 1996, Ladd & Morton 1997, Chen 2003, Savino & Grice 2011, Vanrell 2006, 2011).[8] As mentioned above, the distinction between two levels of pitch height to distinguish statements from questions is very productive in other Romance languages (Savino & Grice 2011 for Bari Italian, Roseano et al. 2011 for Friulian, Estebas-Vilaplana & Prieto 2010 for Castilian Spanish), as well as in other languages, and this distinction needs to be reflected in the intonational phonology of such languages.

[8] The use of higher F0 peaks can be related to the general finding that the average pitch in questions is higher than the average pitch in non-questions (Bolinger 1986), what has been analyzed as a «discretised» manifestation of the so-called Frequency Code (Gussenhoven 1999).

CHAPTER 3

Specific neural traces for intonation-based discourse categories

3.1. Introduction

A series of studies have indicated that segmental and tonal phonological distinctions can be represented in pre-attentive auditory sensory memory. However, there is no conclusive evidence with respect to the neurophysiological representation of intonational discourse contrasts (i.e. between statements and questions), and no previous research has dealt with the processing of intonational within-category and across-category contrasts. In this chapter we report a study that uses the auditory mismatch negativity (MMN) event-related brain potential (ERP) to test the native perception of within-category and across-category intonational contrasts between statement and question interpretations in Catalan. We hypothesize that discrete intonational information — as discrete phonological information — can be represented through symbolic memory traces (in contrast to mere acoustic memory traces) in the brain.

The MMN component is a negative deflection of the auditory ERP occurring between 100 and 250 ms after the onset of a stimulus violating an established acoustic regularity. Traditionally, it is obtained by subtracting the ERP to a standard stimulus from that to a deviant stimulus that is presented in the same block of trials. The MMN is generally elicited in non-attentive conditions and typically argued to reflect pre-attentive detection of auditory changes and higher-level cognitive processes in the auditory system (Näätänen 2001, Pulvermüller & Shtyrov 2006). Following Näätänen (2001), the MMN reflects the early access to stored linguistic representations and indicates the match or mismatch between a stimulus and its corresponding symbolic memory trace in the brain. According to Pulvermüller & Shtyrov (2006), the MMN for language stimuli is composed of at least two parts: a part which reflects the automatic detection of a sound change and a part that reflects the activation of cortical cell assemblies forming the long-term memory traces for learned cognitive representations (see Fournier et al. 2010 for a review of the studies on the lateralization of tonal and intonational pitch processing).

The MMN has been successfully applied in studies of segmental phonetic and phonological analysis (e.g., Sharma & Dorman 2000, Dehaene-Lambertz 1997, Näätänen et al. 1997, Winkler et al. 1999) and abstract phonological features (Eulitz & Lahiri 2004, Phillips et al. 2000; for a review, see Näätänen et al. 2007, Näätänen, 2001). Näätänen et al. (1997) suggested that the identification of the deviant as a native-language vowel enhanced the MMN amplitude, i.e. the phonological representation of a vowel sound can be probed with the mismatch process. Native across-category consonant contrasts also elicit a significant MMN compared to non-native contrasts or within-category contrasts (Dehaene-Lambertz 1997). A series of studies have demonstrated that acoustic contrasts that cross a phonemic boundary lead to larger MMN responses than comparable acoustic contrasts that do not (Aaltonen et al. 1997, Dehaene-Lambertz 1997, Phillips et al. 2000, Sharma & Dorman 1999). In fact, the MMN response is not just larger but rather includes a separate subcomponent when the phoneme boundary is crossed. For example, the same voice onset time span crossing an English phonemic category boundary evokes a far larger MMN than one that does not (Phillips et al. 2000). These results show that discrete phonological representations can be accessed by the auditory cortex, thus providing the basis for lexical storage and further linguistic computation.

Tonal languages have successfully explored experience-dependent effects on the automatic processing of phonologically contrastive pitch (Gandour et al. 1994, Klein et al. 2001, Chandrasekaran et al. 2007, Chandrasekaran et al. 2009, Ren et al. 2009, Xi et al. 2010). Chandrasekaran et al. (2007) showed that early cortical processing of pitch contours might be shaped by the relative saliency of acoustic dimensions underlying the pitch contrasts of a particular language.

However, very few studies have examined suprasegmental prosodic contrasts that convey discursive or pragmatic meanings in intonational languages, like declarative vs. interrogative intent, and their results are controversial. In Doherty et al.'s (2004) study, a set of English speakers made judgments about falling statements (e.g., *She was serving up the meal*), rising declarative questions (with no word order change) and falling questions with the corresponding word order change (e.g., *Was she serving up the meal?*). The authors found an increased BOLD activity for rising declarative questions over the falling counterparts, and they argued that the differences may reflect the presence of a subtle aspect of illocutionary force (conduciveness) in the utterances with rising intonational contours (see also Post et al. *in press*). Fournier et al. (2010) examined the processing of lexical-tonal and intonational contrasts by speakers of an intonational language (standard Dutch) and of a tonal dialectal variety of Dutch (Roermond Dutch). They assumed that the brain responses to the stimuli would depend on the subjects' language experience, but no group differences were found. The authors argued that the expression and recognition of discourse meanings by means of intonation, which is considered universal amongst languages, was not necessarily realized in an identical way in the human brain. Finally, Leitman et al. (2009) employed two artificial sequential sinusoidal tones corresponding to English declaratives and interrogatives. An "interrogative" deviant block and a "declarative" deviant block were presented, and authors found significant MMN responses in both conditions.

In sum, the representation of segmental and tonal phonological distinctions is found to be evident by means of the MMN, but this is not the case of intonational discourse contrasts. The abovementioned MMN results and its magneto-encephalographic (MEG) counterpart on intonational discourse contrasts could be interpreted as detections of acoustic changes in the stimuli, and remain far from signaling intonationally-based phonological distinctions indicating different meanings. Moreover, no previous study has examined the processing of intonational across-category contrasts (e.g. between statements and questions) and within-category contrasts (e.g. between two types of statements or two types of questions). The abovementioned studies exclusively used minimal pairs as their basic stimuli and, furthermore, they did not show any evidence for language-specific phonological representations or traces for intonational contrasts.

Interestingly, in Catalan, a rising-falling intonational sequence can be perceived as an information focus statement (IFS) or as a counter-expectational question (CEQ) depending exclusively on the size of the pitch range interval of the rising movement. The two rising-falling pitch contours consist of a rising movement associated with the stressed syllable followed by a falling F0 movement associated with the posttonic syllables (see Figure 7; also see Chapter 2). The following examples in (3) show two typical discourse contexts in which these intonational configurations could be found. An IFS context is shown in (3a), and a CEQ in (3b). In both cases, the target word *petita* [pə.ˈti.tə] (‘little’-fem.) is typically produced with a low tone on the first syllable, a rising/high tone associated with the second (stressed) syllable followed by a falling/low tone associated with the third (posttonic syllable). The prosodic difference between (3a) and (3b) lies on the pitch range difference between the low and the high tone, which is expanded in the case of CEQ.

(3)	a.	Com la vols, la cullera?	*What type of spoon do you want?*
		Petita, [sisplau].	*[I want a] little [spoon, please].*
	b.	Jo la vull petita, la cullera	*I want a little spoon.*
		Petita? [N’estàs segur?]	*[A] little [one]? [Are you sure?]*

In Chapter 2 we present a set of behavioral experiments (identification and congruity tasks) which confirm that a categorical phonological contrast exists between these two types of rising-falling contours (compressed vs. expanded pitch range) and that they cue an IFS and a CEQ interpretation respectively. These results represent further evidence that pitch range differences can be used to cue intonational distinctions at the phonological level, in line with the findings of other languages (Savino & Grice 2011, Vanrell et al. *in press*). In turn, this finding strengthens the idea that pitch range differences can cue phonological distinctions in the intonational grammar of a non-tonal language like Catalan (Aguilar et al. 2009), thus expanding the inventory of potential grammatical units in the description of pitch movements.

The goal of the present chapter is to test whether the intonational contrasts differentiating IFS and CEQ in Catalan can elicit specific MMN responses, thus providing electrophysiological evidence in favor of the idea that the auditory cortex supports distinctive linguistic representations at the intonational level. The article presents a behavioral identification experiment (Experiment 1) and an ERP study consisting of 3 oddball blocks with the aim of finding electrophysiological evidence for this discrete distinction (Experiment 2).

3.2. Experiment 1

In Experiment 1, subjects participated in an identifications task whose goal was to identify each of the two meanings (Statement and Question) for a set of 16 stimuli in a pitch range continuum. The goal of Experiment 1 was twofold. First, to corroborate the phonological role of pitch range expansion in the interpretation of rising-falling intonational contours in Catalan found in Chapter 2. Second, to determine the pitch region at which the change in categorization occurs and thus select the target stimuli for the MMN oddball experiment. The same set of participants was enrolled in the auditory event-related brain potential experiment several weeks later.

3.2.1. Methodology

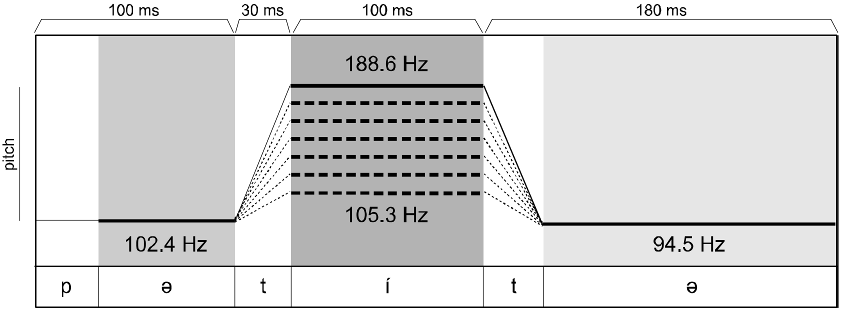
Participants

Fifteen healthy volunteers (3 male, aged 19-42 years, mean age 22.5 years; one left handed) with no history of neurological, psychiatric or hearing impairment and with normal or corrected-to-normal visual acuity participated in the experiment. Subjects reported not having any auditory deficiency and gave informed consent and received monetary compensation for their participation. The study was approved by the Ethics Committee of the University of Barcelona, according to the Code of Ethics of the World Medical Association (Declaration of Helsinki). All participants were native speakers of Central Catalan and musicians were excluded.

Stimuli

To generate the auditory stimuli, a native speaker of Catalan (the first author of this study) read natural productions of the noun phrase *petita* [pə.tí.tə] ('little'-fem) with an IFS pitch contour and a CEQ pitch contour, and these utterances served as the source utterances for our stimuli (Figure 7). The original noun phrase utterances were pronounced with a rising-falling contour. This rising movement was of 0.9 semitones for the IFS and 9.9 semitones for the CEQ. We then converted each syllables' curve to a plateau (taking the mean Hz values for each segment) and normalized the absolute pitch of the pretonic and posttonic syllables of the two utterances (to their mean values). Then, we restored the observed differences of 0.9 and 9.9 semitones, respectively. The height of the accented syllable of the CEQ-based stimuli was then adapted to the value of the IFS stimulus, and no noticeable differences were observed between the stimuli. After this, we normalized the durations of each syllable to the mean values of the two original utterances. The synthesized continuum was created by modifying the F0 peak height in 16 steps (distance between each one = 0.6 semitones; see Figure 7). The speech manipulation was performed by means of Praat (Boersma & Weenink 2008). Each stimulus lasted a total of 410 ms. Rising movements were realized as a 100 ms high plateau starting 30 ms after the onset of the accented syllable /tí/, and were preceded by a low plateau for the syllable [pə] (102.4 Hz, 100 ms). The posttonic syllable [tə] was realized with a low plateau (94.5 Hz, 180 ms). The pretonic and posttonic F0 levels were maintained invariable in all manipulations. The peak height continuum ranged from 105.3 Hz to 188.6 Hz.

Figure 7. Idealized schema of the pitch manipulation in the noun phrase *petita* [pə.ˈti.tə] (‘little’-fem.). Duration of the segments is shown at the top, and the link between each segment is shown at the bottom. The Hz values at the center of the image represent the final frequencies of the extreme stimuli (steps 00 and 15).



Procedure

Stimuli were presented to subjects over headphones and their amplitude was adjusted to a comfortable level. Subjects were instructed to pay attention to the intonation of the stimuli and decide which interpretation was more likely for each stimulus by pressing the corresponding computer key, namely “A” for *Afirmació* (‘Statement’) and “P” for *Pregunta* (‘Question’).

The task consisted of 5 blocks in which all 16 stimuli in the continuum were presented to the subjects in a randomized order, for a total of 80 stimuli. We thus obtained a total of 1,200 responses for Experiment 1 (16 steps × 5 blocks × 15 listeners). The experiment lasted approximately 8 minutes.

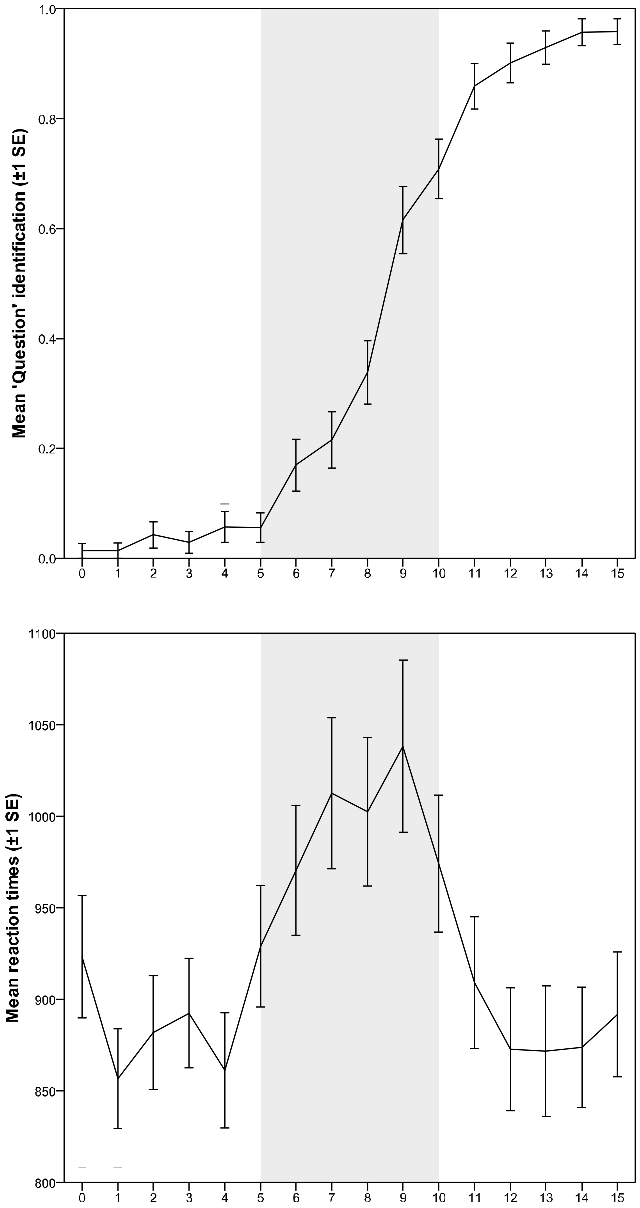
Response frequencies and reaction time (RT) measurements were automatically recorded by means of E-prime version 2.0 (Psychology Software Tools Inc. 2009). The experiment was set up in such a way that the next stimulus was presented only after a response had been given; yet subjects were instructed to press the button as quickly as they could.

3.2.2. Results

A one-way ANOVA was carried out with the proportion of “counterexpectational question” responses as the dependent variable. The data were first checked for the occurrence of possible outliers on the basis of reaction time. Of a total of 1,200 datapoints, 84 cases were treated as outliers, i.e. those cases where the reaction times were at a distance of at least three standard deviations from the overall mean (RTs ≥ 1799). These cases were excluded from the analysis.

Figure 8 shows the identification rate (“y” axis) for the auditory continuum created (“x” axis). This rate is defined as the proportion of “Question” responses that were given over the total. The identification function presents a classic S-shape, revealing that the lowest six stimuli belong to the category “Statement” and the highest five stimuli to “Question”. The perceptual shift from one category to another occurs in the range of stimuli 6 to 11; a full crossover from 16.92% to 85.92% is achieved between these five central steps.

Figure 8. Experiment 1 results. The sixteen stimuli perceived by the listeners are shown in the *x* axis. The left vertical axis represent the mean 'Question' identification responses (Statement = 0 / Question = 1) for all subjects, which are plotted through the black line (error bars showing ±1 Standard Error). The right vertical axis represents the mean reaction times (in ms) for all subjects, which are plotted through the grey area (error bars showing ±1 Standard Error).



The analysis revealed a significant main effect of the auditory stimulus ($F_{15,\ 1100}$ = 117.624, p < .001). Tukey HSD post-hoc tests revealed two main homogeneous subsets, namely between stimulus 0-6 and 11-15, so we can set an area of change of categorization between stimuli 6 and 11. In order to calculate the boundary value between the two categories, the set of data points was fitted to a logistic regression using SPSS (SPSS Inc. 2008). Thus we obtained the boundary value calculated from the "b0" and "b1" values given for the logistic curve using the following formula: boundary = –ln(b0)/ln(b1). Hence, when "y" equals 0.5, "x" is 8.65 (the boundary is therefore located between stimuli 8 and 9).

Figure 8 plots averaged RT responses in ms ("y" axis) for all stimuli ("x" axis). RT were measured from the start of the utterance playback (total length of the utterance = 380 ms). The graph indicates longer RTs for central stimuli, with a clear increase observed for stimuli 7 to 9, which coincides with the area of change reported in the identification function. As expected, listeners displayed faster RTs in identification of within-category exemplars than in exemplars representing the category boundaries.

Results of a univariate ANOVA indicated a statistically significant effect of stimulus type on RT measures ($F_{15,\ 1100}$ = 2.678, p = .001). Duncan post-hoc tests revealed a homogeneous subset between stimuli 0-6 and 10-15 and another one between stimuli 5-10. This second subset between stimuli 5-10 roughly coincides with the area of change of perceptual categorization found in the identification function.

Our behavioral results thus indicate that the variation in pitch range is the main cue that Catalan listeners use to decide between an information focus interpretation (IFS) interpretation and a counter-expectational question (CEQ) interpretation. Taken together, the identification and RT results clearly show that the two intonational categories under examination are categorically perceived. These results replicate the findings presented in

Chapter 2. Experiment 2 will test whether this intonational contrast can be neurophysiologically represented as measured with the MMN.

3.3. Experiment 2

3.3.1. Methodology

The aim of Experiment 2 was to test whether the intonational contrasts differentiating IFS and CEQ in Catalan can elicit a specific MMN response, thus showing electrophysiological evidence supporting that the auditory cortex supports distinctive linguistic representations at the intonational level. We hypothesize that discrete intonational representations, as well as discrete phonological representations, can be represented through symbolic memory traces in the brain (see Pulvermüller & Shtyrov 2006).

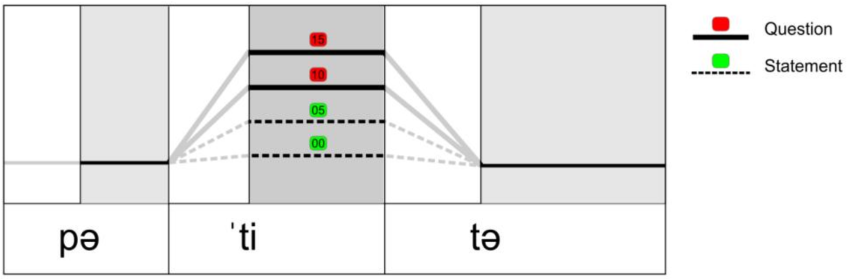
Participants

The same sample of fifteen Catalan speakers that participated in the first experiment volunteered in the present experiment. A period of time of 4 to 9 weeks was elapsed between the two experiments.

Stimuli and procedure

Based on the results of Experiment 1 (i.e., a central area of change of categorization and two tails of within-category variation), four auditory stimuli were selected to be contrasted by pairs in three different oddball blocks (stimuli 00, 05, 10 and 15). The choice was made according to two criteria: 1) the physical distance in semitones between two stimuli within a pair was kept constant (3 semitones); and 2) two stimuli had to be classified as belonging to the "statement" category, and two to the "question" category. Thus, all contrasts involved the same physical difference but the central one (stimuli 05 and 10) involved a categorical difference as well. The idealized intonational contours of the stimuli used are displayed in Figure 9.

Figure 9. Idealized intonational contours of the four stimuli used in the ERP study. Though the same physical difference exists between the four high targets, the extreme pairs represent within-category contrasts, whereas the central pair represents an across-category contrast between statements (IFS) and questions (CEQ), as stated by Experiment 1.



The experiment consisted of 3 oddball blocks presented in random order, with short pauses in between. Each oddball block lasted 21 minutes approximately, and contained 720 standard (STD) stimuli and 180 deviant (DEV) stimuli (80% STD – 20% DEV). STD and DEV stimuli were presented pseudo randomly, with the constraint that a deviant stimulus was preceded by a minimum of two standard stimuli. While the lower pitch stimulus acted as a STD, the higher acted as a DEV, resulting in the following oddball blocks: lower [within-category] (step 00 STD, step 05 DEV), central [across-category] (step 05 STD, step 10 DEV), higher [within-category] (step 10 STD, step 15 DEV).

All stimuli were presented with a fixed SOA of 1400 ms. The onset of the deviance between a pair of stimuli appeared at the second syllable of the token (120 ms after stimulus onset). The use of occlusive phonemes at the beginning of each syllable allowed us to obtain reliably time-locked ERPs (see Pulvermüller 2005). Participants sat in a comfortable chair in a sound-attenuated and electrically shielded room. They were instructed to ignore the sounds delivered by headphones and watch a silent movie with subtitles. The amplitude of the stimuli was adjusted to a comfortable level. The total duration of the experiment was approximately 100 minutes, including the EEG recording preparation.

EEG Recording

The EEG was continuously recorded with frequency limits of 0-138 Hz and digitized at a sampling rate of 512 Hz (EEmagine, ANT Software b.v., Enschede, Netherlands). Ag/AgCl electrodes were used for the EEG acquisition, 33 of which were mounted in a nylon cap (Quik-Cap; Compumedics, Abbotsford, VIC, Australia) according to the international 10-20 system. Vertical and horizontal electrooculogram (EOG) were measured from monopolar electrodes placed respectively below (VEOG) and laterally (HEOG) to the right eye. The ground electrode was located on the chest and the common reference electrode was attached to the tip of the nose. All impedances were kept below 5 kΩ during the whole recording session.

The continuous EEG was further bandpass-filtered off-line between 1 and 20 Hz and cut in epochs of 700ms duration, including a pre-stimulus baseline of 100ms, for each deviant and standard in all 3 conditions (except for the standard following a deviant stimulus; 180 deviant epochs and 540 standard epochs per condition). Epochs with a signal range exceeding 100 µV at any EEG or EOG channel were excluded from the averages, resulting in a mean of 143 deviant epochs (SD = 20.3; 94 minimum) and 325 standard epochs (SD = 47.4; 213 minimum) after rejection.

MMN difference waveforms were obtained by subtracting the ERPs elicited by standard stimuli from those elicited by deviant stimuli. The MMN peak was determined from the Fz electrode as the largest negative peak in the interval of 200-400ms (80-280ms after stimulus onset) for all difference waves and subjects separately. Because MMN peak latencies were not significantly different across conditions, MMN mean amplitudes were derived in a 80ms time window centered on the mean peak latency of the grand-average waveforms for all the 3 conditions (265-345ms).

Data Analysis

The presence of a significant MMN elicited to each intonational contrast was analyzed by means of one-sample t-tests on the MMN amplitude at Fz in each of the three conditions separately. The intonational contrast effects on the MMN peak latencies and mean amplitudes at Fz electrode were evaluated with separate repeated measures ANOVAs including the factor: Contrast (lower [within-category], central [across-category], higher [within-category]). Because MMN inverts its polarity below the Silvian fissure (ref), another repeated measures ANOVA was conducted to assess the effects on the MMN mean amplitude retrieved at Mastoid electrodes, with the factors: Channel (M1, M2) x Contrast (lower, central, higher). The Greenhouse-Geisser correction was applied when appropriate.

In an attempt to relate the electrophysiological responses with behavioral measures, a bivariate correlation analysis was performed between the MMN mean amplitude and the Categorization Index (CI) for all subjects as well as for the grand mean data. For these specific analyses, the EEG data were re-referenced to combined Mastoids in order to better assess the power of the effects. We defined the CI as the difference between the categorization scores to each of the two stimuli in a pair, thus resulting in three measures per subject: lower [within-category] (step 05 – step 00 scores), central [across-category] (step 10 – step 05 scores) and higher [within-category] (step 15 – step 10 scores). The higher the CI, the higher the categorical difference a subject made between a pair of stimuli (please note that we have steps of 0.2 CI because each stimulus in experiment 1 was presented five times to each subject). To further test the significance of the obtained correlation values, we estimated the variability of the correlation statistic (Pearson's correlation coefficient) with the bootstrap method. Bootstrapping is a resampling method that helps to perform statistical inferences without assuming a known probability distribution for the data. In short, the correlation index was calculated for 10000 randomly chosen samples (with replacement) of N=45 (15 subjects x 3 conditions) of MMN amplitude values and CI scores respectively. The obtained distribution (H1; centered at the Pearson's coefficient value that is obtained performing a simple correlation with the raw data) was tested for significance against the null hypothesis distribution (H0), which arises from performing the correlation analysis in 10000 random samples of MMN and CI scores (N=45) pooled together. Thus, the bootstrap method yields a mean of the correlation statistic for the H0 centered at 0, with the confidence intervals (95%) that are used to test the significance of the obtained H1.

3.3.2. Results

Grand average waveforms elicited to STD (dotted line) and DEV (continuous line) stimuli at Fz, M1 and M2 electrodes are shown in Figure 10. DEV minus STD stimuli difference waveforms are shown in Figure 11. The mean values of the DEV minus STD waveforms at the 266-346 ms window (and their standard deviations) are shown in Table 4. The amplitude enhancement of the DEV stimuli AEPs compared to the STD stimuli ERPs, around 180 ms post-deviance onset and identified as the MMN, was statistically significant in each intonational contrast (lower [within-category] contrast, t_{14} = –6.217, p < .0005; central [across-category] contrast, t_{14} = –8.875, $p < 10^{-6}$; higher [within-category] contrast, t_{14} = –6.551, p < .0005). A repeated measures ANOVA on the MMN peak latencies did not yield any difference between the three conditions ($F_{2, 28}$ = 2.828, p = n.s., η^2 = .168). As we hypothesized, the mean amplitude of the MMN was larger for the central [across-category] intonational contrast (steps 05 - 10) compared to the within-category contrasts: Intonational contrast effect at Fz, $F_{2, 28}$ = 3.417, p < .05, η^2 = .196

(within subject contrasts: lower vs. central, $F_{1,\ 14}$ = 6.256, p < .05, η^2 = .309; central vs. higher, $F_{1,\ 14}$ = 4.898, p < .05, η^2 = .259; lower vs. higher , $F_{1,\ 14}$ = 0.172, p = n.s., η^2 = .012). The analysis at the Mastoid electrodes yielded similar results to those obtained at Fz: $F_{2,\ 28}$ = 6.978, ε = .679, p = .01, η^2 = .333 (within subject contrasts: lower vs. central, $F_{1,\ 14}$ = 43.403, p < .00001, η^2= .756; central vs. higher, $F_{1,\ 14}$ = 4.323, p = .056, η^2 = .236; lower vs. higher, $F_{1,\ 14}$ = 1.203, p = n.s., η^2 = .079). The scalp distribution maps of the MMN are shown in Figure 12.

Figure 10. Grand-average waveforms elicited to STD and DEV stimuli and their difference waves. The first row (in red) represents the lower [within-category] contrast, the second row (in green) represents the central [across-category] contrast, and the third row (in blue) represents de higher [within-category] contrast. In each plot, STD and DEV responses are represented by colored lines, STD with dotted lines and DEV with continuous lines. Also, DEV minus STD stimuli difference waveforms are plotted in black. Columns indicate the measures at Fz, M1, and M2 (left, center and right columns, respectively).

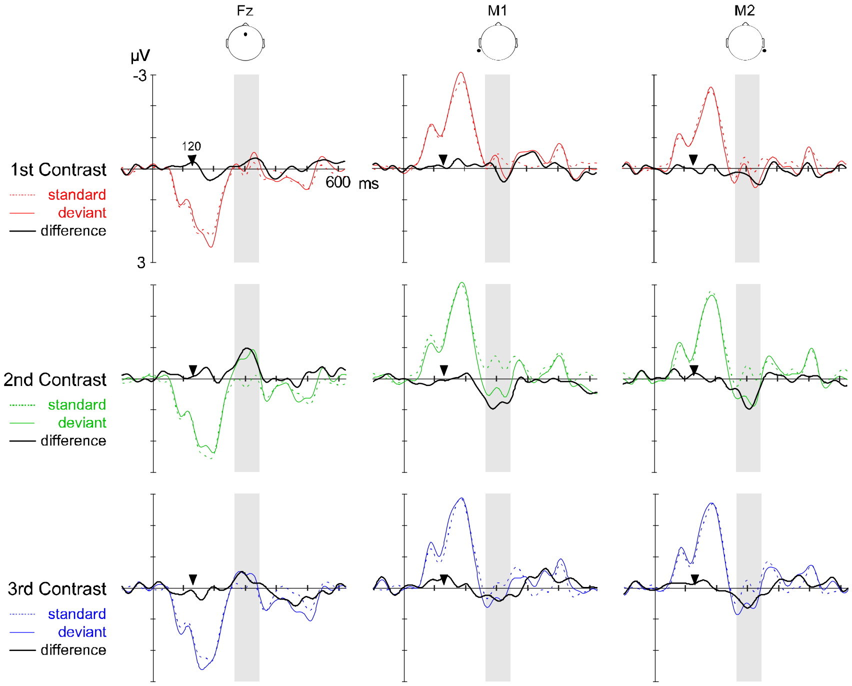


Figure 11. DEV minus STD stimuli difference waves of each contrast, measured at Fz, M1 and M2 electrodes (left, center and right columns, respectively). MMN processes are observed at frontocentral electrodes (Fz) as negative deflections of the ERP, and at mastoid electrodes as positive deflections, as MMN inverts polarity below the Silvian fissure when the reference electrode is placed on the tip of the nose (Näätänen & Michie 1979).

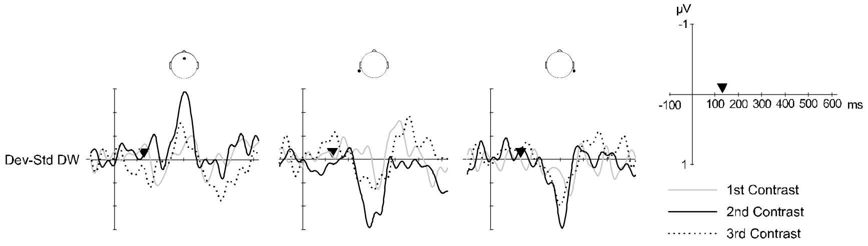
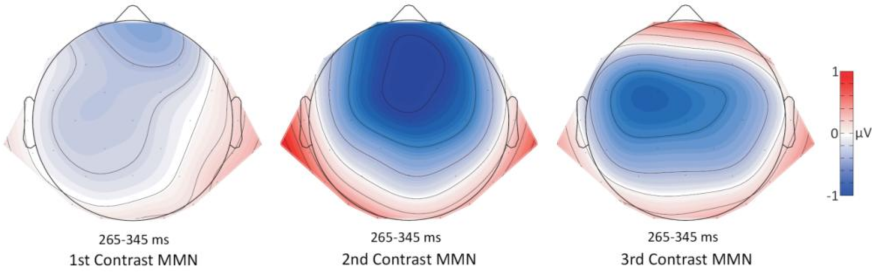


Table 4. Mean MMN amplitudes and their standard deviations for the three experimental contrasts (lower [within-category], central [across-category], and higher [within-category]).

	Mean (Std. Deviation)		
Contrast	FZ	M1	M2
lower (00-05)	-.21 (.726)	.17 (.584)	.33 (.603)
central (05-10)	-.73 (.474)	.96 (.606)	.73 (.396)
higher (10-15)	-.31 (.765)	.38 (.875)	.52 (.671)

Figure 12. Scalp potential distribution maps at the MMN time window extracted from the DEV minus STD difference waves (265-345 ms).

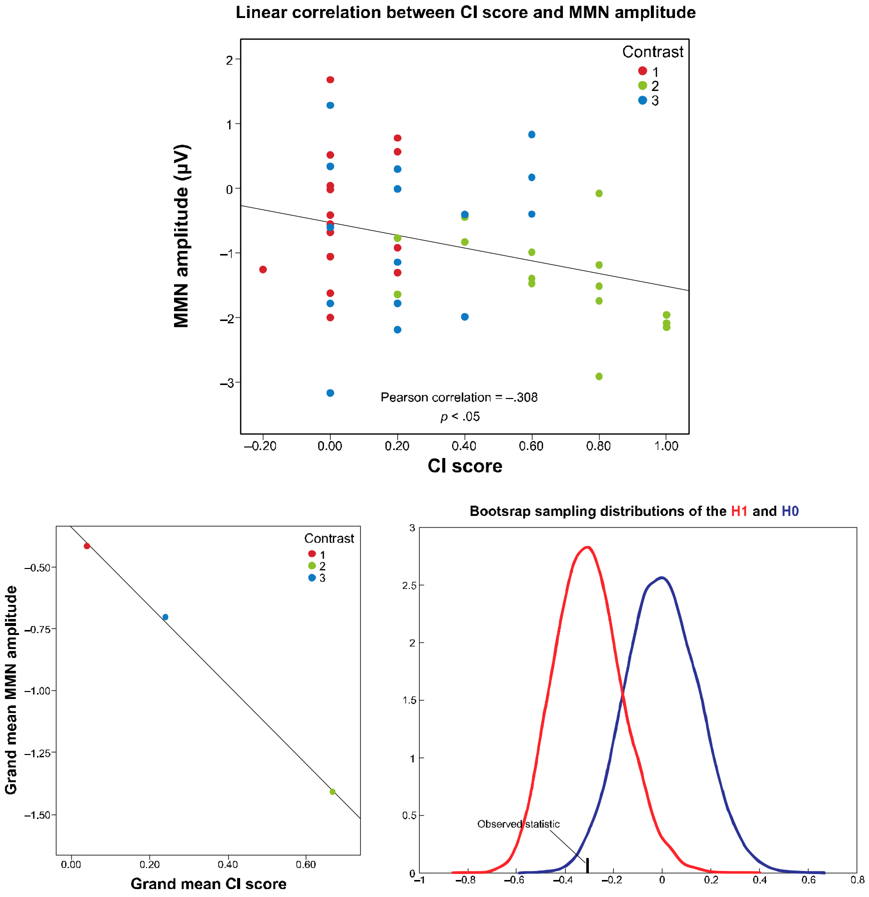


Furthermore, an analysis between the CI and the MMN mean amplitude (electrophysiological measure) yielded a significant negative correlation: Pearson's correlation statistic = –.308; *p* < .05 (one-tailed). This means that the higher the amplitude of the MMN elicited in an oddball sequence with that pair of stimuli acting as DEV and STD stimuli, the more a subject categorized differently the two stimuli within a pair. The significance of this correlation was further supported by an analysis using the bootstrap method: Pearson's correlation statistic sampling distribution centered at –.308; confidence interval of the null hypothesis with 95% confidence bounds, [–.289, .297]; *p* = .018. Additionally, we performed a bivariate correlation between the grand mean of the CI and the grand mean of the MMN, yielding a significant Pearson's correlation of –.999; *p* = .011. We acknowledge that the statistics on the grand mean cannot be taken as a real proof of the existence of a correlation between the CI and the MMN; however, it serves us to illustrate more clearly the direction of the effects. Bivariate correlations between CI and MMN for all subjects and grand means respectively, and the bootstrap sampling distributions of the alternative and null hypotheses can be seen in Figure 13.

3.4. Discussion

Previous electrophysiological studies on vocalic and consonantal phonological contrasts have found evidence that native linguistic contrasts elicit significantly larger MMN responses than non-native contrasts (Näätänen et al. 1997, Winkler et al. 1999, Eulitz & Lahiri 2004). In addition, acoustic contrasts that cross a category boundary lead to larger MMN responses than comparable acoustic contrasts that did not cross these category boundaries (Dehaene-Lambertz 1997, Sharma & Dorman 2000; Phillips et al. 2000). Similarly, it is an established result that tone contrasts in tonal languages obtain larger MMN responses when listeners are

Figure 13. Bivariate correlations between CI and MMN, for all subjects (top) and grand means (botoom-left), and the bootstrap sampling distributions of the alternative and null hypotheses (bottom-right).



exposed to native tonal contrasts (Gandour et al. 1994, Klein et al. 2001, Chandrasekaran et al. 2009, Ren et al. 2009) and also in tonal stimuli crossing the category boundaries (Chandrasekaran et al. 2007, Xi et al. 2010). Thus a substantial set of empirical results demonstrate the larger activation of memory traces for linguistic elements in the human brain. In line with this, Näätänen (2001) proposed that the MMN reflects the early access to stored linguistic representations. In the recent years, more evidence has been accumulating that MMN reflects the early access of linguistic information, reflecting early automatic processes of lexical access and selection, semantic information processing and syntactic analysis (see Pulvermüller & Shtyrov 2006 for a review). Yet previous electrophysiological results on the representation of phonological contrasts at the level of intonation are still controversial. Doherty et al. (2004) and Leitman et al. (2009) argued that the large MMN elicited only by interrogative stimuli (and not by the declarative stimuli) "may underlie the ability of questions to automatically capture attention even when the preceding declarative information has been ignored" (Leitman et al. 2009: 289). Fournier et al. (2010) argued that the recognition of discourse meanings by means of intonation was not necessarily clear by looking at the human brain.

Our results go beyond the body of evidence presented by previous experiments and provide electrophysiological evidence that phonological contrasts at the intonational level (based on a pitch range difference) are encoded in the auditory cortex. The empirical data in our study was based on an intonational contrast between statements and questions in Catalan. The results of Experiment 1, which tested the participants' interpretation of isolated stimuli in a binary way (statement vs. counterexpectational question), corroborated the findings presented in Chapter 2 by indicating a clear nonmonotonic identification. Specifically, a perceptual shift from one category to another occurred in the range of stimuli 6 to 11, with a full crossover from 16.92% to 85.92% achieved between these five central steps. Moreover, post-hoc tests revealed two main homogeneous subsets, namely between stimulus 0-6 and 11-15. Concerning reaction times, listeners displayed faster RTs in identification of within-category exemplars than in exemplars representing the category boundaries (specially for stimuli 7 to 9).

For Experiment 2, four auditory stimuli were selected to be contrasted by pairs in three different oddball blocks. Though the physical distance between each pair of stimuli was kept constant, the central pair represented an across-category contrast whereas the other pairs represented within-category contrasts. The mean amplitude of the MMN was found to be larger for the across-category contrast compared to the other contrasts, suggesting that intonational contrasts in the target language can be encoded automatically in the auditory cortex. Moreover, our results showed that the activation of these auditory cortex intonational representations was related to the individuals' subjective perception and performance. As Pulvermüller & Shtyrov (2006) proposed, the MMN might reflect not only the automatic detection of a change, but also the activation of a certain symbolic memory trace in the brain. Finding a MMN for within-category contrasts would indicate that a change in the acoustic environment has been detected, but the symbolic memory trace is still the same called by the standard. By contrast, finding a significantly larger MMN in an across-category contrast would thus not only indicate a reactivation of the attentional system, but also an activation of different cortical cell assemblies supporting another long-term memory trace.

It is also important to note that our data can also support an alternative explanation, i.e., that the MMN results may reflect perceptual saliencies or distinctiveness that may be consistent across languages. While external evidence suggests that the MMN may reflect symbolic memory traces, others have suggested that the MMN robustness may reflect individual differences in dimensional weighting (e.g. Chandrasekaran et al. 2007, Chandrasekaran et al. 2009). For example, animals show categorical perception (Kuhl & Miller 1978), and thus the increased MMN for across-category contrasts may reflect auditory discontinuities (e.g. Holt et al. 2004, for voice onset time), i.e., natural boundaries within which distinctiveness is enhanced, reflecting a warped acoustic space (Kuhl & Miller 1975). One possibility for demonstrating the explanation based on symbolic memory traces would be the application of a cross-language design, but this should be addressed in future studies.

The present experiment design does not allow us to draw any conclusions regarding the specific neural network supporting the across-category intonation contrasts observed here as enhanced MMNs and therefore we can only speculate. The MMN has multiple cerebral sources, including the auditory cortex (Alho 1995, Escera et al. 2000) and frontal regions (Deouell 2007), and recent results from animal (Ulanovsky et al. 2003, Pérez-González et al. 2005, Malmierca et al. 2009, Antunes et al. 2010) and human studies (Grimm et al. 2011, Slabu et al. 2010) have suggested that deviance detection yielding to MMN generation might encompass the whole auditory hierarchy (Grimm & Escera 2012). Moreover, recent studies have suggested that processing linguistic deviant features recruits not only auditory but also motor cortical regions in a somatotopic fashion (Hauk et al. 2006, Shtyrov et al. 2004), and that category-based enhancement is often found in prefrontal regions (Freedman et al. 2001). In addition, Raizada & Poldrack (2007) found that lower-level auditory areas show little enhancement of across-category phonetic pairs relative to higher order areas, and Zhang et al. (2011) have shown that across-category variation on a lexical tonal continuum activated the left middle temporal gyrus, apparently reflecting abstract phonological representations, whereas the within-category contrasts activated the superior temporal and Heschl gyri bilaterally. Therefore, it is possible that the cross-category intonational effects observed here as a frontally distributed enhanced MMN, compared to the within category one, might reflect the activation of a distributed cortical network including higher-order auditory areas, such as the posterior superior temporal gyrus and the middle temporal gyrus, and frontal regions.

In sum, the MMN findings reported in this chapter show that a distributed auditory-frontal cortical network supports not only phonological representations at the segmental level but also at the intonational level. Catalan listeners showed a larger MMN response to differences in pitch activating the semantic contrast between a question and a statement. To our knowledge, this is the first study showing a clear electrophysiological response to a change of intonational category. This result agrees with Pulvermüller & Shtyrov's (2006) hypothesis that MMN responses reflect early automatic processes not only affecting lexical access and selection, but also semantic and discourse information processing.

CHAPTER 4

The role of facial gestures in establishing discourse categories

4.1. Introduction

The strong influence of visual cues upon speech perception in normal verbal communication has increasingly been recognized. Audiovisual speech studies have revealed that the visual component plays an important role in various aspects of communication typically associated with verbal prosody. The visual correlates of prominence and focus (movements such as eyebrow flashes, head nods, and beat gestures) boost the perception of these elements (Cavé et al. 1996, Hadar et al. 1983, Krahmer & Swerts 2007, Swerts & Krahmer 2008, Dohen & Lœvenbruck 2009). Similarly, audiovisual cues for prosodic functions such as face-to-face grounding (Nakano et al. 2003) and question intonation (Srinivasan & Massaro 2003) have been successfully investigated, as have the audiovisual expressions of affective meanings such as uncertainty (Krahmer & Swerts 2005) and frustration (Barkhuysen et al. 2005).

In the last few decades, an important research topic in the field of audiovisualprosody has been the relative importance of facial cues with respect to auditory cues for signaling communicatively relevant information. A large number of studies have described a correlated mode of processing, whereby vision partially duplicates acoustic information and helps in the decoding process. For example, it is well known that visual information provides a powerful assist in decoding speech in noisy environments, particularly for the hearing impaired (Sumby & Pollack 1954, Breeuwer & Plomp 1984, Massaro 1987, Summerfield 1992, Grant & Walden 1996, Grant et al. 1998, Assmann & Summerfield 2004).

Another set of studies has found a weak visual effect relative to a robustly strong auditory effect. For example, it has been found that observers extract more cue value from auditory features when it comes to marking prominent information in an utterance (Scarborough et al. 2009). Krahmer et al. (2002) found that people pay much more attention to auditory than to the eyebrow information when they have to determine which word in an utterance represents new information, and other follow-up studies confirmed the relatively weak cue value of these visual features, yet at the same time provided evidence that visual cues do have some perceptual importance (given that a visual-cue-only identification task yielded 92.4% correct guesses; see Krahmer & Swerts 2004).

Srinivasan and Massaro (2003) showed for English that statements and questions are discriminated both auditorily (on the basis of the F0 contour, amplitude and duration) and visually (based on the eyebrow raise and head tilt), but they also found a much larger influence of the auditory cues than visual cues in this judgment. Their results were consistent with those reported by House (2002) for Swedish, who found that visual cues (consisting of a slow up-down head nod and eyebrow lowering for questions, and a smile throughout the whole utterance, a short up-down head nod and eye narrowing for statements) did not strongly signal interrogative meanings, compared to auditory information like pitch range and peak alignment differences. Dohen and Lœvenbruck (2009) showed that adding vision to audition for perception of prosodic focus in French can both improve focus detection and reduce reaction times. When the experimental paradigm was applied to whispered speech, results showed an enhanced role for visual cues in this type of speech. However, when evaluating the auditory-visual perceptual processes involved in normal speech, they found that auditory-only perception was nearly perfect, which suggests a ceiling effect for visual information. These results were in line with those from Krahmer and Swerts (2004), which showed that prosodic prominence was very well perceived auditorily only for normal speech in Dutch and Italian. In relation to this, fMRI studies have shown that when visual and audio channels share time-varying characteristics this results in a perceptual gain which is realized by subsequent amplification of the signal intensity in the relevant sensory-specific cortices (auditory and visual) (see Calvert & Campbell 2003, Colin et al. 2002).

The abovementioned results could lead to the conclusion that visual information from the face is essentially redundant to auditory information, by using a set of audiovisual properties that can be found in most intonational languages. However, there are a few studies that have found that visual information is crucial in signaling certain types of attitudinal or emotional correlates. Studies like those of Swerts and Krahmer (2005), Dijkstra et al. (2006) and Mehrabian and Ferris (1967) have found that visual information is far more important for communicative purposes than acoustic information. In the first study, Dijkstra et al. (2006) studied speakers' signs of uncertainty about the correctness of their answer when answering factual questions. They noted the use of prosodic cues such as fillers ("uh"), rising intonation contours or marked facial expressions. Results showed that, while all three prosodic factors had a significant influence on the perception results, this effect was by far the largest for facial expressions. Similarly, Swerts and Krahmer (2005) showed that there are clear visual cues for a speaker's uncertainty and that listeners are more capable of estimating their feeling of an interlocutor's uncertainty on the basis of combined auditory and visual information than on the basis of auditory information alone. When visual expressions such as funny faces and eyebrow movements occurred, they seemed to offer a very strong cue for estimating uncertainty.[9] Mehrabian and Ferris (1967) analyzed

[9] Authors refer to uncertainty with the term "feeling of knowing", which is defined as the ability to monitor the accuracy of one's own knowledge or the ability to monitor the feeling of knowing of someone else ("feeling of another's knowing") (see, e.g., Litman & Forbes-Riley 2009).

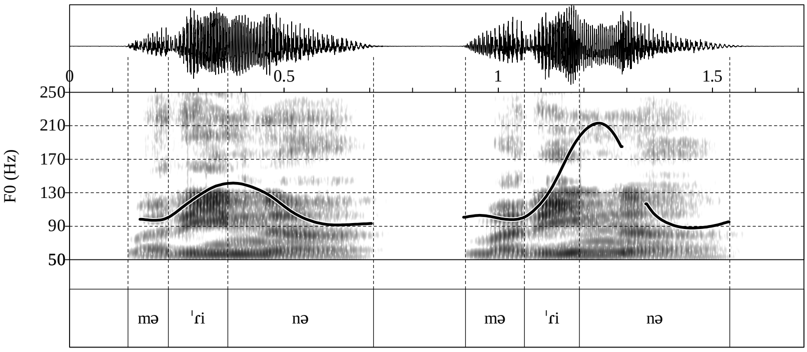
how listeners got their information about a speaker's general attitude in situations where the facial expression, tone of voice and/or words were sending conflicting signals.[10] Three different speakers were instructed to say "maybe" with three different attitudes towards their listener (positive, neutral or negative). Next, photographs of the faces of three female models were taken as they attempted to convey the emotions of like, neutrality and dislike. Test groups were then instructed to listen to the various renditions of the word "maybe," with the pictures of the models, and were asked to rate the attitude of the speakers. Significant effects of facial expression and tone were found such that the study suggested that the combined effect of simultaneous verbal, vocal and facial attitude communications is a weighted sum of their independent effects with the coefficients of .07, .38 and .55, respectively. Nevertheless, these results do not mean that the coefficients derived may not vary greatly depending upon a number of other factors, such as actions, context of the communication and how well the interpreting individual knew the other person (see also Lapakko 1997).

Thus, an overview of the literature reveals that visual cues are potentially useful as markers of prosodic information, yet it is still unclear how important they are compared to auditory cues. In the present chapter, we address this question by analyzing the patterns of prosodic perception of contrastive focus statements vs. counter-expectational questions in a group of Catalan speakers. The main goal of the chapter will be to investigate the relative contribution of visual and pitch accent cues in conveying this specific prosodic distinction in Catalan. In this language, as presented in Chapters 2 and 3, a pitch range difference in a rising-falling nuclear configuration is the main intonational cue for the

[10] The term 'tone of voice' has to be understood in a non-technical way. In this experiment, subjects were asked to listen to a recording of a female saying the single word 'maybe' in three tones of voice conveying liking, neutrality and disliking.

distinction between statements (both information and contrastive focus statements) and counter-expectational questions (see Chapter 2). Figure 14 shows the waveforms and F0 contours of the proper noun *Marina* produced with a CFS meaning (left) and a CEQ meaning (right). In line with this, a L+H* L% nuclear configuration for the expression of contrastive focus statements (CFS) and a L+¡H* L% nuclear configuration for a counter-expectational question (CEQ) (see the Cat_ToBI proposal in Prieto *in press* and Aguilar et al. 2009).

Figure 14. Waveforms and F0 contours of the proper noun *Marina* 'person's name' produced with a CFS meaning (left) and a CEQ meaning (right).



This chapter addresses two related questions regarding the perceptual processing of the audiovisual markers of CFS vs. CEQ meanings in Catalan. First, how important are facial gestural correlates to this distinction with respect to pitch accent cues? Second, are there differences in the relative weight of the acoustic information when facial cues are less prominent and thus more ambiguous? The advantage of using the Catalan distinction between CFS and CEQ meanings is that we will be assessing the relative perceptual importance of a well-known pitch accent contrast in the intonational phonology of Catalan (L+H* for statements and L+¡H* for questions) in conjunction with congruent and incongruent facial gesture information. To our knowledge, no previous studies have examined the bimodal perception of a prosodic contrast by using congruent and incongruent pitch accent and facial cue information. This methodology will allow us to create a very controlled situation where both pitch accent contrasts and visual information are carefully controlled for in a bimodal identification task.

The following sections describe the two experiments that were conducted to address these questions. Experiment 1 tackled the relative contribution of visual and auditory information to the target prosodic contrast by means of an identification experiment. For this task, subjects were presented with two video clips of a person's face as they spoke the word *Petita(?)* 'small' with their expression conveying one or the other of the two target meanings. The visual material was coupled with an audio track selected from a continuum of varying degrees of pitch range for the rising-falling configuration (the main acoustic cue to the distinction between the two meanings). Subjects were thus presented with either congruent or incongruent audio and visual target stimuli. Experiment 2 also investigated the role of auditory and visual information using the same stimuli but this time the continuum of audio cues was combined with a continuum of facial expressions created using a digital image-morphing technique. The task of the participants was again to identify the intended meaning (CFS or CEQ), for each combined audio + visual stimulus.

4.2. Recordings

Little research has been undertaken on the description of gestural patterns in Catalan. Most of the studies have been devoted to the description of Catalan emblems, i.e. specific hand/arm gestures which convey standard meanings that are used as substitutes for words (for example, holding up the hand with all fingers closed except the index and middle finger, which are extended and spread apart, can mean ‘V for victory’ or ‘peace’).[11]

There has been no previous research dealing specifically with the facial gestures that characterize CFS and CFS meanings in Catalan. Thus in order to decide which gestural patterns would be used as target facial expressions in our visual materials, ten native speakers of Catalan between the ages of 20 and 47 were videotaped pronouncing both possible interpretations of the utterance. Two of the ten speakers were the authors, and the other eight were graduate students and professors, with no previous experience in audiovisual research. In order to prompt the corresponding answer, subjects were asked to read in an expressive way the two dialogues in (4), with dialogue (4a) involving CFS and dialogue (4b) exemplifying a CEQ. As is well known, in this type of echo questions, the listener repeats information that s/he has just heard, and these questions are sometimes marked by a nuance of surprise or incredulity. Subjects were given no instructions as to how to express these pragmatic meanings in audiovisual prosody. The audiovisual recordings of all ten speakers were carried out in quiet research rooms at the Universitat Autònoma de Barcelona and the Universitat Pompeu Fabra. Speakers were seated on a chair in front of a digital camera that recorded their upper body and face at 25 frames per second.

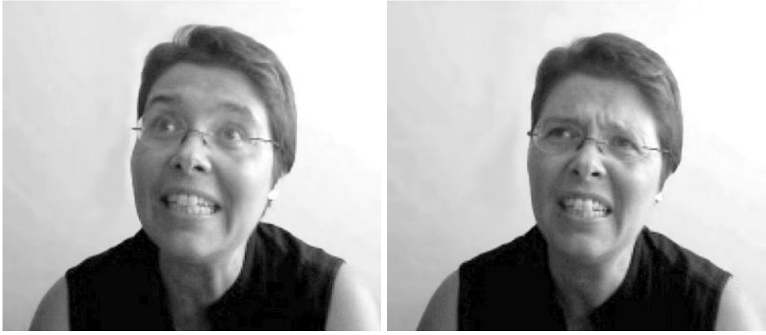
(4) a. Volies una cullera gran, no? *You wanted a big spoon, didn’t you?*
Petita, [la vull, i no gran]. *[I want a] little [one, not a big one].*
b. Jo la vull petita, la cullera. *I want a little spoon.*
Petita? [N’estàs segur?] *[A] little [one]? [Are you sure?]*

[11] Of particular note is the work by Amades (1957), Mascaró (1978, 1981) and especially Payrató (1989, 1993), which contains a description of a repertoire of 221 emblems and pseudoemblems of Central Catalan. Since the 1990s, two projects lead by Lluís Payrató and financed by the varcom and pragmaestil have analyzed the system of Catalan gestures but have mainly focused on coverbal manual gestures (see e.g. Payrató et al. 2004).

From these twenty visual tokens (ten for each pragmatic meaning), the authors assessed qualitatively the facial gesture correlates that were most effective and representative for each pragmatic meaning. One of the facial expressions that correlate most clearly with the perception of CFS is the upward eyebrow movement and forward head movement. For a CEQ, the facial expression is characterized by a furrowing of the brows and a squinting of the eyes, often accompanied by a head shake. Figure 15 shows two representative stills of the facial expression as one of our speakers spoke a CFS (left panel) and a CEQ (right panel). For describing the facial gestures, we have used the Facial Action Coding System (FACS), developed by Paul Ekman and his colleagues, which allows coding of all visually distinguishable facial expressions (Ekman & Friesen 1978, Ekman et al. 2002). FACS groups muscle activity into so-called Action Units (AUs) that bundle uniquely identifiable facial movements, the articulatory basis of these movements can thus be the activity of one or multiple muscles. Three AUs are relevant in the production of eyebrow movements (see also De Vos et al. 2009): AU 1, the Inner Brow Raiser; AU 2, the Outer Brow Raiser; and AU 4, the Brow Lowerer. For CFS interpretations, the most common facial expression consisted of a combination of action units AU1+2 (Inner and Outer Brow Raisers) and M57 (Head Forward). For CEQ interpretation, the most common pattern was a combination of AU4 (Brow Lowerer) and M58 (Head Backward).[12]

[12] Please note that there is a noticeable lip stretching in the case of the CFS gesture. It is interesting to point that the gestural overarticulation of the segments in accented position (in our case, the vowel /i/) is a common phenomenon among the production of CFS (as described by Dohen & Lœvenbruck 2009, Prieto et al. 2011, and Borràs-Comes et al. 2011). In fact, this specific aspect lead us to compare CEQ with CFS, as both categories are produced in face-to-face communication with a specific facial configuration (which is not the case of their nonbiased counterparts, i.e., IFS and ISQ).

Figure 15. Representative stills of a facial expression of one of our speakers while producing a CFS (left panel) and a CEQ (right panel).



From the results of the production test it was thus clear that one of the most effective gestural cues for the distinction between CFS and CEQ was the pattern of eyebrow movements. A number of crosslinguistic studies have shown that eyebrow movements combine with facial gestures (Beskow et al. 2006, Cavé et al. 1996, Graf et al. 2002, Scarborough et al. 2009, Armstrong 2012) or head movements (Beskow et al. 2006, Graf et al. 2002, Hadar et al. 1983, Scarborough et al. 2009, Munhall et al. 2004) to express prosodic focus. For instance, it has been found that focus production is accompanied by eyebrow raising and/or a head nod (Krahmer & Swerts 2004 for Dutch, Dohen et al. 2006 for French).

It is also interesting to note that in sign languages, eyebrow movements serve various grammatical functions. For example, eyebrows are furrowed in *wh*-questions and raised in yes/no questions in American Sign Language (Baker-Shenk 1983, Grossman 2001, Grossman & Kegl 2006), Swedish Sign Language (Bergman 1984), British Sign Language (Kyle & Woll 1985) and Sign Language of the Netherlands (Coerts 1992) — see Pfau and Quer (2010) for a review.

The prosodic information obtained in this set of audiovisual recordings was used as a basis for the preparation of audiovisual stimuli for use in our two perception experiments. While the acoustic information was almost identical in the two experiments (a set of either 11 or 6 pitch range differences created with PSOLA manipulation), the visual information was different, in that we used two unmanipulated video recordings for the contrast for Experiment 1 but used six videos in Experiment 2, with four of these clips being digitally-generated interpolations between part of the two used in Experiment 1.

4.3. Experiment 1

4.3.1. Methodology

The first experiment tested the role of auditory and visual information in pragmatic identification of CFS and CEQ by means of an auditory continuum of pitch range which was combined with two video clips depicting the facial gestures characteristic of the two pragmatic meanings in such a way that the audio cue might be congruent or incongruent to a greater or lesser degree with the visual cue.

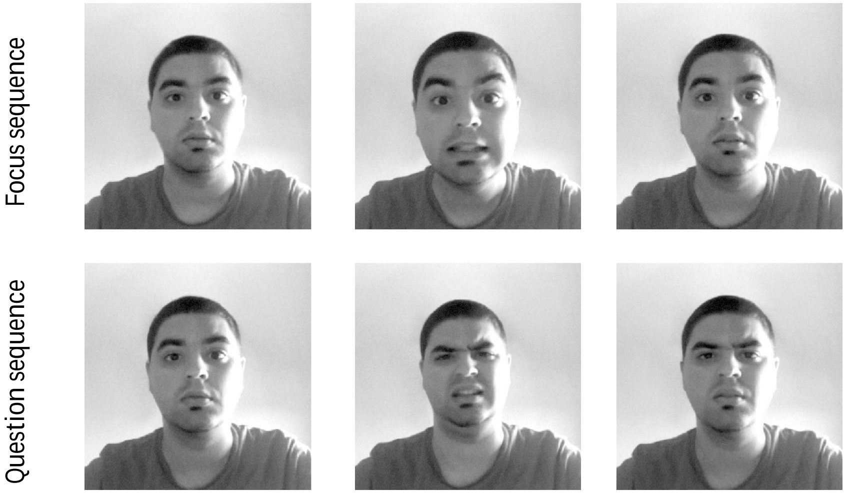
Participants

A total of twenty native speakers of Central Catalan participated in the experiment. The ages of the participants ranged from 18 to 36. All of them were undergraduate or graduate students with no previous experience in audiovisual research.

Materials

To make sure that participants in our experiments could focus as much as possible on the audiovisual correlates of the two target pragmatic meanings, we selected a very short utterance that would contain the target intonational cues and facial gestures. To generate the audiovisual stimuli for the experiment, a native speaker of Catalan was videotaped several times producing natural productions of the noun phrase *petita* [pə.ˈti.tə] (‘small’-fem) with either a CFS contour or a CEQ contour. The author tried to imitate the two gestural patterns selected from among our preliminary video recordings as representative of the CFS and CEQ meanings. The two authors of the original paper then selected the two exemplars that best characterized the contrast, while at the same time making sure that syllabic durations were similar in the two recordings. Figure 16 shows three representative stills from the video clips as the subject utters first a CFS (upper panels) and then a CEQ (lower panels). The three images in each set correspond to three different stages of the facial gesture: initial expression (left), central expression (centre; approximately coinciding with the beginning of the stressed syllable) and final expression (right).

Figure 16. Stills from video clips depicting facial gestures during the utterance of a CFS (upper panels) and a CEQ (lower panels). The three images correspond to three different stages of the gestures: initial expression (left), central expression (centre) and final expression (right).



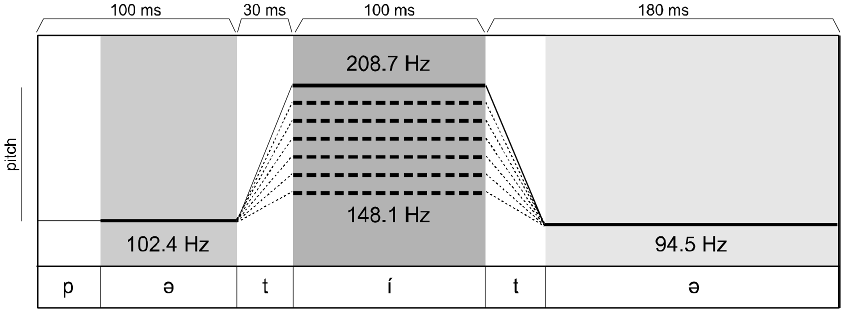
The target utterances were inspected for their prosodic properties. As expected, both target sentences were pronounced with a rising-falling intonational contour (L+H* L%) but differed in pitch range. The observed values for the high tone were 148.1 Hz for the CFS example and 208.7 Hz for the CEQ example. As noted above, duration patterns had been controlled for in the original materials. Table 5 shows the duration values of each of the target segments of the utterance *petita* in both readings (CFS and CEQ), revealing very small differences across the two utterances.

Table 5. Original values of the duration (in ms.) of the target segments in the auditory sequence *petita* 'small' and their difference.

	original CFS	original CEQ	difference
p	13	17	4
ə	68	80	13
t	41	39	2
ˈi	116	110	6
t	35	39	3
ə	116	124	8
Sum	389	409	

To prepare the target auditory stimuli for the experiments, we chose one of the two auditory recordings (the CEQ) and manipulated the pitch by means of Praat (Boersma & Weenink 2008). A synthesized continuum was created by modifying the F0 peak height in 11 steps (distance between each one = 0.6 semitones). The pitch values corresponding to the accented syllable of the word *petita* were manipulated so that they would be realized as a 110 ms plateau starting 39 ms after the onset of the accented syllable /'ti/, and were preceded by a low plateau for the syllable [pə] (102.4 Hz, 97 ms). The posttonic syllable [tə] was produced with a low plateau (94.5 Hz, 163 ms). A schematic diagram of these manipulations is shown in Figure 17.

Figure 17. Schematic diagram with the pitch target manipulation.



Each one of the auditory steps was then combined with the two target visual stimuli (see Figure 16), for a total of 22 target audiovisual stimuli. Since the video materials were recorded at 25 frames per second and the observed differences between natural auditory stimuli never surpassed 40 ms., no visual manipulations were needed to prepare the final audiovisual stimuli. An informal inspection of the data did not reveal cases of undesired lip-sync problems and visually the manipulated stimuli appeared natural. To confirm these impressions, we asked a panel of two independent judges to check all the stimuli in terms of whether they felt that either auditory or visual signals lagged behind, or instead appeared perfect synchronized. This additional check did not reveal any problematic cases of audiovisual mismatches.

Procedure

Experiment 1 consisted of 5 blocks in which all 22 stimuli were presented to the subjects in a randomized order. A brief training session was conducted prior to the task in order to get subjects accustomed to the stimuli and the task. In this session, subjects were shown two repetitions of the fully congruent and fully incongruent audio + visual combinations.

Stimuli were presented to subjects using a laptop computer equipped with headphones. Subjects were instructed to pay attention to the auditory stimuli and facial gestures as a whole and decide which interpretation was more likely for each stimulus by pressing the corresponding computer key, “0” for CFS and “1” for CEQ.

The experiment was set up by means of E-Prime version 2.0 (Psychology Software Tools Inc. 2009), which allowed us to record response frequencies automatically. A timer with 1 ms accuracy was activated at the beginning of each stimulus, and the time that elapsed from the beginning of each playback to the striking of a response key was recorded, thus giving reaction time (RT) measurements. Subjects were instructed to press one of the two computer keys as quickly as they could. The experiment was set up in such a way that the next stimulus was presented only after a response had been given.

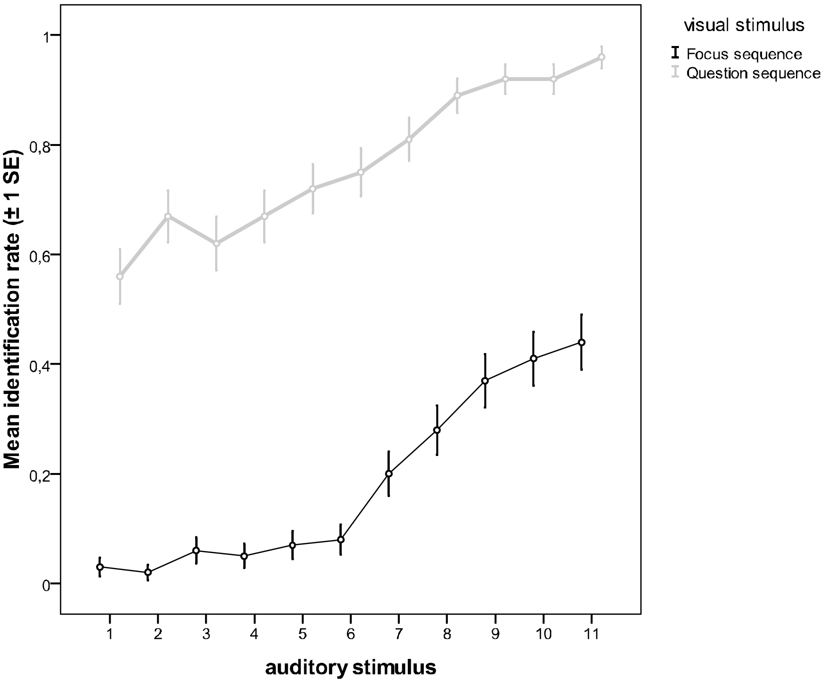
The experiment was set up in a quiet research room at the Universitat Pompeu Fabra. We obtained a total of 2,200 responses (11 auditory steps × 2 visual sequences × 5 blocks × 20 listeners). The experiment lasted approximately 8 minutes.

4.3.2. Results

Identification responses

The graph in Figure 18 shows the mean “CEQ” identification rate as a function of video stimulus (solid black line = CFS video; solid gray line = CEQ video) and auditory stimulus (x-axis), for the 20 subjects. The graph reveals that subjects mostly decided on the interrogativity of the utterance by relying on the visual materials, as the CEQ video and the CFS video responses are clearly separated in the graph (the CEQ video elicited from 56% to 96% of “CEQ” identification responses and the CFS video elicited from 3% to 45% “CEQ” identifications). Interestingly, there is also a clear effect of the auditory information but it is less robust: the preference for identifications is stronger for congruent audio + visual combinations (that is, a CEQ video combined with a CEQ pitch contour obtains a 96% of “CEQ” responses, and a CFS video combined with a CFS pitch contour obtains a 3% of “CEQ” responses). By contrast, most confusion arises in cases where the auditory cue is incongruent with the visual cue (that is, a CEQ video with a CFS audio track, or a CFS video with CEQ audio track). In other words, the congruent stimuli reveal more accurate responses than the incongruent ones. The clear congruity effects can be interpreted as evidence for a bimodal integration process.

Figure 18. Mean “CEQ” identification rate as a function of video stimulus (solid black line = CFS video; solid gray line = CEQ video) and auditory stimulus (x-axis), for the 20 listeners. Error bars show ± 1 Standard Error. In the x-axis, stimulus 1 is a CFS and stimulus 11 is a CEQ.



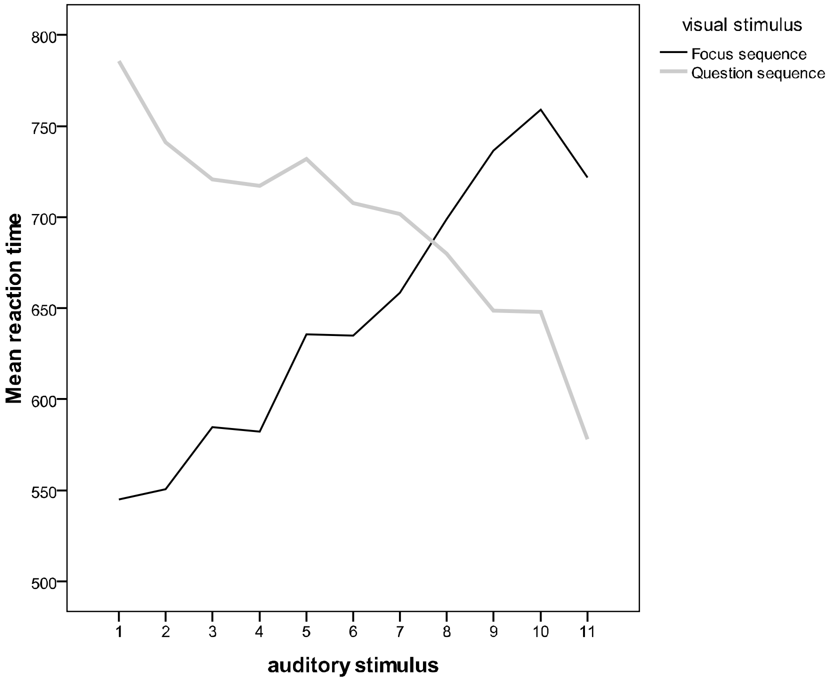
A two-factor ANOVA with a 2 × 11 design was carried out with the following within-subjects independent factors: visual stimulus (two levels: CFS, CEQ) and audio stimulus (eleven levels: 11 steps in the pitch range). The dependent variable was the proportion of “CEQ” responses. The data were first checked for the occurrence of possible outliers on the basis of reaction time. Of a total of 2200 datapoints, 193 cases were treated as outliers, i.e. those cases where the reaction times were at a distance of at least three standard deviations from the overall mean. These cases were excluded from the analysis.

The analysis revealed a significant main effect of visual stimulus ($F_{1, 2007}$ = 1306.798, $p < .001$) and of auditory stimulus ($F_{10, 2007}$ = 31.119, $p < .001$) on statement/question identification. The interaction between the two factors was not significant ($F_{10, 2007}$ = 1.059, $p = .391$), meaning that the effects of both factors are consistent across factor groups. Thus we can observe a clear preference for visual cues in the listener's main decisions, but also a crucial effect of the auditory stimuli.

Reaction times

Figure 19 shows mean reaction times (in ms) as a function of video stimulus (solid black line = CFS video; solid gray line = CEQ video) and auditory stimulus (1 = CFS contour; 11 = CEQ contour), for the 20 listeners. In general, mean RT patterns show that congruent audiovisual stimuli differ significantly from incongruent ones in that the latter trigger consistently slower reaction times. That is, when a CEQ-based visual stimulus occurred with a low-pitched auditory stimulus, this triggered an important time delay in the response (mean RT: 786 ms). This is also the case when CFS-based visual stimuli occurred with high-pitch auditory stimuli (mean RT: 722 ms). By contrast, congruent audio + visual combinations triggered very fast responses, namely in the combinations of a CEQ video with the highest peak (mean RT: 578 ms) and of a CFS video with the lowest peak (mean RT: 545 ms).

Figure 19. Mean reaction times in ms as a function of video stimulus (solid black line = CFS video; solid gray line = CEQ video) and auditory stimulus (1 = CFS contour; 11 = CEQ contour), for the 20 listeners.



To get a first insight into the patterns of the reaction times, we conducted a *t*-test which compared averages for congruent and incongruent stimuli. Thus, for this test, we combined the two conditions for the extreme congruent stimuli (CFS video with auditory stimulus 1 & CEQ video with auditory stimulus 11) and paired those with that for the most incongruent stimuli (CFS video with auditory stimulus 11 & CEQ video with auditory stimulus 1). This *t*-test revealed that congruent stimuli differed significantly from incongruent ones in that the latter yielded consistently slower reaction times (congruent: 670 ms; incongruent: 979 ms) (t_{183} = –3.619, p < .001).

A two-factor ANOVA was carried out on the results. The dependent variable was reaction time measures. The within-subject independent variables were the visual stimulus (two levels:

CFS, CEQ) and the auditory stimuli (eleven steps in the pitch range). The analysis revealed a clear effect of the visual factor for reaction times ($F_{1,\ 2173}$ = 6.362, p = .012), and no effect for the auditory stimuli ($F_{10,\ 2173}$ = .671, p = .752). The interaction between the two factors was statistically significant ($F_{10,\ 2173}$ = 2.815, p = .002). Thus we clearly observe a preference for visual cues in the listener's main decisions, but also a crucial interaction between the visual and auditory information.

4.4. Experiment 2

4.4.1. Methodology

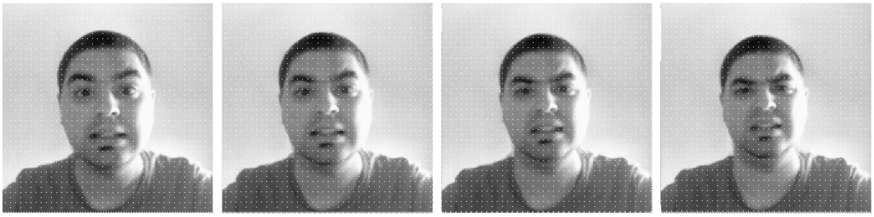
Experiment 2 analyzed the identification of CFS and CEQ by means of the same auditory continuum used in Experiment 1 but this time in combination with a continuum of facial gestures produced using a digital image-morphing technique. The goal of this experiment was to test whether the creation of intermediate steps in facial gestures would affect the interpretation of the stimulus materials and how this gradient visual information would interact with the processing of the auditory information.

Materials

To produce the target visual materials for Experiment 2, four static images were extracted from the target recordings used in Experiment 1, namely the first one for the initial neutral facial gesture, the second at the beginning of the stressed syllable, the third at the beginning of the post-tonic syllable and the last one at the end of the utterance (see Figure 16 above, which illustrates the first, second and fourth moments in time for each gesture pattern). Then, a face morphing technique was applied to the second, third and fourth stills selected (since the first one represented a neutral facial gesture; see Figure 16) in order to create four intermediate videos in between the two original video clips. The morphing was performed by means of Sothink SWF

Quicker version 3.0 software (SourceTec Software Co. 2007). With this technique, one can morph one face into another by marking key points on the first face, such as the contour of the nose or location of an eye, and mark where these same points are located on the second face. The program will then create an intermediate frame between the first and second face. The drawings between the key frames are called inbetweens. Once we had the four inbetweens for each moment in time, we concatenated each set of key frames or inbetweens and synchronized them with the auditory materials. Figure 20 illustrates the 4 inbetweens resulting from the face morph manipulation from the CFS gesture pattern (left) to the CEQ gesture pattern (right). The total number of target visual stimuli was six.

Figure 20. Inbetween frames resulting from the digital morphing of the central facial expression between the CFS gesture sequence (left) to the CEQ gesture sequence (right).



(visual stimulus 2) (visual stimulus 3) (visual stimulus 4) (visual stimulus 5)

The duration of this experiment was longer because the auditory materials had to be combined with the set of six video stimuli (instead of the two videos in Experiment 1). Because of this, we selected a subset of the auditory continuum used for Experiment 1, specifically, stimuli numbers 1-3-5-7-9-11 (the distance between each peak height thus becoming 1.2 semitones rather than 0.6). As in Experiment 1, each auditory stimulus was combined with each visual stimulus (6 videotapes), for a total of 36 target stimuli.

Procedure

Experiment 2 consisted of 5 blocks in which all stimuli (36 in total) were presented to the subjects in a randomized order. Again, a brief training session was conducted prior to the task, in which participants were shown two repetitions of the most congruent and incongruent audio + visual stimuli.

The conditions for Experiment 2 and the instructions for subjects were the same as for Experiment 1, and the same group of twenty native Catalan speakers participated. We obtained a total of 3,600 responses (6 auditory steps × 6 visual sequences × 5 blocks × 20 listeners). The order of the two tasks was counterbalanced. The experiment lasted approximately 10 minutes.

4.4.2. Results

Identification responses

Figure 21 shows the mean “CEQ” identification rate as a function of video stimulus (different types of lines, ranging from the solid black line = CFS video to the solid gray line = CEQ video) and auditory stimulus (x-axis), for the 20 listeners. The graph reveals a very similar pattern of responses to that obtained in Experiment 1. First, it is clear that the visual materials were crucial in the participants’ decision on the interrogativity of the utterance, as again the CEQ video responses and the CFS video responses are clearly separated in the graph (the CEQ video elicits from 58.2% to 96% of “CEQ” responses while the CFS video elicits from 1% to 47.5% of “CEQ” responses). Table 6 shows the mean “CEQ” identification rate for each visual stimulus (visual stimulus 1 = CFS video; visual stimulus 6 = CEQ video) when combined with auditory stimuli from both ends of the continuum, i.e. lowest pitch range and highest pitch range.

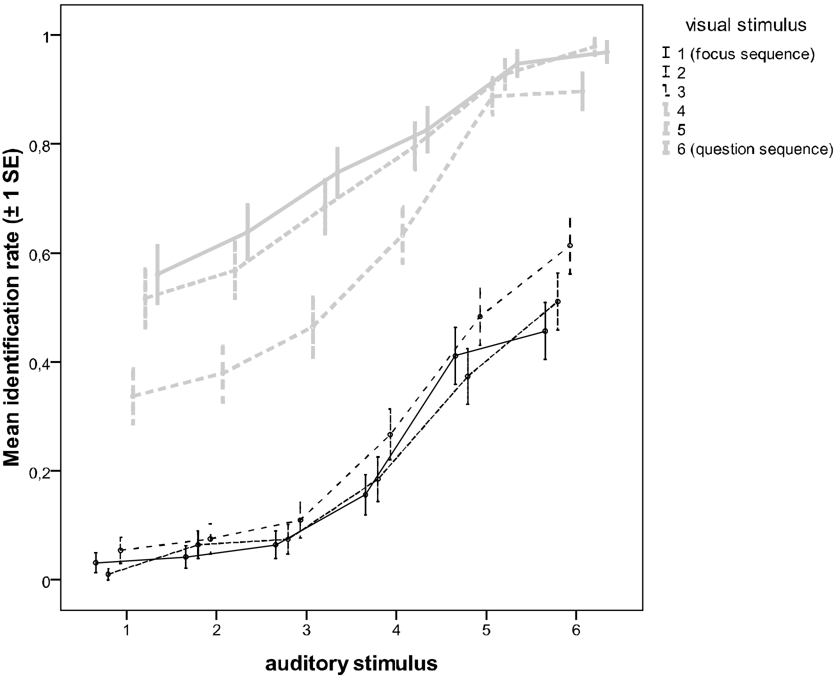
Table 6. Mean 'CEQ' identification rates for each visual stimulus when combined with stimuli from each end of the auditory continuum in Experiment 2

	lowest aud. stim. (CFS)	highest aud. stim. (CEQ)
v1 (CFS)	.010	.475
v2	.030	.515
v3	.050	.592
v4	.340	.888
v5	.536	.970
v6 (CEQ)	.582	.960

Importantly, in all cases we obtain the same effect of the auditory information as in Experiment 1: the preference for interrogativity is stronger for congruent audiovisual combinations (that is, a CEQ video combined with a CEQ pitch contour obtains 96% of "CEQ" responses, and a CFS video combined with a CFS pitch contour obtains 1% of "CEQ" responses). By contrast, most confusion arises in cases where the auditory cue is incongruent with the visual cue.

Interestingly, the tendency to rely on acoustic input is more detectable when the ambiguity of the visual stimulus is more extreme (see Table 6) as can be seen with visual stimulus 4. This elicits 88.8% of "CEQ" responses when the audio cue shows an F0 contour with the highest peak (i.e. when the audio track is indeed a CEQ), and 34% of "CEQ" responses when the F0 contour has the lowest peak (i.e. the audio track is a CFS).

Figure 21. Mean "CEQ" identification rate as a function of video stimulus (different types of lines, ranging from the solid black line = CFS video to the solid gray line = CEQ video) and auditory stimulus (x-axis), for the 20 listeners. In the x-axis, stimulus 1 is a CFS and stimulus 6 is a CEQ.



After completion of the task, several participants reported having seen facial expressions that looked "angry", especially for the most ambiguous visual stimuli. We argue that this collateral identification is an indicator of the ambiguity of the central visual stimuli, which thus increases the effect of the auditory information. In order to compare the curves obtained for the six visual stimuli, we calculated the slope value by means of a logistic regression. This slope value *per se* is not given directly by the function, but the term "b1" is related to the slope, with higher values reflecting shallower curves (Keating 2004). Table 7 shows the b1 value for all tasks. What we can see is that the slope for visual stimulus 4 is the shallowest.

Table 7. b1 values of the logistic regression applied to the six visual stimuli across the six auditory stimuli.

	v1	v2	v3	v4	v5	v6
b1	.482	.418	.489	**.525**	.472	.511

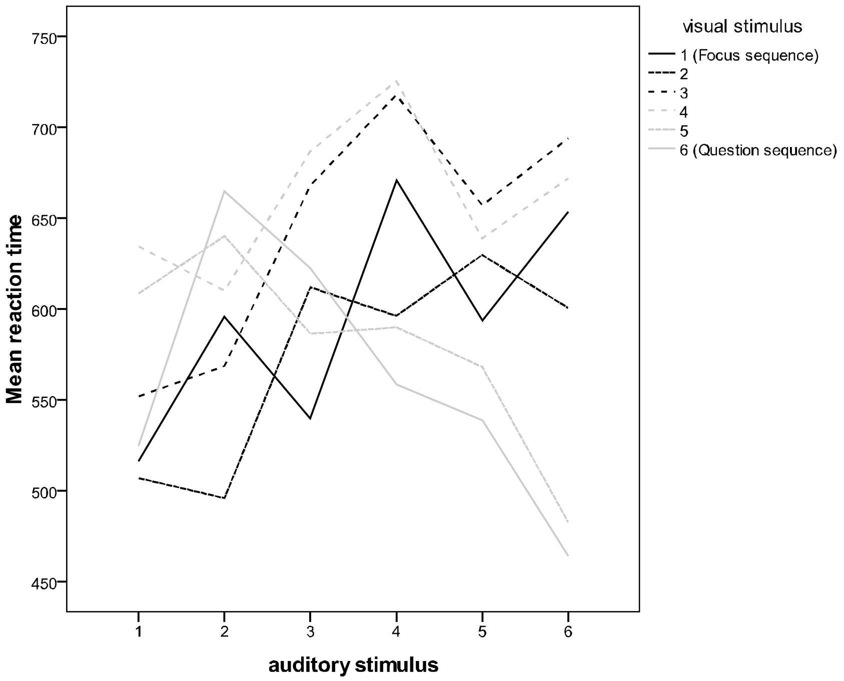
A two-factor ANOVA with a 6 × 6 design was carried out with the following within-subjects independent factors: visual stimulus (six levels: 6 steps from CFS to CEQ) and audio stimulus (six levels: 6 steps in the pitch range). The dependent variable was the proportion of "CEQ" responses. Again, the data were first checked for the occurrence of possible outliers on the basis of reaction time. Of a total of 3600 datapoints, 280 cases were treated as outliers.

Parallel to the results of Experiment 1, the analysis revealed an effect of visual stimulus ($F_{5, 3404}$ = 289.617, $p < .001$) and an effect of auditory stimulus ($F_{5, 3404}$ = 149.821, $p < .001$). However, the interaction between the two factors was not significant ($F_{25, 3404}$ = 1.391, $p = .093$).

Reaction times

Figure 22 shows the mean reaction times (in ms) as a function of video stimulus (different types of lines, ranging from the solid black line = CFS video to the solid gray line = CEQ video) and auditory stimulus (1 = CFS contour; 6 = CEQ contour), for the 20 listeners. Mean RTs patterns show that congruent audiovisual stimuli differ significantly from incongruent ones in that the latter trigger consistently slower reaction times. First, the visual sequences closer to the focus gesture pattern (1 and 2) show an increasing function across the auditory stimuli; second, the visual sequences closer to the question gesture pattern (5 and 6) show a decreasing function across the auditory stimuli;[13] third, the most ambiguous visual stimuli (3 and 4) show larger reaction times when combined with almost all auditory stimuli and an quite increase when the auditory stimuli are more ambiguous. Table 8 shows the mean RT values for each visual stimulus, across all auditory stimuli, when combined with the lowest and highest auditory stimuli.

Figure 22. Mean reaction time measures as a function of video stimulus (black different types of lines, ranging from the solid black line = CFS videotape to the solid gray line = CEQ videotape) and auditory stimulus (1 = CFS contour; 6 = CEQ contour), for the 20 listeners.



[13] As for the specific result in the RT values in the incongruent stimulus audio 1 - video 6, we obtain, as Reviewer 1 points out, an unexpected result of a very low RT. This unexpected value is due to the deletion of the outliers for RT values (the ones that were at a distance of at least three standard deviations from the overall mean), which eliminated very high RT values and lead, in this case, to an unexpected mean RT value.

Table 8. Mean RTs in ms for each visual stimulus (v1-6) across auditory stimuli when combined with auditory stimuli from each end of the continuum.

	mean	lowest aud. stim. (CFS)	highest aud. stim. (CEQ)
v1 (CFS)	712	**604**	779
v2	687	**575**	743
v3	**792**	730	883
v4	**900**	853	925
v5	691	766	**580**
v6 (CEQ)	739	685	**505**

As with the results of Experiment 1, we conducted a *t*-test which compared averages for congruent and incongruent stimuli, the difference being that in this case the auditory stimulus representing the CEQ end of the continuum was stimulus 6 (identical to stimulus 11 in Experiment 1). As in Experiment 1, again, this *t*-test revealed that congruent stimuli differed significantly from incongruent ones in that the latter yielded consistently slower reaction times (congruent: 591 ms; incongruent: 803 ms) ($t_{(180)}$ = –2.194, p = .029).

A two-factor ANOVA was carried out on the results with the dependent variable again reaction time. The within-subject independent variables were visual stimulus (six steps from CFS to CEQ) and audio stimulus (six levels this time, not eleven). The analysis again revealed a clear effect of the visual factor for reaction times ($F_{5,\ 3564}$ = 11.608, p = .012), and no effect for the auditory stimuli ($F_{25,\ 3564}$ = .730, p = .601). The interaction between the two factors was again statistically significant ($F_{25,\ 3564}$ = 1.579, p = .034). Thus, we again observe a main effect of visual cues but also important interaction between the visual and auditory input.

4.5. Discussion

To what extent can gestural cues be crucial in encoding a linguistically relevant contrast such as the perception of statements and questions? This is a question that is still subject to debate among linguists and psycholinguists and has important consequences for models of multimodal language processing. In this chapter, we have explored the relative importance of pitch accent contrasts and facial gestures in the perception of the contrast between contrastive focus statements (CFS) and counter-expectational questions (CEQ) in Catalan, by using congruent and incongruent multimodal stimuli. Our general goal is to understand interaction in the linguistic processing of audio and visual cues during speech perception.

This chapter has presented the results of two perceptual tasks that investigated how Catalan listeners use pitch accent information and facial gestures in making this linguistic distinction. Experiment 1 analyzed whether visual information is a more important cue than auditory information when a continuum of pitch range differences (the main acoustic cue to the distinction between CFS and CEQ) co-occur with congruent and non-congruent facial gestures. Experiment 2 analyzed whether the role of auditory information is stronger when visual information is particularly ambiguous. In this case the visual stimuli were created by means of a digital image-morphing technique. Several important conclusions can be drawn from the results of these experiments with regard to the perception of statement and question prosody.

First, in both experiments, the response frequencies given by Catalan listeners revealed a clear preference for giving priority to visual cues when deciding between a CFS and CEQ interpretation. In both experiments, the listeners' decisions were mainly dependent on whether the video component of the audio + visual material they were watching show facial expressions corresponding to a CFS or a CEQ. Thus the present results show that CFS and CEQ can be discriminated predominantly from visual information, with auditory information (on the basis of an F0 pitch range contrast) probably playing a secondary reinforcing role. In these experiments, the facial gesture acts as an integral part of language comprehension and, as such, provides insight into fundamental aspects of prosodic interpretation.

A second result that is obtained in the two experiments (and which can be observed in Figures 20 and 22) is the effect of bimodal audio + visual congruity. In both experiments, stimuli were identified as a "CEQ" more quickly and more accurately when CEQ-based visual stimuli occurred with a congruent audio stimulus (i.e. the upstepped pitch accent configuration L+¡H* L%). By contrast, identification became slower and less accurate (more chance-like) when the visual stimuli occurred with exemplars of the incongruent nuclear pitch configuration (i.e. L+H* L%). That is, when Catalan listeners saw a CEQ-based visual stimulus occurring with an incongruent low-pitched auditory stimulus, an important time delay appeared in the response, and vice versa. Importantly, the strong effects of congruity/incongruity both in patterns of results and in reaction time measures represent a clear argument in favor of the view that facial gestures and speech form a single integrated system.

Third, another important result refers to the enhanced importance of acoustic stimuli when visual input is ambiguous. Attenuating the differences in the visual stimuli in Experiment 2 triggered a stronger influence of the auditory signals. Concerning the theories of speech perception, integration models predict that both auditory and visual information are used together in a pattern recognition process. On the one hand, the weighted averaging model of perception (WTAV; see Massaro 1998) predicts that the sources are averaged according to weight assigned to each modality. On the other hand, the fuzzy logical model of perception (FLMP) predicts, moreover, that the influence of one modality will be greater than the other when the latter is more ambiguous. According to the results of our Experiment 2, and in line with the findings of Massaro and Cohen (1993), we argue that this model of speech perception accounts for the processing of prosodic information better than competing models of perception (see also Srinivasan & Massaro 2003).

Our results showing a strong role for visual information in the perception of interrogativity seems to partially contradict the results of a large number of studies in audiovisual prosody (e.g. Krahmer et al. 2002, Swerts & Krahmer 2004, Srinivasan & Massaro 2003, House 2002, Dohen & Lœvenbruck 2009, and others). We believe that it is in fact surprising that previous literature on audiovisual speech perception has not found more evidence of the role of visual information in linguistic interpretation. One possible explanation is that the use of real audiovisual recordings is better than the use of embodied conversational agents in avoiding the uncanny valley (the hypothesis in the field of robotics and 3D computer animation which holds that when facsimiles of humans look and act almost, but not perfectly, like actual humans, it causes a response of revulsion among human observers; Mori 1970, Prieto et al. 2011). Moreover, the claim that visual cues simply provide redundant information seems to be at odds with the famous McGurk audiovisual 'illusion' discovered by McGurk and MacDonald (1976). The basic McGurk effect found that an auditory [ba] stimulus combined with a visual [ga] stimulus resulted in a [da] percept. This effect is quite robust and has been replicated for many languages (see Burnham 1998, for an extensive review), thus suggesting that the brain tries to find the most likely stimulus given the conflicting auditory and visual cues, and that visual and auditory information are fused rather than the visual information being superimposed on the auditory one (see also MacDonald & McGurk 1978).

Virtually all studies that have found a complementary effect of visual cues have dealt with the perception of prominence or focus. Yet the studies that have focused on the role of facial expressions as salient indicators of the individual's emotional state (such as incredulity, counter-expectation or surprise in echo questions, degree of uncertainty, etc.) have found a very strong effect of these cues. For example, the studies by Dijkstra et al. (2006), Swerts and Krahmer (2005), and Mehrabian and Ferris (1967), found that visual information is far more influential than acoustic information. Dijkstra et al. (2006) dealt with speakers' signs of uncertainty about the correctness of their answer and showed that facial expressions were the key factor in perception. Similarly, Swerts and Krahmer (2005) showed that there are clear visual cues for a speaker's uncertainty and that listeners are better capable of estimate another person's uncertainty on the basis of combined auditory and visual information than on the basis of auditory information alone.

Nevertheless, Srinivasan and Massaro (2003) showed that statements and echoic questions were discriminated auditorily and visually, but they also found a much larger influence of auditory cues than visual cues in these judgments. We argue that the discrepancies between our results and theirs might be related to the audiovisual materials used. First, their visual materials were based on a synthetic talking head. The question face was characterized by a significant eyebrow raise and head tilt which extended dynamically across the length of the utterance. Yet it is well known that the eyebrow raise can also mark focalized constituents in statements, thus rendering the visual cues ambiguous between a question interpretation and a focus statement interpretation. Second, their auditory materials were manipulated on the basis of the F0 contour, amplitude and duration. Crucially, their difference in F0 contour implied changing a larger structure of nuclear and prenuclear tonal configurations (e.g. *We owe you a yo-yo / Pat cooked Pete's breakfast / We will weigh you / Chuck caught two cats*), leading to large modifications in the F0 stimulation, whereas our F0 changes were limited to changes in the pitch range of a single tonal target that always created a rising-falling intonation sequence. Listeners might have paid more attention to the sentential intonation contour than to the facial cues. As the authors themselves point out, “to assess whether the extended length of the sentence was responsible for nonoptimal integration, a shorter test stimulus (e.g.: “Sunny. / Sunny?”) might be used. A short utterance might make statement / question identification a more automatic perceptual task, and less of a cognitive decision-making process. This task might engage an optimal bimodal integration process.” (Srinivasan & Massaro 2003:20)

Summarizing, our results provide clear evidence for the importance of visual cues in the perception of linguistic contrasts (in our case, the perception of statements and questions) and open the way to new investigations in this area. One of the research questions is the relevance of potential facial cues and their contributions to the judgments of statements and questions. We have also tested this question by using computer-generated 3D talking heads to simulate face gestures during speech production (Borràs-Comes et al. 2011). In that study, the visual stimuli are implemented in a computer-generated 3D avatar in which each intended facial gesture — in that case, eyebrow furrowing, eyelid closure, and head movement — is manipulated separately and appears on a continuum of four levels of strength.

CHAPTER 5

Audiovisual correlates of interrogativity: a crosslinguistic study

5.1. Introduction

The world's languages have different grammatical means to mark an utterance as a *yes-no* question (e.g., *Are you hungry?*, *Does the shop open on Saturday?*), including the use of different lexical items or morphemes, changes in the syntactic structure, or prosodic and gestural marking. While declaratives are considered to be the unmarked sentence type, primarily used to convey information with no special illocutionary force markers (Levinson 2010: 2742), questions are primarily used to seek information.

Crosslinguistically, morphosyntactic features have been shown to constitute a common way to identify *yes-no* questions. Among these strategies, we find the presence of question particles (*est-ce que* in French, ли [*li*] in Russian), the presence of interrogative clitics (*ne* in Latin, 까 [*ka*] in Korean), a specific interrogative word order (as in most Germanic languages), or a combination of such strategies. As Dryer (2008) states, most languages using these morphosyntactic strategies also employ a distinct intonation pattern, though some do not (e.g., Imbabura Quechua, spoken in Ecuador).

Prosody is also a very common resource to signal *yes-no* questions across languages. It can be used to assign question status to a declarative-formatted sentence (Stivers & Rossano 2010 for Italian), even in those languages that use morphosyntactic strategies (as happens with the so-called *declarative questions*, which are those sentences that maintain the typical word order of a declarative sentence but have been produced with a specific interrogative intonation contour; see Englert 2010 for Dutch).

Bolinger (1986) argued that the presence of high pitch in questions may even be considered a linguistic universal (i.e., the fact that the average pitch in questions tends to be higher than the average pitch in non-questions). Moreover, Cruttenden (1981) suggested that the universal dichotomy between falling and rising tunes may be associated with the abstract notions of *closed* (for falls) vs. *open* status (for rises). However, some recent descriptive studies like Englert's (2010) have pointed out that this prosodic feature is not exclusively tied to interrogativity but is also a common device for signaling continuation in statements or, at the level of discourse, both turn-giving and turn-keeping. In contrast with Bolinger's claim mentioned above, Rialland's (2007) analysis of 78 Central African languages showed that question prosodies without any high-pitched correlates are widespread and include falling intonations or low tones, lengthening, breathy termination, and open vowels.

Though the analysis of morphosyntactic and prosodic markers of *yes-no* questions has received considerable attention in the linguistics literature, less is known about the relevance of nonverbal cues. Nonetheless, various studies in the last three decades have taken into account the potential importance of eye gaze and certain facial and manual gestures. In fact, backchannel signals like facial expressions, head movements and gaze, seem to be critically linked to listeners' attention, perception, and comprehension (Peters et al. 2005, Lysander & Horton 2012). Argyle and Cook (1976) argued that gaze serves three main purposes during face-to-face communication: seeking information, receiving signals that accompany the speech, and controlling the flow of the conversation. Cosnier's (1991) study of French spontaneous speech revealed that the gestural traits that characterize information-seeking questions are those that normally accompany informative verbal expressions, namely, eye gaze to the interlocutor, head elevation, an optional suspended hand gesture facing the interlocutor, and a variety of facial expressions which are then frozen while the speaker awaits a response. Cosnier in fact argued that gaze is as important as intonation and pauses for question marking and turn-taking. As Vilhjálmsson pointed out (1997: 21-22), since the primary function of the eyes is to gather sensory input, the most obvious function of gaze is perhaps information-seeking, since the speaker will at least look at the listener when feedback is expected.

Eyebrow movements have also been associated with questioning, though the results appear to be somewhat inconclusive. For instance, Srinivasan and Massaro (2003) made use of "talking heads" (synthetic representations of a human face) in which they varied specific auditory and visual characteristics to investigate whether these could differentiate statements from declarative questions in English. They found that both eyebrow raising and head tilting could increase the perceivers' detection of a question, though participants tended to rely more on auditory cues. However, Flecha-García's (2010) analysis of English spontaneous speech materials found that speakers do not use eyebrow raises in questions more often than in other types of utterances. Yet, incidentally, she also suggested that eyebrow raises may add a questioning meaning to any utterance — somewhat like adding a tag question at the end — even if the utterance does not express a question or request, whether verbally or prosodically (Flecha-Garcia 2010: 553).

In line with this crosslinguistic variation, recent studies prefer to look at question marking as the set of features that contribute to response mobilization (Stivers & Rossano 2010: 29, Haan 2002). Stivers and Rossano (2010) found for both English and Italian that no single feature is present in all cases and thus conclude that no such feature appears to be intrinsic to the action of requesting information (2010: 8). They state that if an assessment is accompanied by several response-mobilizing features, this increases the response relevance of the action (2010: 28). From a crosslinguistic point of view, even though speakers of different languages rely on different question marking correlates, the same response-mobilizing resources — gaze, lexico-morphosyntax, prosody, as well as contextual epistemic asymmetry — seem to be available across languages, ethnicities, and cultures (Stivers & Rossano 2010: 29). In general, Rossano (2010) observed a trade-off relationship between mobilizing cues, and observed that Italian speakers tend to look more often at recipients when those utterances do not have a clear intonational marking. In addition, he found that speakers looked more at recipients during *yes-no* questions and alternative questions than during *wh*– questions, which can also be linked to the fact that the latter show a greater use of interrogative verbal cues than the other two types of questions (i.e., *wh*- words). Moreover, Levinson (2010, see also Stivers 2010) has shown that pragmatic inference is a crosslinguistic cue for interrogativity detection and can even represent the main question marker in a language (Levinson 2010, for Yélî Dnye). If the speaker makes a statement about anything of which the recipient has greater knowledge, this routinely attracts the recipient’s response (Labov & Fanshel 1977, Pomerantz 1980).

To our knowledge, no controlled experimental studies have been undertaken to explore what role verbal and nonverbal cues play in the production and perception of questions and whether there exists a trade-off relationship between different mobilizing correlates. To date, the majority of descriptions have been based on the analysis of controlled or natural corpora, and some perception studies have assessed the audiovisual identification of ‘biased’ questions (i.e., those conveying, for instance, counter-expectation, incredulity, or surprise), most of them by means of synthetic materials (House 2002, Srinivasan & Massaro 2003, Borràs-Comes et al. 2011, Crespo-Sendra 2011, see also Chapter 4). There are still a number of open questions that have not received a complete answer, such as: Can we differentiate an information-seeking *yes-no* question from a broad focus statement by means of visual information alone? How does visual information contribute to question identification when added to auditory information? Does the simultaneous use of several questioning cues increase the perceiver’s identification of an utterance as a question? Do nonverbal cues have a major role in those languages in which intonation and syntactic cues do not play a defining role?

The present chapter aims to compare interrogativity-marking strategies in Dutch and Catalan, two European languages that have been argued to rely on different resources for this distinction. One the one hand, Dutch *yes-no* interrogatives are characterized by subject/verb inversion, without making use of an auxiliary verb as is the case in English (Dut. *Heb je een man?*, lit. 'Have you a man?', 'Do you have a man?'; Englert 2010: 2668). By contrast, subject/verb inversion is not available for *yes-no* questions marking in Catalan (Cat. **Té ell bigoti?*, lit. 'Has he moustache?'), and grammatical subjects are generally not produced when related to a known referent (Cat. *Té bigoti?*, lit. 'Has moustache?', 'Does he have a moustache?') or appear dislocated to a postfocal position (Cat. *Té bigoti, ell?*, lit. 'Has moustache, he?') both in statements and questions. In terms of prosody, speakers of Dutch appear to draw on the overall set of phonological devices of their language for question-marking, though certain configurations are more likely to occur in questions than in statements, as happens with rising tunes (Haan 2002: 214). By contrast, in order to convey information-seeking *yes-no* questions most Catalan dialects have been claimed to use a specific intonational contour which consists of a low pitch accent followed by a rising boundary tone (Prieto & Rigau 2007).[14] Drawing on Rossano's (2010) hypothesis, we expect that the use of prosodic and gestural cues by speakers of Catalan will be more productive than the use of such cues by speakers of Dutch, since the latter language uses an additional syntactic strategy to mark questions (see also Geluykens 1988).

This chapter has two related goals. First, we aim to describe the combination of syntactic, prosodic, and gestural cues used by

[14] Catalan can also mark interrogativity through the expletive particle *que* (cf. *est-ce que* in French and *é que* in Portuguese), which is especially found for Central, Balearic, and North-western Catalan in confirmation-seeking questions (Prieto & Rigau 2007).

Dutch and Catalan speakers for the marking of information focus statements (IFS) and information-seeking *yes-no* questions (ISQ). In order to collect a series of IFS and ISQ for our perception experiment, we conducted a production task using two variants of the *Guess Who* game. As Ahmad et al. (2011) point out, the dynamic nature of games make them a good tool for investigating human communication in different experimental setups, especially if the outcome of a game is controllable in a systematic manner.

The second goal of the chapter is to test whether and how listeners of the two languages differentiate (ISQ) questions from (IFS) statements, as well as to evaluate the relative importance of the different cues used in production and perception. A random set of the stimuli obtained by means of the production task was therefore used as stimulus materials for a test in which participants had to guess whether an utterance was a statement or a question. Participants were presented with materials in three perceptual conditions: one in which only the auditory information was available (AO), another one in which only the visual information was available (VO), and a third one which presented simultaneously the full auditory and visual information of the actual recordings (AV). This identification test allowed us to assess the relevance of the various features and their potential interaction effects.

5.2. Experiment 1

5.2.1. Methodology

In order to obtain a set of natural productions of statements and questions in Dutch and Catalan, we designed a production task based on two variants of the *Guess Who* game which would allow us to observe which prosodic cues are used by people when giving instructions or asking yes-no questions. Especially, we are interested in their perceptual importance when another group of native listeners have to judge the specific materials in terms of whether they are statements or questions.

Participants

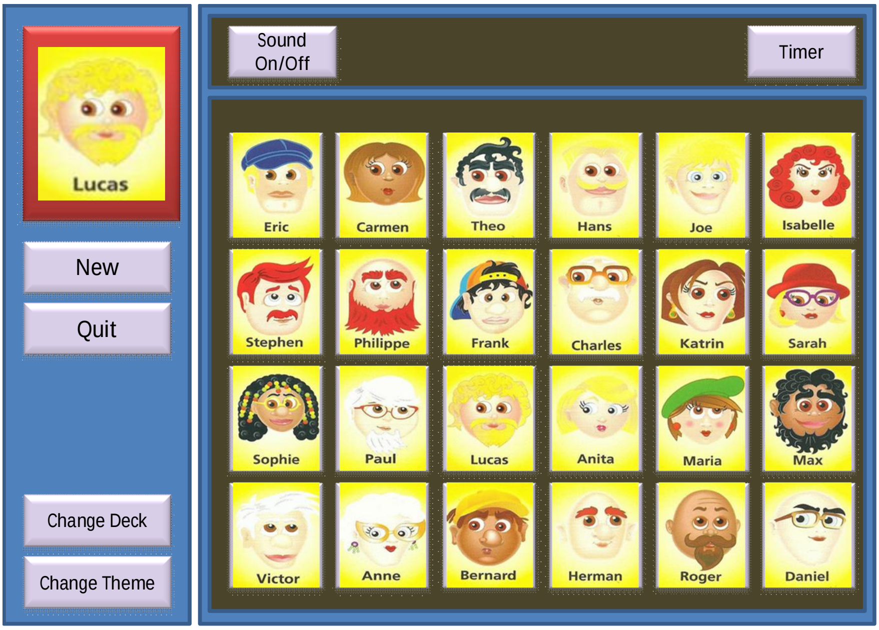
Sixteen Dutch speakers and sixteen Central Catalan speakers participated in the production task. Participants played the game in pairs, taking turns in adopting the roles of participant A and B in the two procedures described below. Participants only played the game with other native speakers of their own language. All subjects were undergraduates at either the Tilburg University, The Netherlands, or the Universitat Pompeu Fabra in Barcelona, Spain. All participants played both variants of the game.

Procedure

In order to elicit IFS and ISQ in a natural manner, we used two digital variants of the "Guess Who" board game as created by Suleman Shahid, from Tilburg University, and some colleagues (see Ahmad et al. 2011). In this game, participants were presented with a board containing 24 colored drawings of human faces (see an example in Figure 23). These faces differed regarding various parameters, such as gender or color of their skin, hair, and eyes. Some faces were bald, some had beards or moustaches, and some were wearing hats, glasses, or earrings. In the traditional version of "Guess Who", the purpose of the game is to try to guess the opponent's mystery person before s/he guesses yours.[15]

[15] This experimental setup provides a clear advantage over real situations. As Richardson et al. (2009) state, a question typically implies turn transition, and several studies have shown that gaze is related with turn-giving (Kendon 1967, Kendon 1990, Argyle & Cook 1976, Duncan & Fiske 1977). Moreover, Englert (2010) has shown for Dutch that questioners rely overwhelmingly on speaker gaze (90%) for next speaker selection. Thus, in order to describe the nonverbal patterns that characterize questions one has to focus on those cases in which gaze plays no addressee-selection role, and this is controlled in our study since participants are engaged in dyadic situations.

Figure 23. Example of the screen image used in the game procedure. At the left, the mystery person of our opponent is shown (top) and buttons for starting a new game or quitting it (middle). The 24 faces make up the main game panel.



Given our need to elicit either information focus statements or information-seeking questions, we asked participants to play one of two different variations of the game. In the *question-elicitation* variation, participant A had to ask Participant B questions to try to determine the mystery person on B's face card. Players took turns asking questions about the physical features of their respective "mystery persons" in an effort to eliminate the wrong candidates. The winner is the player who guesses his/her mystery person first. In the *statement-elicitation* variation of the game, participants take turns making statements about their mystery person, while the other player listens and eliminates all characters that do not exhibit a particular physical feature. Again, it is the player who guesses the identity of their “mystery person” first that wins.[16] Note that both participants within a pair took turns in the course of both variations of the game and therefore both provided examples of questions and statements. Prototypical dialogs of these two procedures are shown in (5); target sentences appear in boldface.

(5) a. Question-elicitation procedure

(A looks at his/her board and thinks of a question that may be useful for him/her to narrow down the number of candidates for “mystery person”)

A: **Does your mystery person have brown eyes?**

(B checks for this feature on his/her mystery person)

B: Yes.

(A unchecks all the faces on his/her screen that do not have brown eyes. Now it is B’s turn to ask a question)

b. Statement-elicitation procedure

(A thinks of a physical feature that will help participant B eliminate some candidates)

A: **He has brown eyes.**

(B unchecks all the faces on his/her screen that do not have brown eyes. Then B tries to guess who the mystery person is)

B: Could it be Bob?

(A checks to see if the mystery person is called Bob)

A: No.

(Now it is B’s turn to describe a feature of his/her own mystery person for A)

Participants sat in the same room, facing each other across a table and in front of two laptop computers arranged so that they could not see each other’s screen. Two camcorders were placed in such a way that they could record the upper part of each participant’s

[16] In order to increase the number of interactions and communication flow between participants — and to avoid continuation rises in the intonation patterns they produced — we added an additional rule to the game: at the end of each turn, players had to try to guess the mystery person’s name. This additional set of questions was not subjected to analysis.

body (see Figures 24 and 25). Before the start of each experiment, the camera was raised or lowered according to the participant's height. Once the participants were seated, the experimenter gave spoken instructions, telling the participants about the game and procedure to be followed for each variation. Each game lasted approximately twenty minutes, the time it took for both variants of the game to be played and won (4 to 6 times each).

Figure 24. Schematic (birdseye) drawing of the experimental set up.

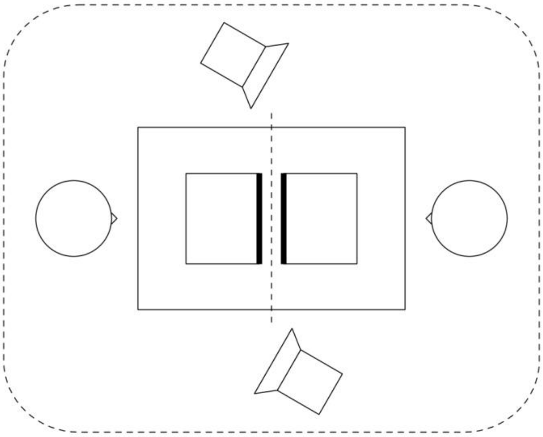
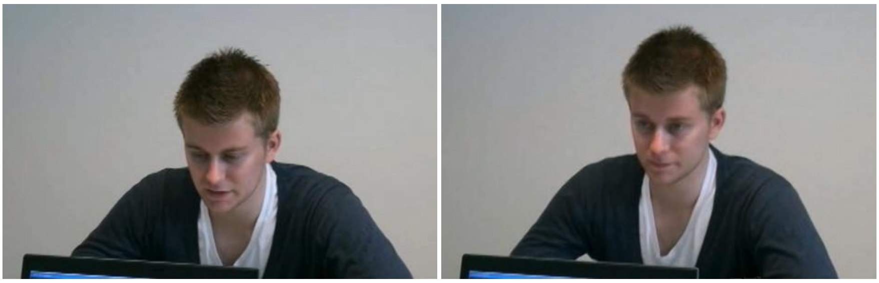


Figure 25. Stills depicting one of the Dutch-speaking participant's video recordings while uttering a statement (left) and a question (right).



Analysis

From the production recordings, 35 statements and 35 questions related to gender (e.g., *It is a man* vs. *Is it a man?*) were randomly selected for each language in order to be included in the subsequent rating task. One participant from each language group involved in the production experiment did not produce any of these utterances, so the final set of materials came from 15 Dutch speakers and 15 Central Catalan speakers. Whenever available, we guaranteed that each speaker provided a similar number of statements and questions.

With the aim of assessing the discrimination power of prosodic and gestural cues, the first two authors of the original article — native speakers of Catalan and Dutch, respectively, but with some knowledge of each other's language — independently coded the selected audiovisual materials (a total of 70 utterances) in terms of the following cues (based on Cosnier 1991):

- order of the sentence constituents (SV, VS, V)
- intonation (falling or rising boundary tone; i.e., L% vs. H%)
- gaze to interlocutor (presence, absence)
- eyebrow raising movement (presence, absence)

The inter-transcriber agreement between the two labelers' coding was quantified by means of the Cohen's kappa coefficient (Cohen 1960), which gave an overall coefficient of .838, which means that the strength of the agreement was very good (Landis & Koch 1977). The coefficient was .855 for Dutch and .822 for Catalan. Concerning the different cues, it was .721 for the boundary contour, .914 for gaze, and .701 for eyebrow raising.

5.2.2. Results

Table 9 presents the results of the presence of cues found in the database. Regarding syntax, the subject was omitted in all Catalan sentences, which only displayed the verb and predicate (Cat. *És una dona*, lit. 'Is a woman', 'It is a woman'). In turn, all Dutch statements presented a SV order (Dut. *'t is een vrouw*, 'It is a woman') and all Dutch questions presented a VS order (Dut. *is 't een vrouw?*, 'Is it a woman?'). In terms of intonation, the same pattern of results was attained for statements in the two languages, showing a great number of falling tones (mostly L* L% and some H* L%).[17] Rising tones (L* H%) were found more often in Dutch questions than in Dutch statements (though Dutch questions exhibited a larger number of falling tones than rising tones; see Geluykens 1988). In turn, Catalan showed a clear majority of questions produced with a rising tone (L* H%, as in the case of Dutch).

Concerning the two visual cues labeled (presence of gaze, eyebrow raising), the two languages showed similar distributions of their uses in statements and questions. Crucially, the presence of gaze and eyebrow raising were found to be more present in questions. Overall, Catalan speakers also seem to use more non-syntactic cues than Dutch speakers.

Table 9. Number of utterances containing the four labeled cues, for each meaning, in Dutch and Catalan.

	Dutch		Catalan	
	statements	questions	statements	questions
VS order	0	35	0	0
rising intonation	4	13	4	33
eye gaze	9	21	12	24
eyebrow raising	5	9	6	16

[17] Please note that, although a broad ToBI analysis was applied for analyzing Dutch intonation and it has properly accounted for the variation observed in the present study, a language-specific system for transcribing Dutch intonation has been proposed in the literature, namely ToDI (Gussenhoven *in press*). ToDI parallels our broad ToBI analysis by showing a distinction between L% and H% IP-final tones, which can be specified or not. Moreover, the falling patterns observed in our study can be transcribed as H*L L% and the rising patterns as L*H H%.

5.3. Experiment 2

5.3.1. Methodology

Participants

In the perception experiment, twenty Dutch listeners (between 18 and 35, average = 24.6, standard deviation = 3.82) and twenty Catalan listeners (between 18 and 25, average = 22.1, standard deviation = 1.80) rated the selection of 70 stimuli in their own L1 as being statements or questions. As the stimuli were excerpts from recordings made during the first experiment. None of the participants in the first experiment took part in the second one.

Materials

A selection 35 statements and 35 questions related to gender (e.g., *It is a man* vs. *Is it a man?*), for each language, randomly selected from the production recordings.

Procedure

The target 70 stimuli were presented to each group of same-language participants in three different conditions in a within-subjects design: Auditory-Only (AO), Visual-Only (VO), and AudioVisual (AV). In order to control for a possible learning effect, the AV condition was always the last to be presented to the participants, and the order of the two unimodal conditions was counterbalanced among subjects. Inside each condition, the different sentences were presented in a randomized order.

Stimuli were presented to subjects using a desktop computer equipped with headphones. Subjects were instructed to pay attention to the stimuli and decide which interpretation was more likely for each stimulus by pressing the corresponding computer key for *statement* and *question*: ‘A’/‘P’ (*afirmació*, *pregunta*) for Catalan, and ‘S’/‘V’ (*stelling*, *vraag*) for Dutch. No feedback was given on the “correctness” of their responses. Participants could take as much time as they wanted to make a decision, but could not return to an earlier stimulus once they had made a decision on it.

The experiment was set up by means of E-Prime version 2.0 (Psychology Software Tools Inc. 2009), which allowed us to record responses automatically. A new stimulus was presented only after a response to the previous one had been given. The experiment was set up in a quiet research room at either Tilburg University and or the Universitat Pompeu Fabra, respectively. It lasted approximately 17 minutes. The total number of responses obtained was 8,400 (70 stimuli × 20 subjects × 3 conditions × 2 languages).

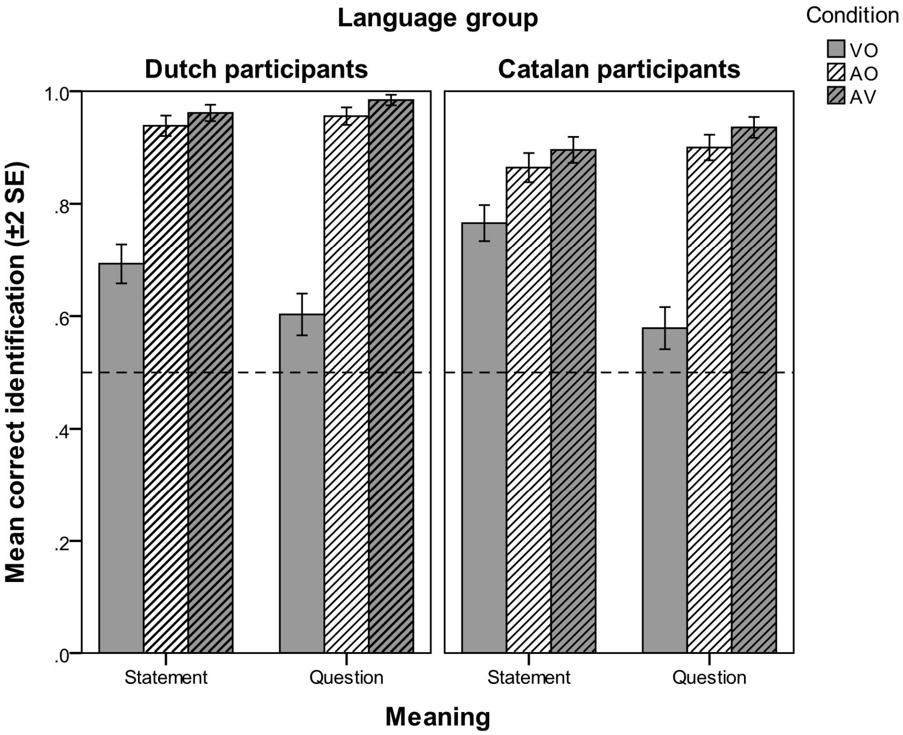
5.3.2. Results

General perception results

Figure 26 shows the mean correct identification rates of the perception experiment broken down by language (Dutch, Catalan), condition (AO, VO, AV), and meaning (statement, question). The results in the graph show that participants in both languages were able to identify the two categories above chance level in all three presentation conditions. However, materials that included auditory information (i.e., VO and AV) were consistently more reliable conveyors of question identification.

A Generalized Linear Mixed Model (GLMM) analysis was run with the correct identification of the utterance category as the dependent variable, with language, condition, meaning, and all the possible interactions as fixed factors and subject and item(speaker) as random factors. Main effects for language ($F_{1, 155}$ = 6.578, p = .011) and condition ($F_{2, 8388}$ = 417.403, p < .001) were found, but not for meaning ($F_{1, 152}$ = 0.462, p = .498). Two interactions were also found to be significant: language × condition ($\mathrm{F}_{2, 8388}$ = 21.504, p < .001) and condition × meaning ($F_{2, 8388}$ = 33.481, p < .001).

Figure 26. Mean correct identification rate (y-axis) as a function of language group (Dutch, Catalan), condition (different bars: VO, AO, AV), and intended meaning (x-axis: statement, question).



Bonferroni post-hoc tests were extracted in order to know the direction of the significant main effects and interactions. They show an effect of condition such that AV > AO > VO (all paired comparisons, *p* < .001). Concerning the interaction language × condition, Dutch participants were more accurate than Catalan participants only when auditory information was available: AO (*p* = .002) and AV (*p* < .001), and not in VO (*p* = .529). Concerning the interaction condition × meaning, statements were more accurately identified than questions only when visual information was available: VO (*p* = .001) and AV (*p* = .006), and not in the AO condition (*p* = .128).

In sum, the perception results shown here reveal that participants could identify questions and statements above chance level in all conditions. Specifically, participants' responses were better when auditory information was present, but a beneficial effect of visual cues was also shown when they were added to the auditory ones. In addition, Dutch participants' perception of auditory materials was found to be better than that of Catalan participants, with less of a difference between language groups when they were presented with VO materials, which allows us to hypothesize that language differences were most pronounced when the auditory components of the experiment materials were involved. Importantly, our results show that when visual information is present, statements are better identified than questions. These questions are further investigated in the next section, where we analyze the materials in terms of their specific auditory and visual features.

Unimodal perception of auditory and visual features

The lack of syntactic marking in Catalan (i.e., zero degrees of freedom) makes it impossible for us to compute the interactions in which language and syntax are implied.[18] As for the perception of these intonation differences, a GLMM analysis was conducted on the results of the AO task, with identification as the dependent variable, language, contour, and their interaction as fixed effects, and subject and speaker as random factors. There were main effects for language ($F_{1,\ 26}$ = 11.665, p = .002), contour ($F_{1,\ 2796}$ = 601.409, p < .001), and their interaction ($F_{1,\ 2796}$ = 79.249, p < .001).

[18] In order to know the effect of both syntax and intonation within Dutch, a language-specific GLMM analysis of the AO task was performed, with IDENTIFICATION as the dependent variable, SYNTAX, CONTOUR, and their interaction as fixed effects, and SUBJECT and SPEAKER as random factors. All factors were significant: SYNTAX ($F_{1,\ 107}$ = 331.192, p < .001), CONTOUR ($F_{1,\ 32}$ = 16.989, p < .001), and their interaction ($F_{1,\ 59}$ = 6.087, p = .017). Bonferroni paired contrasts crucially showed that the interaction SYNTAX × CONTOUR was related to the fact that a rising contours caused more question identifications when applied to a SV structure (p < .001), but not when applied to a VS structure (p = .180).

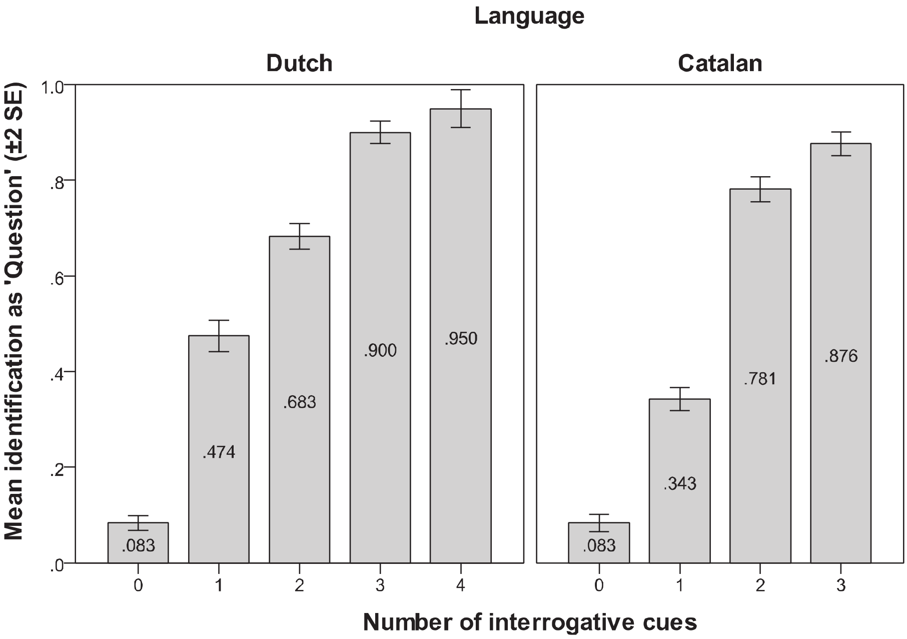
The significant interaction is due to the fact that Catalan listeners rated more falling contours as statements than Dutch listeners (*p* < .001) but this difference does not hold for rising contours (*p* = .328), suggesting that rising contours are perceived equally often as question-conveyors for both language groups. This is consistent with the patterns found in production.

Another GLMM analysis was conducted on the results of the VO task, with identification as the dependent variable, and subject and speaker as random factors. The fixed effects were language, gaze, eyebrow, and all the possible interactions. Main effects were found for gaze ($F_{1,\ 2080}$ = 283.044, *p* < .001), eyebrow ($F_{1,\ 2792}$ = 21.042, *p* = .004) and language ($F_{1,\ 37}$ = 8.879, *p* = .005). Two interactions were also found to be significant: gaze × eyebrow ($F_{1,\ 2792}$ = 16.094, *p* < .001), and the triple interaction gaze × eyebrow × language ($F_{1,\ 2792}$ = 4.425, *p* = .035). The main effects of gaze and eyebrow are related to the patterns observed in production, i.e., that the presence of these cues increased ‘question’ responses. The main effect of language suggests that Dutch participants gave overall more ‘question’ responses than Catalan participants. As for the gaze × eyebrow interaction, eyebrow had a significant effect on ‘question’ identification when in the presence of gaze (*p* < .001), but not in its absence (*p* = .678). Regarding the triple interaction, a language difference is found, such that Dutch participants provided more ‘question’ responses than Catalan participants when gaze (*p* = .003) and eyebrow (*p* = .006) appeared alone in the perceived materials, but not when these features co-appeared (*p* = .331) or were both absent (*p* = .058).

Auditory and visual features combined

A main question related to cue interaction is whether the presence of different cues related to questioning can significantly increase the detection of questions. To this end we created a new column in our results database that contained the sum of the different cues to questioning found in both languages (i.e., VS syntax, rising intonation contour, presence of gaze, and eyebrow raising). The graph in Figure 27 shows that the incremental presence of cues to questioning does increase participants' 'question' responses in both languages.

Figure 27. Mean identification as 'question' (y-axis) of the materials in the perception experiment divided by the number of interrogative cues that they contain, in both Dutch (i.e., VS + rise + gaze + eyebrow) and Catalan (i.e., rise+ gaze + eyebrow).



A Pearson correlation (2-tailed) was conducted between the number of interrogative cues and the identification responses. The test identified a positive correlation of .736 in the case of Dutch and a correlation of .709 in the case of Catalan (in both cases, *p* < .001), which means that there is a high correlation of the two variables in each language.

5.4. Discussion

The first goal of the present chapter was to describe the syntactic, prosodic, and gestural strategies used by Dutch and Catalan speakers for marking information-seeking *yes-no* questions (ISQ) and information focus statements (IFS). These two languages have been argued to mark interrogativity in two different ways. Whereas Dutch *yes-no* questions are characterized by the use of a syntactic verb fronting strategy and optional intonational marks (e.g., *Hij heeft een baard* vs. *Heeft hij een baard?*, lit. 'He has a beard' vs. 'Has he a beard?'), Catalan *yes-no* questions do not allow SV inversion and the main strategy in this language is the use of specific intonational patterns (e.g., *Té barba* vs. *Té barba?*, lit. 'Has beard' vs. 'Has beard?'). On the one hand, the fact that Dutch indeed has a systematic syntactic strategy as described in the literature was confirmed by the results of our production task. As for prosody, both languages showed a great number of rising tones in questions, though Catalan (because of the lack of any lexico-morphosyntactic distinction in our target sentences) showed a stronger effect of intonation for interrogativity marking. Concerning gestures, both languages showed similar distributions of the use of gaze and eyebrow raisings, which were mainly found in questions.

The second and main goal of this investigation was to test whether listeners of the two languages could differentiate questions from statements in the different presentation conditions (AO, VO, AV), as well as to evaluate the relevance of the different cues used in perception. The results of our perception experiment with 20 Dutch listeners and 20 Catalan listeners confirmed that participants can identify questions and statements above chance level in all conditions. Importantly, perceivers showed a great reliance on auditory information, but also showed that (*a*) visual-only utterances were classified above chance; and (*b*) better accuracy in responses was exhibited when visual information was added to auditory information. This result confirms the importance of nonverbal cues in speakers' identification of pragmatic intentions but also suggests a higher importance of auditory cues in the perception of interrogativity.

Focusing on the auditory-only perception, Dutch participants were found to be more accurate than Catalan participants, which can be linked to the fact that Dutch uses an unambiguous syntactic strategy. With respect to the perceptual importance of syntax and intonation in Dutch, an analysis of the Dutch listeners' perception of AO information revealed that both factors were significant. Moreover, there was an interaction between the two, in the sense that rising contours led to more 'question' identification responses only when applied to an unmarked (SV) syntactic structure. This demonstrates that when both markings are available syntax has greater importance relative to intonation.

When focusing on the visual-only perception, gaze played an especially strong role in 'question' identification responses in both languages. This is in line with Rossano's (2010) production results for Italian, which showed that the occurrence of speaker gaze towards the recipient in dyadic interactions increases the likelihood of obtaining a response. As for eyebrow raising, a secondary role was found such that it powered 'question' responses only when in the presence of gaze.

More crucially, in the AV presentation, we found a positive correlation between the concentration of mobilizing cues in a sentence and its rating as an interrogative utterance, for both languages. This result is especially relevant for the theory of response relevance put forward by Stivers and Rossano (2010). While suggesting four main response-mobilizing features — namely interrogative lexico-morphosyntax, interrogative prosody, recipient-directed speaker gaze, and recipient-tilted epistemic asymmetry — they argue that the inclusion of multiple response-mobilizing features leads to higher response relevance than the inclusion of fewer or no features. In their own words, "a request (or an offer or information request) is high in response relevance, but a request designed 'directly' (e.g., with interrogative morphosyntax and/or prosody) would be still higher. Similarly, an assessment (or a noticing or announcement) would be low in response relevance. However, if it were designed with multiple response-mobilizing features, this would increase the response relevance of the action" (Stivers & Rossano 2010: 27–28). In our data, a higher concentration of lexico-morphosyntactic, prosodic, and gestural cues increases the chances that utterances will be perceived as questions.

To our knowledge, the present chapter provides the first results of a controlled investigation on the crosslinguistic perception of information-seeking *yes-no* questions compared with broad focus statements. First, we have found that auditory information has a greater effect in question identification (auditory cues > visual cues). As for visual cues, we have empirically shown that both auditory and visual cues play a role in this distinction in both Catalan and Dutch. Specifically, the addition of non-verbal cues to auditory cues enhances the perception of information-seeking questions. Also, a visual-only presentation of the materials led to successful interrogativity detection. In terms of its perceptual relevance, a greater effect was found for gaze compared to eyebrow raising. This pattern of results suggests, at least when taking into account Dutch and Catalan data, a cue value scale for interrogativity marking such that syntax > intonation > gaze > eyebrow. In conclusion, this chapter shows how several verbal and nonverbal cues are systematically used in the production of interrogativity and how they crucially interact in its perception.

CHAPTER 6
General discussion and conclusions

6.1. The phonological status of pitch range

One of the main goals of this thesis was to describe the role of pitch range in conveying interrogativity. In Catalan, the same sequence of low and high tones in a nuclear pitch configuration can express three different pragmatic meanings depending on its pitch range properties: information focus statement (IFS), contrastive focus statement (CFS), and counter-expectational question (CEQ). Given this three-way contrast in meaning potentially triggered by pitch range, we ran a series of behavioral and electrophysiological experiments in order to find out whether the difference between these three meanings is cued by pitch range in a discrete fashion.

Our investigation of the role of pitch range in the intonational grammar of this language has been couched in the Autosegmental-Metrical (AM) model of prosodic analysis, which takes as a central assumption that only two tones, Low and High, are necessary to distinguish intonational categories in a language like English. In this regard, the role of pitch range has often been relegated to express differences in emphasis or prominence (Pierrehumbert 1980, Beckman & Pierrehumbert 1986). However, work on different Romance and Germanic languages has revealed that pitch range variation can express categorical differences in meaning (Hirschberg & Ward 1992, Ladd & Morton 1997, Savino & Grice 2011, Vanrell 2011), and some authors have suggested that the AM framework has to take this tonal feature explicitly into account as conveyor of categorical distinctions (Ladd 1994, Face 2011). Chapters 2 and 3 were devoted to investigate the phonological role of pitch range in Catalan.

Chapter 2 described two behavioral experiments in which participants were presented with an acoustic continuum of pitch range and had to decide among three possible responses (IFS, CFS, CEQ). From these two experiments we analyzed response frequencies and subjects' reaction times (RTs). In the first experiment, participants had to identify which meaning was understood for each isolated stimulus; on the other hand, in the second one participants had to rate the degree of perceived appropriateness between the stimulus and corresponding congruent (and potentially incongruent) discourse contexts, for each of the three potential meanings. In both experiments, participants associated IFS and CEQ with the low and high ends of the pitch range continuum respectively, while CFS was less consistently associated with a specific range though skewed towards an IFS interpretation.

As for reaction times patterns, the first experiment showed a clear peak in the perceived acoustic boundary between CEQ and the other two types of statements (namely IFS and CFS) and in the second experiment a RT peak emerged only for IFS and CEQ, but not for CFS. Following Chen (2003), if a RT peak located at an identification boundary is taken as an indication of the discreteness of a perceived contrast, we cannot claim that participants' decisions on the appropriateness of CFS sentences are discretely distributed depending on pitch range.

Therefore, the results of Chapter 2 reveal that IFS interpretations are induced by contours with narrow pitch range, whereas CEQ interpretations are triggered by contours with a wider pitch range. Concerning the role of pitch range in CFS marking, our results show that CFS behaves approximately like IFS in terms of pitch range values. The congruity experiment showed that there is no RT peak between the 'appropriate' and 'inappropriate' decisions that affect the role of pitch range for CFS marking, which means that these two responses are not discretely divided by native listeners and so the role of pitch range for CFS marking is simply a gradient phenomenon. The IFS-like behavior and absence of a RT peak might thus be interpreted as meaning that pitch range distinguishes CFS from IFS in a gradient fashion. We argue that the detection of an utterance as being a CFS relies to a greater extent on a pragmatic inferencing process, such that CFS is understood when contrastive information is added to the discourse in normal conversation. Finally, the speaker can also mark the corrective status of that utterance with morphosyntactic strategies like focus fronting, as well as with postfocal prosodic reduction.

Chapter 3 presented two experiments intended to show that the perceived discreteness between IFS and CEQ described in Chapter 2 have a significant electrophysiological correlate. Previous electrophysiological studies of segmental phonological contrasts and tone contrasts from tone languages found evidence that native linguistic contrasts of this sort elicited significantly larger mismatch negativity (MMN) responses than non-native contrasts (Näätänen et al. 1997, Gandour et al. 1994) and that acoustic contrasts that crossed a category boundary lead to larger MMN responses than comparable acoustic contrasts that did not cross these category boundaries (Dehaene-Lambertz 1997, Chandrasekaran et al. 2007). Such results have not yet been obtained for intonational contrasts. Doherty et al. (2004) and Leitman et al. (2009) argued that the large MMN elicited only by interrogative stimuli (and not by the declarative stimuli) "may underlie the ability of questions to automatically capture attention even when the preceding declarative information has been ignored" (Leitman et al. 2009: 289). Fournier et al. (2010) argued that electrophysiological information taken from the human brain did not provide clear evidence for the recognition of discourse meanings by means of intonation.

However, the findings presented in Chapter 3 confirmed the results reported in Chapter 2. In a first identification experiment, a clear nonmonotonic identification of the contrast between IFS and CEQ was found, as well as faster RTs in the identification of within-category exemplars than in more ambiguously-interpreted exemplars. In the second experiment presented in Chapter 3, the mean amplitude of the MMN was found to be larger for the across-category contrast compared to the within-category contrasts, suggesting that intonational contrasts in the target language can be encoded automatically in the auditory cortex. Moreover, our results showed that the activation of these auditory cortex intonational representations was related to the individuals' subjective perception and performance (i.e., that a significant correlation was obtained between the electrophysiological responses and the behavioral measures obtained in the first experiment, both for individuals as well as for the grand mean data). Thus, our results provided electrophysiological evidence that phonological contrasts at the intonational level (based on a pitch range difference) are also encoded in the auditory cortex, which is in line with a substantial set of empirical results that demonstrate the larger activation of memory traces for linguistic elements in the human brain.

Taken together, Chapters 2 and 3 showed that variation in pitch range is the main cue that Catalan listeners use to discriminate between IFS and CEQ, i.e., there is a threshold along a continuum of pitch range beyond which a CEQ meaning is consistently attained. This contrast in pitch range for distinguishing questions and statements has been shown to also signal phonological distinctions in other Romance languages (Savino & Grice 2011 for Bari Italian, Roseano et al. 2011 for Friulian, Estebas-Vilaplana & Prieto 2010 for Castilian Spanish, etc.), as well as in other languages.

These results indicate that an accurate prosodic transcription system for these languages — at least for Catalan — needs to signal the distinction between the IFS patterns (L+H*) and the CEQ patterns (L+¡H*) (Aguilar et al. 2009 for Catalan). In line with this, and following recent work by Vanrell (2011), the inclusion of a tone like [L+¡H*] (with the upstep diacritic), has been proposed to expand the inventory of available pitch-accent phonological contrasts (i.e., three phonologically different tones are thus available in the intonational transcription system for Catalan: L, H, and ¡H).

6.2. Interaction between prosodic and gestural cues in sentence processing

The main goal of Chapters 4 and 5 was to understand the interaction between acoustic and visual cues in the linguistic perception of interrogativity. In Chapter 4, we explored the relative importance of pitch accent range and facial gestures in the perception of the contrast between CFS and CEQ by using congruent and incongruent multimodal stimuli.

The main question to be answered by Chapter 4 was to what extent gestural cues could be central in encoding a linguistically relevant distinction between CFS vs. CEQ. In the two identification experiments included in that chapter, Catalan listeners were presented with congruent and incongruent audiovisual materials. The analysis of their response frequencies revealed a clear preference for visual cues when deciding between a CFS and CEQ interpretation, whereas the pitch range contrast in intonation was observed to play a secondary reinforcing role. These results show that in some circumstances facial gestures can act as central conveyors of prosodic interpretation and compete with prosodic cues, which seems to partially contradict the results of a large number of studies in audiovisual prosody that have found a complementary effect of visual cues (Krahmer et al. 2002, Swerts & Krahmer 2004, Srinivasan & Massaro 2003, House 2002, Dohen & Lœvenbruck 2009, and others).

It is worth mentioning that audiovisual integration effects have been well observed at the segmental level in other research, mostly since the publication of McGurk & MacDonald’s (1976) study. That study showed that when hearing [ba] while looking at lip movements pronouncing [ga] adult English-speakers perceived [da], a phonematic sequence which was not actually present in either the acoustic or the visual input provided to participants. Yet when these same subjects were presented with the same materials unimodally, [ba] and [ga] were perceived respectively. Our results are related to the McGurk effect in the sense that both modalities compete and interact in our participants' decisions, but are different from a 'classic' McGurk effect in that we do not obtain a category that is intermediate between our contrasted statements and questions.

Another interesting result from the two experiments in Chapter 4 is that the role of auditory information is stronger when visual information is particularly ambiguous, which suggests a pattern of audiovisual integration in normal face-to-face communication. This means that when participants were presented with unclear exemplars of CFS and CEQ gestures their reliance on acoustic information was enhanced. Another study using synthetic materials comparing the perception of IFS vs. CEQ in Catalan (Borràs-Comes et al. 2011) provides additional evidence for the pattern observed here. In that study, the reliance on acoustic cues was generally enhanced when they co-occurred with an IFS facial configuration and decreased when presented with a CEQ facial configuration. As expected, given that IFS is a neutral type of statement and a CEQ is a biased type of question, participants relied more heavily on the CEQ facial gestures than on the practically nonexistent IFS gestures.

On the other hand, when gestural and intonational features are salient, listeners tend to rely on both acoustic and visual signals in a more balanced way. Support for this explanation comes from the analysis of the distinction between IFS and CFS in Central Catalan using avatars reported in Prieto et al. (2011). The difference found between IFS and CFS is based both on a gradient activation in pitch range and on the strength of activation of two specific gestures: forward head movement and eyebrow raising. Because both modalities showed a gradient and equally salient distinction concerning the linguistic contrasts studied (IFS vs. CFS), a balanced use of auditory and visual cues was found in the participants' identification of both categories (with head movement being a clearer correlate of CFS marking than eyebrow raising in terms of gestural correlates).

This compensatory interpretation is linked to Crespo-Sendra's (2011) results regarding the audiovisual perception of IFS vs. CEQ in Valencian Catalan and Dutch. Whereas the facial gestures characteristic of the two meanings are found to be similar to those discussed in the present thesis, a clear difference between the two languages is reported concerning intonational marking: whereas Valencian Catalan marks the distinction between the two types of interrogatives with pitch scaling differences over the same rising configuration (L* H%), Dutch uses two very different contours to distinguish between the two meanings (namely L* H% for ISQ and L+H* LH% for CEQ). When both populations were presented with congruent and incongruent combinations of those audiovisual materials, Valencian Catalan speakers relied significantly more on visual cues, whereas Dutch speakers crucially showed a more balanced effect between the two cues in interaction.

6.3. The role of verbal and nonverbal cues in question detection

In Chapter 5, we explored the relative importance of different types of boundary tones and both eye gaze and eyebrow raising in the perception of the contrast between IFS and ISQ in two types of languages, one that exhibits a syntactic strategy (i.e., subject/verb inversion) for question marking (Dutch) and one that does not (Catalan). The results of our perception experiment showed that both Dutch and Catalan participants can identify questions and statements above chance level in all conditions. Importantly, they showed a great reliance on auditory information, but also better accuracy in identification responses when the visual information was added to the auditory one.

This pattern of results, though, is partially in contradiction with those reported in Chapter 4. When participants had to distinguish between IFS and ISQ — nonbiased types of statements and questions respectively — they showed a greater reliance on auditory information compared to visual information (though a visual-only presentation of the materials also yielded a significantly accurate identification rate). In line with what is mentioned above, I suggest that these partially contradictory results are related with the properties of the acoustic and visual cues analyzed in both sets of experiments.

Concerning the contrast between CFS and CEQ (Chapter 4), the visual information contained in each of the two facial patterns was very different, though both are characterized by salient head and eyebrow movements (a forward/backward head movement and a raising/furrowing eyebrow movement); as for the acoustic properties of the two utterance types, though they represent a phonological contrast in the intonational phonology of Catalan (see Chapters 2 and 3), they are based on a single difference in the pitch range properties of the intonational contour. Concerning the contrast between IFS and ISQ analyzed in Chapter 5, the visual information characterizing this difference was perceptually less salient and determined only by the presence or absence of a single feature, namely eye gaze, whose role was found to be improved when adding a raising eyebrow movement; as for the acoustic information, it was based on one of the most commonly applied crosslinguistic dichotomies in intonational languages for question marking, the rising vs. falling distinction within the boundary tone domain, and even syntactic differences when available.

In this regard, it can be argued that the difference found in the perceptual **weight** of auditory and visual information in these two chapters is especially linked to the saliency expressed by these cues. For instance, the difference between two types of falling tones (even if they show a difference in pitch range) will be less salient than the difference existing between a falling tone and a rising one. In addition, the difference between a raised eyebrow and a furrowed brow will be more salient than the difference between a raised brow and its default configuration.

Interestingly, the results described in Chapter 5 showed that interaction effects such as those existing between acoustic and visual information were found within a single modality when comparing the perception of IFS vs. ISQ. Even though in our Dutch materials there were no ISQ were produced that did not manifest subject/verb inversion, Dutch participants significantly classified SV utterances with a final rising intonation as being exemplars of questions. What is also important to note, though, is that this preference for 'question' responses had no effect when rising intonation co-occurred with a (syntactically marked) VS structure, which also suggests a kind of **hierarchical weight** of the available cues which plays a role in the detection of interrogativity. Finally, the same result was obtained when comparing eye gaze with intonation both in Dutch and Catalan, namely, the presence of gaze significantly increased participants' 'question' responses only when gaze co-occurred with a falling contour.

This result is in line with recent investigations on the role of verbal and nonverbal cues as response-mobilizing features using corpus analysis. Stivers & Rossano (2010) stated that "a request (or an offer or information request) is high in response relevance, but a request designed 'directly' (e.g., with interrogative morphosyntax and/or prosody) would be still higher [in response relevance]. Similarly, an assessment (or a noticing or announcement) would be low in response relevance. However, if it were designed with multiple response-mobilizing features, this would increase the response relevance of the action" (Stivers & Rossano 2010: 27–28).

This principle of response relevance takes into consideration the role location of "interrogative morphosyntax and/or prosody" at a higher rank in the hierarchy, but also takes into consideration the **incremental** effect of other available cues. Stivers & Rossano (2010) found for both English and Italian that no single feature is present in all cases and thus concluded that no feature appeared to be intrinsic to the action of requesting information. Moreover, they stated that the use of a number of response-mobilizing features increases the response relevance of an action. In fact, when analyzing the AV perception results in our Chapter 5, we found a positive correlation between the concentration of interrogative cues in a sentence and its rating as an interrogative utterance, for both languages. This pattern of results suggests — at least when taking into account in terms of the data for Dutch and Catalan — that there exists a cue value scale for interrogativity marking such that syntax > intonation > gaze (eyebrow).

In sum, the present thesis has provided results that are relevant for the issue of the interaction between auditory and facial cues in speakers' perception of an utterance as being a statement or a question, which I suggest can ultimately be linked to concepts such as *hierarchical weight*. The results presented here allow for a better understanding of human communication and the role that facial gestures and intonational features — especially pitch range — play in this system.

References

Aaltonen, O., Eerola, O., Hellström, Å., Uusipaikka, E., & Lang, A. H. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *Journal of the Acoustical Society of America*, 101(2), 1090–1105.

Abramson, A. S. (1979). The noncategorical perception of tone categories in Thai. In B. Lindblom & S. Öhman (Eds.), *Frontiers of speech communication research*. London: Academic Press, 127–134.

Aguilar, L., De-la-Mota, C., & Prieto, P. (Coords.) (2009). *Cat_ToBI Training Materials*. Web page. http://prosodia.upf.edu/cat_tobi/

Ahmad, M. I., Tariq, H., Saeed, M., Shahid, S., & Krahmer, E. (2011). Guess Who? An interactive and entertaining game-like platform for investigating human emotions. *Human-computer interaction. Towards mobile and intelligent interaction environments. Lecture Notes in Computer Science*, 6763, 543–551.

Alho, K. (1995). Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. *Ear and Hearing*, 16(1), 38–51.

Amades, J. (1957). El gest a Catalunya. *Anales del Instituto de Lingüística*, VI, 88–148.

Antunes, F. M., Nelken, I., Covey, E., & Malmierca, M. S. (2010). Stimulus-specific adaptation in the auditory thalamus of the anesthetized rat. *PLoS one*, 5(11), e14071.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Armstrong, M. E. (2012). *The development of yes-no question intonation in Puerto Rican Spanish*. PhD dissertation. Columbus: The Ohio State University.

Arvaniti, A., & Baltazani, M. (2004). Greek ToBI. In S. A. Jun (Ed.), *Prosodic models and transcription: Towards prosodic typology*. Oxford: Oxford University Press, 84–117.

Assmann, P., & Summerfield, Q. (2004). The perception of speech under adverse conditions. In S. Greenberg & W. A. Ainsworth (Eds.), *Speech processing in the auditory system*. New York: Springer Verlag, 231–308.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Baker-Shenk, C. L. (1983). *A microanalysis of the nonmanual components of questions in American Sign Language*. Berkeley: University of California.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2005). Problem detection in human-machine interactions based on facial expressions of users. *Speech Communication*, 45(3), 343–359.

Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255–310.

Beckman, M. E., Díaz-Campos, M., McGory, J. T., & Morgan, T. A. (2002). Intonation across Spanish, in the Tones and Break Indices framework. *Probus*, 14, 9–36.

Beckman, Mary E. & Gayle Ayers Elam. 1997. Guidelines for ToBI labeling (version 3). Manuscript. Ohio State University.

Bergman, B. (1984). Non-manual components of signed language: Some sentence types in Swedish Sign Language. In F. Loncke, P. Boyes Braem & Y. Lebrun (Eds.), *Recent research on European sign languages (Proceedings of the European Meeting of Sign Language Research, Brussels)*. Lisse: Swets & Zeitlinger, 49–59.

Beskow, J., Granström, B., & House, D. (2006). Visual correlates to prominence in several expressive modes. *Proceedings of Interspeech 2006* (Pittsburgh), 1272–1275.

Billmyer, K., & Varghese, M. (2000). Investigating instrument-based pragmatic variability: Effects of enhancing Discourse Completion Tests. *Applied Linguistics*, 21(4), 517–52.

Boersma, P., & Weenink, D. (2008). *Praat: doing phonetics by computer* (version 5.0.09). Computer Program.

Bolinger, D. L. (1986). *Intonation and its uses: Melody in grammar and discourse*. Palo Alto: Stanford University Press.

Borràs-Comes, J., Puglesi, C., & Prieto, P. (2011). Audiovisual competition in the perception of counter-expectational questions. In G. Salvi, J. Beskow, O. Engwall & S. Al Moubayed (Eds.), *Proceedings of the 11th International Conference on Auditory-Visual Speech Processing 2011* (Volterra, Italy), 43–46.

Borràs-Comes, J., Vanrell, M. M., & Prieto, P. (2010). The role of pitch range in establishing intonational contrasts in Catalan. *Proceedings of the Fifth International Conference on Speech Prosody* (Chicago), 100103, 1–4.

Braun, B. (2006). Phonetics and phonology of thematic contrast in German. *Language and Speech*, 49(4), 451–493.

Breeuwer, M. & Plomp, R. (1984). Speechreading supplemented with frequency-selective soundpressure information. *Journal of the Acoustical Society of America* 76(3), 686–691.

Burnham, D. (1998). Language specificity in the development of auditory–visual speech perception. In R. Campbell, B. Dodd & D. Burnham (Eds.), *Hearing by eye II: advances in the psychology of speechreading and auditory-visual speech*. New York: Psychology Press, 29–60.

Calhoun, S. (2004). Phonetic dimensions of intonational categories: The case of L+H* and H*. In B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2004*, 103–106. Nara (Japan).

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of Cognitive Neuroscience*, 15(1), 57–70.

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., & Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In T. Bunnell & W. Idsardi (Eds.), *Proceedings of the Fourth International Conference on Spoken Language Processing* (Philadelphia), 2175–2179.

Chandrasekaran, B., Krishnan, A., & Gandour, J. (2007). Mismatch negativity to pitch contours is influenced by language experience. *Brain Research*, 1128(1), 148–156.

Chandrasekaran, B., Krishnan, A., & Gandour, J. (2009). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain and Language*, 108(1), 1–9.

Chen, A. (2003). Reaction time as an indicator of discrete intonational contrasts in English. *Proceedings of the Eighth European Conference on Speech Communication and Technology* (Geneva), 97–100.

Coerts, J. (1992). *Nonmanual grammatical markers. An analysis of interrogatives, negations and topicalisations in Sign Language of the Netherlands*. Amsterdam: Universiteit van Amsterdam.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Cohen, A. (1996). Investigating the production of speech act sets. In S. M. Gass & J. Neu (Eds.), *Speech Acts across cultures*. Berlin, Germany: Mouton de Gruyter, 23–43.

Cohen, A. (2007). Incredulity questions. In R. Artstein & L. Vieu (Eds.), *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue* (Rovereto), 133–140.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology*, 113(4), 495–506.

Cosnier, J. (1991). Les gestes de la question. In C. Kerbrat-Orecchioni (Dir.), *La question*. Lyon: Presses Universitaires de Lyon, 163–171.

Crespo-Sendra, V. (2011). *Aspectes de l'entonació del valencià*. PhD dissertation. Barcelona: Universitat Pompeu Fabra.

Cruschina, S. (2011). Focalization and word order in Old Italo-Romance. *Catalan Journal of Linguistics*, 10, 92–135.

Cruttenden, A. (1981). Falls and rises: meanings and universals. *Journal of Linguistics*, 17(1), 77–91.

de Vos, C., van der Kooij, E., & Crasborn, O. (2009). Mixed Signals: Combining linguistic and affective functions of eyebrows in questions in Sign Language of the Netherlands. *Language and Speech*, 52, 315–339.

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport*, 8, 919–924.

Deouell, L. Y. (2007). The frontal generator of the Mismatch Negativity revisited. *Journal of Psychophysiology*, 21(3-4), 188–203.

Dijkstra, C., Krahmer, E., & Swerts, M. (2006). Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. In R. Hoffmann & H. Mixdorff (Eds.), *Proceedings of the Third International Conference on Speech Prosody* (Dresden). Dresden: TUDpress.

Dilley, L. C. (2010). Pitch range variation in English tonal contrasts: Continuous or categorical? *Phonetica*, 67, 63–81.

Dohen, M., & Lœvenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech*, 52(2–3), 177–206.

Dohen, M., Lœvenbruck, H., & Hill, H. (2006). Visual correlates of prosodic contrastive focus in French: Description and inter-speaker variabilities. In R. Hoffmann & H. Mixdorff (Eds.), *Proceedings of the Third International Conference on Speech Prosody* (Dresden). Dresden: TUDpress, 221–224.

Doherty, C. P., West, W. C., Dilley, L. C., Shattuck-Hufnagel, S., & Caplan, D. (2004). Question/statement judgments: an fMRI study of intonation processing. *Human Brain Mapping*, 23, 85–98.

Dryer, M. S. (2008). Polar questions. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The World Atlas of Language Structures Online* (chapter 116). Munich: Max Planck Digital Library. http://wals.info/feature/116

Duncan, S., & Fiske, D. W. (1977). *Face-to-Face Interaction: Research, Methods, and Theory*. New York: Wiley.

Ekman, P., & Friesen, W. V. (1978). *The Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *The Facial Action Coding System CD-ROM*. Salt Lake City: Research Nexus.

Englert, C. (2010). Questions and responses in Dutch conversations. *Journal of Pragmatics*, 42(10), 2666–2684.

Escera, C., Alho, K., Schröger, E., & Winkler, I. (2000). Involuntary attention and distractibility as evaluated with event-related brain potentials. *Audiology & Neuro-Otology*, 5(3-4), 151–166.

Estebas-Vilaplana, E. (2009). *The use and realization of accentual focus in Central Catalan with a comparison to English*. Munich: Lincom Europa.

Estebas-Vilaplana, E., & Prieto, P. (2010). Castilian Spanish intonation. In P. Prieto & P. Roseano (Eds.), *Transcription of intonation of the Spanish language*. München: Lincom Europa, 17–48.

Eulitz, C., & Lahiri, A. (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of Cognitive Neuroscience*, 16(4), 577–583.

Face, T. L. (2005). F0 peak height and the perception of sentence type in Castilian Spanish. *Revista de Lingüística Iberoamericana*, 2(6), 49–65.

Face, T. L. (2007). The role of intonational cues in the perception of declaratives and absolute interrogatives in Castilian Spanish. *Estudios de Fonética Experimental*, 16, 185–225.

Face, T. L. (2011). *Perception of Castilian Spanish Intonation: Implications for Intonational Phonology*. Munich: Lincom Europa.

Face, T. L., & D'Imperio, M. (2005). Reconsidering a focal typology: Evidence from Spanish and Italian. *Italian Journal of Linguistics*, 17(2), 271–289.

Falé, I., & Hub Faria, I. (2005). Intonational contrasts in EP: a categorical perception approach. *Proceedings of the Ninth*

European Conference on Speech Communication and Technology (Lisboa), 1705–1708.

Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, 52, 542–554.

Fournier, R., Gussenhoven, C., Jensen, O., & Hagoort, P. (2010). Lateralization of tonal and intonational pitch processing: an MEG study. *Brain Research*, 1328, 79–88.

Francis, A. L., Ciocca, V., & Ng, B. K. C. (2003). On the (non)categorical perception of lexical tones. *Perception & Psychophysics*, 65(7), 1029–1044.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502), 312–316.

Gandour, J., Dechongkit, S., Ponglorpisit, S., & Khunadorn, F. (1994). Speech timing at the sentence level in Thai after unilateral brain damage. *Brain and Language*, 46(3), 419–438.

Geluykens, R. (1988). On the myth of rising intonation in polar questions. *Journal of Pragmatics*, 12, 467–485.

Golato, A. (2006). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics*, 24(1), 90–121.

Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition* (Washington), 396–401.

Grant, K. W., & Walden, B. E. (1996). Spectral distribution of prosodic information. *Journal of Speech and Hearing Research*, 39, 228–238.

Grant, K. W., Walden, B. E., &, Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America*, 103, 2677–2690.

Grice, M., D'Imperio, M., Savino, M., & Avesani, C. (2005). Towards a strategy for labeling varieties of Italian. In S. A. Jun (Ed.), *Prosodic models and transcription: Towards prosodic typology*. Oxford: Oxford University Press, 55–83.

Grimm, S., & Escera, C. (2012). Auditory deviance detection revisited: Evidence for a hierarchical novelty system. *International Journal of Psychophysiology*, 85(1), 88–92.

Grimm, S., Escera, C., Slabu, L., & Costa-Faidella, J. (2011). Electrophysiological evidence for the hierarchical organization of auditory change detection in the human brain. *Psychophysiology*, 48(3), 377–384.

Grossman, R. B. (2001). *Dynamic facial expressions in American Sign Language: Behavioral, neuroimaging, and facial coding analyses for deaf and hearing participants*. PhD dissertation. Boston: Boston University.

Grossman, R. B., & Kegl, J. (2006). To capture a face: A novel technique for the analysis and quantification of facial expressions in American Sign Language. *Sign Language Studies*, 6(3), 273–305.

Gussenhoven, C. (1999). Discreteness and gradience in intonational contrasts. *Language and Speech*, 42(2-3), 283–305.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Gussenhoven, C. (2007). Types of Focus in English. In C. Lee, M. Gordon & D. Büring (Eds.), *Topic and focus: Cross-linguistic perspectives on meaning and intonation*. Heidelberg/New York/London: Springer, 83–100.

Gussenhoven, C. (in press). Transcription of Dutch intonation. In S.-A. Jun (Ed.), *Prosodic Typology 2. The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Haan, J. (2002). *Speaking of questions. An exploration of Dutch question intonation*. *LOT Dissertation* Series 52. Utrecht: LOT.

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1983). Head movements correlates of juncture and stress at sentence level. *Language and Speech*, 26, 117–129.

Hauk, O., Shtyrov, Y., & Pulvermüller, F. (2006). The sound of actions as reflected by mismatch negativity: rapid activation of cortical sensory-motor networks by sounds associated with finger and tongue movements. *The European Journal of Neuroscience*, 23(3), 811–821.

Hirschberg, J., & Ward, G. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics*, 20, 241–251.

Holt, H. H., Lotto, A. J., Diehl, R. L. (2004). Auditory discontinuities interact with categorization: implications for speech perception. *Journal of the Acoustical Society of America*, 116(3), 1763–1773.

House, D. (2002). Perception of question intonation and facial gestures. *Fonetik*, 44(1), 41–44.

IBM Corporation (2010). *IBM SPSS Statistics* (version 19.0.0). Computer Program.

Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition*, 18(21), 49–69.

Keating, P. A. (2004). Statistics. Manuscript. UCLA Phonetics Lab. Web page. http://www.linguistics.ucla.edu/faciliti/facilities/statistics/statistics.html

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26, 22–63.

Kendon, A. (1990). *Conducting interaction: patterns of behavior in focused encounters*. New York: Cambridge University Press.

Klein, D., Zatorre, R., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *NeuroImage*, 13, 646–653.

Krahmer, E., & Swerts, M. (2004). More about brows: a cross-linguistic analysis-by-synthesis study. In Z. Ruttkay & C. Pelachaud (Eds.), *From brows to trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers, 191–216.

Krahmer, E., & Swerts, M. (2005). How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech*, 48(1), 29–54.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396–414.

Krahmer, E., Ruttkay, Z., Swerts, M., & Wesselink, W. (2002). Pitch, eyebrows and the perception of focus. In B. Bel & I. Marlien (Eds.), *Proceedings of the First International Conference on Speech Prosody*, Aix en Provence.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. *Science*, 190(4209), 69–72.

Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, 63(3), 905–917.

Kyle, J. G., & Woll, B. (1985). *Sign language: The study of deaf people and their language*. Cambridge: Cambridge University Press.

Labov, W., & Fanshel, D. (1977). *Therapeutic Discourse*. New York: Academic Press.

Ladd, D. R. (1994). Constraints on the gradient variability of pitch range, or, pitch level 4 lives! In P. Keating (Ed.), *Phonological structure and phonetic form. Papers in Laboratory Phonology III*. Cambridge: Cambridge University Press, 43–63.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics*, 25, 313–342.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–74.

Lapakko, D. (1997). Three cheers for language: A closer examination of a widely cited study of nonverbal communication. *Communication Education*, 46, 63–67.

Lee, H.-Y. (2004). H and L are not enough in intonational phonology. *Eoneohag*, 39, 71-79.

Leitman, D., Sehatpour, P., Shpaner, M., Foxe, J., & Javitt, D. (2009). Mismatch negativity to tonal contours suggests preattentive perception of prosodic content. *Brain Imaging and Behavior*, 3, 284–291.

Levinson, S. C. (2010). Questions and responses in Yélî Dnye, the Papuan language of Rossel Island. *Journal of Pragmatics*, 42(10), 2741–2755.

Liberman, M. Y., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. T. Oehrle (Eds.), *Language Sound Structure. Studies in phonology presented to Morris Halle*. Cambridge: MIT Press, 157–233.

Litman, D. & Forbes-Riley, K. (2009). Spoken tutorial dialogue and the Feeling of Another's Knowing. *Proceedings of the Tenth Annual Meeting of the Special Interest Group in Discourse and Dialogue* (London), 286–289.

Lysander, K., & Horton, W. S. (2012). Conversational grounding in younger and older adults: the effect of partner visibility and referent abstractness in task-oriented dialogue. *Discourse Processes*, 49(1), 29–60.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception and Psychophysics*, 24(3), 253–257.

Malmierca, M. S., Cristaudo, S., Perez-Gonzalez, D., & Covey E. (2009). Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *The Journal of Neuroscience*, 29(17), 5483–5493.

Mascaró, J. (1978). *Expresión y comunicación no verbal. Metodología y crítica*. PhD dissertation. Barcelona: Universitat de Barcelona.

Mascaró, J. (1981). Notes per a un estudi de la gestualitat catalana. *Serra d'Or*, 259, 25–28.

Massaro, D. W. (1987). *Speech perception by ear and by eye*. Hillsdale: Erlbaum.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge: MIT Press.

Massaro, D. W., & Cohen, M. M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, 122(1), 115–124.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices: A new illusion. *Nature*, 264, 746–748.

McHugh, B. D. (1990). The phrasal cycle in Kivunjo Chaga tonology. In S. Inkelas & D. Zec (Eds.), *The phonology-syntax connection*. Chicago: University of Chicago Press, 217–242.

Mehrabian, A., & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31, 248–252.

Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33–35.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. Head movement improves auditory speech perception. *Psychological Science*, 15, 133–137.

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, 38, 1–21.

Näätänen, R., & Michie, P. T. (1979). Early selective-attention effects on the evoked potential: a critical review and reinterpretation. *Biological Psychology*, 8(2), 81–136.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour-Luhtanen, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., Allik, J., Sinkkonen, J., & Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385(6615), 432–434.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, 118, 2544–2590.

Nakano, Y. I., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (Sapporo), 553–561.

Nurani, L. M. (2009). Methodological issue in pragmatic research: is discourse completion test a reliable data collection instrument? *Jurnal Sosioteknologi*, 17(8), 667–678.

(The) Ohio State University Department of Linguistics (1999). ToBI. Web page. http://www.ling.ohio-state.edu/~tobi/

Payrató, L. (1989). *Assaig de dialectologia gestual. Aproximació pragmàtica al repertori bàsic d'emblemes del català de Barcelona*. PhD dissertation. Barcelona: Universitat de Barcelona.

Payrató, L. (1993). A pragmatic view on autonomous gestures: A first repertoire of Catalan emblems. *Journal of Pragmatics*, 20, 193–216.

Payrató, L., Alturo, N., & Payà, M. (Eds.) (2004). *Les fronteres del llenguatge. Lingüística i comunicació no verbal*. Barcelona: Promociones y Publicaciones Universitarias.

Pérez-González, D., Malmierca, M. S., & Covey, E. (2005). Novelty detector neurons in the mammalian auditory midbrain. *The European Journal of Neuroscience*, 22(11), 2879–2885.

Peters, C., Pelachaud, C., Bevacqua, E., Mancini, M., & Poggi, I. (2005). A model of attention and interest using gaze behavior. *Proceedings of Intelligent Virtual Agents 2005* (Kos), 229–240.

Pfau, R., & Quer, J. (2010). Nonmanuals: their grammatical and prosodic roles. In D. Brentari (Ed.), *Sign Languages* (*Cambridge Language Surveys*). Cambridge: Cambridge University Press, 381–402.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M., & Roberts, T. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. *Journal of Cognitive Neuroscience*, 12(6), 1038–1055.

Pierrehumbert, J. (1980). *The Phonetics and Phonology of English Intonation*. PhD dissertation. Cambridge: Massachusetts Institute of Technology.

Pomerantz, A. M. (1980). Telling my side: "Limited access" as a "fishing" device. *Sociological Inquiry*, 50, 186–198.

Post, B. Stamatakis, E., Bohr, I., Nolan, F. & Cummins, C. (in press). Categories and gradience in intonation: an fMRI study. In J. Romero & M. Riera (Eds.), *Phonetics and Phonology in Iberia*.

Prieto, P. (2002). Entonació. In J. Solà, M. R. Lloret, J. Mascaró & M. Pérez Saldanya (Eds.), *Gramàtica del català contemporani*, vol. 1. Barcelona: Empúries, 1395–1462.

Prieto, P. (2004). The search for phonological targets in the tonal space: H1 scaling and alignment in five sentence-types in Peninsular Spanish. In T. L. Face (Ed.), *Laboratory approaches to Spanish phonology*. Berlin: Mouton de Gruyter, 29–59.

Prieto, P. (2005). Stability effects in tonal clash contexts in Catalan. *Journal of Phonetics*, 33(2), 215–242.

Prieto, P. (in press). The intonational phonology of Catalan. In S.-A. Jun (Ed.), *Prosodic Typology 2. The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press.

Prieto, P., & Cabré, T. (Coords.) (2007-2012). *Atles interactiu de l'entonació del català*. Web page. http://prosodia.upf.edu/atlesentonacio/

Prieto, P., & Rigau, G. (2007). The syntax-prosody interface: Catalan interrogative sentences headed by *que*. *Journal of Portuguese Linguistics*, 6(2), 29–59.

Prieto, P., Aguilar, L., Mascaró, I., Torres-Tamarit, F., & Vanrell, M. M. (2009). L'etiquetatge prosòdic Cat_ToBI. *Estudios de Fonética Experimental*, 18, 287–309.

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., & Blat, J. (2011). Crossmodal prosodic and gestural contribution to the perception of contrastive focus. *Proceedings of the 12th Annual Conference of the International Speech Communication Association* (Florence), 977–980.

Psychology Software Tools Inc. (2009). *E-Prime* (version 2.0). Computer Program.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6, 576–582.

Pulvermüller, F., & Shtyrov, Y. (2006). Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes. *Progress in Neurobiology*, 79(1), 49–71.

Quené, H., & van der Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59, 413–425.

Raizada, R. D., & Poldrack, R. A. (2007). Selective amplification of stimulus differences during categorical processing of speech. *Neuron*, 56(4), 726-740.

Rathcke, T., & Harrington, J. (2010). The variability of early accent peaks in Standard German. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory Phonology 10*, 533–555. Berlin – New York: Mouton de Gruyter.

Ren, G.-Q., Yang, Y., & Li, X. (2009). Early cortical processing of linguistic pitch patterns as revealed by the mismatch negativity. *Neuroscience*, 162, 87–95.

Rialland, A. (2007). Question prosody: an African perspective. In C. Gussenhoven & C. Riad (Eds.), *Tones and tunes*, vol. 2. Mouton: Berlin, 35–62.

Richardson, D. C., Dale, R., & Tomlinson, J. M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science*, 33(8), 1468–1482.

Roseano, P., Vanrell, M. M., & Prieto, P. (2011). Fri_ToBI. *Workshop on Romance ToBI. Transcription of intonation of the Romance languages* (Tarragona).

Rossano, F. (2010). Questioning and responding in Italian. *Journal of Pragmatics*, 42(10), 2756–2771.

Savino, M., & Grice, M. (2011). The perception of negative bias in Bari Italian questions. In S. Frota, P. Prieto & G. Elordieta (Eds.), *Prosodic categories: production, perception and comprehension*. Springer Verlag, 187–206.

Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech*, 52(2–3), 135–175.

Sharma, A., & Dorman, M. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *Journal of the Acoustical Society of America*, 106(2), 1078–1083.

Sharma, A., & Dorman, M. (2000). Neurophysiologic correlates of cross-language phonetic perception. *Journal of the Acoustical Society of America*, 107(5.1), 2697–2703.

Shtyrov, Y., Hauk, O., & Pulvermüller, F. (2004). Distributed neuronal networks for encoding category-specific semantic information: the mismatch negativity to action words. *The European Journal of Neuroscience*, 19(4), 1083–1092.

Slabu, L., Escera, C., Grimm, S., & Costa-Faidella, J. (2010). Early change detection in humans as revealed by auditory brainstem and middle-latency evoked potentials. *The European Journal of Neuroscience*, 32(5), 859–865.

Solà, J. (1990). L'ordre dels mots en català. Notes pràctiques. In J. Solà (Ed.), *Lingüística i normativa.* Barcelona: Empúries, 91–124.

SourceTec Software Co. (2007). *Sothink SWF Quicker* (version 3.0). Computer Program.

Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice. Distinguishing statements from echoic questions in English. *Language and Speech*, 46(1), 1–22.

Stivers, T. (2010). An overview of the question–response system in American English conversation. *Journal of Pragmatics*, 42(10), 2772–2781.

Stivers, T., & Rossano, F. (2010). Mobilizing Response. *Research on Language and Social Interaction*, 43(1), 1–31.

Stockwell, R. P., Bowen, D. J., & Silva-Fuenzalida, I. (1956). Spanish juncture and intonation. *Language*, 32(4), 641–665.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212–215.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. In V. Bruce, A. Cowey, W. Ellis & D. I. Perrett (Eds.), *Processing the facial image*. Oxford: Oxford University Press, 71–78.

Swerts, M., & Krahmer, E. (2004). Congruent and incongruent audiovisual cues to prominence. *Proceedings of the Second International Conference on Speech Prosody* (Nara), 69–72.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1), 81–94.

Swerts, M., & Krahmer, E. (2008). Facial expressions and prosodic prominence: Comparing modalities and facial areas. *Journal of Phonetics*, 36(2), 219–238.

Trager, G. L., & Smith, H. L. (1951). *An outline of English structure. Studies in Linguistics occasional papers 3*. Norman: Battenberg Press.

Ulanovsky, N., Las, L., & Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, 6(4), 391–398.

Vallduví, E. (1991). The role of plasticity in the association of focus and prominence. In Y. No & M. Libucha (Eds.), *Proceedings of the Seventh Eastern States Conference on Linguistics*. Columbus: Ohio State University Press, 295–306.

Vanrell, M. M. (2006). A tonal scaling contrast in Majorcan Catalan interrogatives. *Journal of Portuguese Linguistics*, 6(1), 147–178.

Vanrell, M. M. (2011). *The phonological relevance of tonal scaling in the intonational grammar of Catalan*. PhD dissertation. Cerdanyola del Vallès: Universitat Autònoma de Barcelona.

Vanrell, M. M., Mascaró, I., Torres-Tamarit, F., & Prieto, P. (in press). Intonation as an encoder of speaker certainty: information and confirmation yes-no questions in Catalan. *Language and Speech*.

Vanrell, M. M., Stella, A., Gili-Fivela, B., & Prieto, P. (2012). Prosodic cues for the recognition of contrastive focus. In B. Gili-Fivela, A. Stella, L. Garrapa & M. Grimaldi (Eds.), *Contesto comunicativo e variabilità nella produzione e percezione della lingua. Atti del 7 Convegno AISV*. Lecce: Bulzoni.

Vilhjálmsson, H. H. (1997). Autonomous communicative behaviors in avatars. Master of Science Thesis. Cambridge: Massachusetts Institute of Technology

Ward, G., & Hirschberg, J. (1985). Implicating uncertainty. The pragmatics of fall-rise intonation. *Language*, 61, 747–776.

Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., Csépe, V., Aaltonen, O., Raimo, I., Alho, K., Lang, A. H., Iivonen, A., & Näätänen, R. (1999). Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations. *Cognitive Brain Research*, 7, 357–369.

Xi, J., Zhang, L., Shu, H., Zhang, Y., & Li, P. (2010). Categorical perception of lexical tones in Chinese revealed by mismatch negativity. *Neuroscience*, 170, 223–231.

Zhang, L., Xi, J., Xu, G., Shu, H., Wang, X., & Li, P. (2011) Cortical dynamics of acoustic and phonological processing in speech perception. *PLoS one*, 6(6): e20963.

Appendix 1

Introducció en català

L'objectiu principal d'aquesta tesi és aprofundir en el coneixement de la interrogativitat. Concretament, es pretén esbrinar com els parlants la marquen i, especialment, com la detecten. És a dir, volem saber quins són els elements que ens permeten diferenciar una oració interrogativa d'una declarativa, tant en la fase de producció de la parla com en la fase de percepció. D'aquesta manera, la motivació central de la tesi és entendre millor un dels aspectes centrals de la comunicació humana: el mecanisme segons el qual sabem si se'ns dóna informació o si se'ns en demana.

És ben sabut que moltes llengües empren l'entonació per marcar la interrogació. No obstant això, encara que una de les funcions principals de l'entonació siga vehicular significats pragmàtics, molts dels estudis sobre entonació l'han descrit sense tindre explícitament en compte aquests contextos pragmàtics. A part, diferents estudis previs sobre l'entonació de les llengües s'han basat en parla llegida i han ignorat molt sovint altres correlats lingüístics que acompanyen l'entonació, com és el cas dels gestos. En aquesta tesi es tracten dos tipus de declaratives i dos tipus d'interrogatives absolutes, que poden ser classificades com a neutres (és a dir, no marcades) i marcades depenent de la manera com vehiculen el seu contingut semàntic.

Pel que fa a les declaratives, distingim entre *declaratives de focus informatiu* (IFS, *information focus statements*) i *declaratives de focus contrastiu* (CFS, *contrastive focus statements*). Les IFS són considerades les declaratives neutres, és a dir, aquelles oracions que vehiculen informació nova i que contenen un constituent que es focalitza respecte al *background*. En canvi, en una CFS es marca un dels constituents com a "desestimació directa d'una alternativa" (Gussenhoven 2007); es corregix "el valor de l'alternativa assignant un valor diferent" (Cruschina 2011). Per tant, la principal diferència entre els dos tipus de focus és que, mentre el CFS depèn de l'asserció prèvia, que és rebutjada o corregida, l'IFS no mostra aquest requisit. Aquest rebuig o correcció s'explicita sovint pels mitjans entonatius i gestuals de la majoria de llengües entonatives.

Pel que fa a les interrogatives, distingim entre *interrogatives de cerca d'informació* (ISQ, *information-seeking questions*) i *interrogatives antiexpectatives* (CEQ, *counter-expectational questions*). D'una banda, una ISQ és aquella interrogativa que té la funció específica d'obtenir informació d'un receptor, sense cap matís especial que indique les expectatives del parlant. D'altra banda, les CEQ estan relacionades amb les interrogatives ecoiques, que són aquelles en què l'oient repetix informació que acaba de sentir per diverses raons possibles, com ara perquè no ho ha sentit bé o no ha entès bé el que se li ha dit o perquè el que implica aquella informació entra en conflicte amb les seues expectatives prèvies. Les CEQ constituïxen aquest darrer tipus, i poden ser marcades amb un matís de sorpresa o d'incredulitat. Tal com establix Cohen (2007: 133), "una interrogativa d'incredulitat expressa la noció que la declaració de què es fa eco no és certa en cap dels mons creguts (o normatius) pels parlants — d'ací la incredulitat (o indignació) expressada cap a aquella declaració" (v. Cohen 2007 per a més distincions entre interrogatives *ecoiques* i *d'incredulitat*). Com en el cas de les CFS, els matisos d'antiexpectació, sorpresa o incredulitat d'una CEQ es marquen sovint amb patrons entonatius i gestuals específics en moltes llengües entonatives.

Per tal d'analitzar els patrons entonatius, hem emprat el sistema de transcripció *Tone and Break Indices* (ToBI), basat en el model Mètric i Autosegmental (AM). De manera breu, aquesta aproximació descriu l'entonació d'una oració distingint entre aquells tons associats a les síl·labes accentuades (*accents tonals*) i aquells altres que s'alineen a la frontera prosòdica de les frases entonatives (*tons de frontera* i *accents de frase*). Les dues unitats bàsiques que conformen els accents tonals i els tons de frontera són els tons H (*high*, 'alt') i L (*low*, 'baix'), interpretats respectivament com un augment o una davallada del to en el transcurs d'una melodia oracional. En la majoria de llengües, els accents tonals estan compostos d'un o dos tons, el més prominent dels quals és marcat amb un asterisc (T*). Els tons de frontera són percebuts generalment com a excursions descendents o ascendents, o com una combinació d'aquestes, i són generalment transcrits amb el símbol del percentatge (T%) o amb un guionet (T–). Pel fet de ser un sistema de transcripció fonològica, el ToBI requerix un coneixement humà expert per tal de caracteritzar els esdeveniments prosòdics de cada llengua, i és per això que s'han desenvolupat diferents sistemes de transcripció ToBI específics per a cada llengua des de l'aparició de la tesi de Pierrehumbert (1980) sobre el sistema entonatiu de l'anglès (v. Ohio State University Department of Linguistics 1999).

Aquesta tesi s'organitza en quatre estudis principals, presentats dels capítols 2 al 5. Primer, analitzem el rol que exercix una propietat específica de l'entonació en la distinció entre declaratives i interrogatives antiexpectatives en català. Aquesta propietat entonativa és el camp tonal, que fa referència a la distància tonal entre els valors d'f0 més baix i més alt observats en l'accent tonal d'una oració (és a dir, una vall i un pic; v. Gussenhoven 2004). La raó per triar el català per analitzar aquest fenomen és que en aquesta llengua, com en d'altres llengües romàniques, un contorn entonatiu nuclear ascendent-descendent —és a dir, un accent tonal ascendent associat amb la darrera síl·laba accentuada d'una oració seguit d'un to de frontera baix— és emprat per vehicular 'IFS', 'CFS' i 'CEQ' depenent de les característiques de camp tonal que presenta. Aquest contrast entonatiu s'analitza en els capítols 2 i 3. Atès que aquests contrastos també es poden expressar mitjançant gestos facials específics, en el capítol 4 analitzem la interacció entre els indicis acústics i visuals en la percepció de la interrogativitat. Així com en els experiments presentats als capítols 2, 3 i 4 s'han comparat declaratives amb interrogatives marcades, en el capítol 5 analitzem com es detecten les interrogatives neutres (ISQ) quan es comparen amb declaratives neutres (IFS).

La Taula 1 mostra un resum dels tipus de declaratives i interrogatives que s'analitzen en aquesta tesi.

Taula 1. Significats oracionals analitzats en aquesta tesi.

declaratives	neutres	declarativa de focus informatiu (IFS)
	marcades	declarativa de focus contrastiu (CFS)
interrogatives	neutres	interrogativa de cerca d'informació (ISQ)
	marcades	interrogativa antiexpectativa (CEQ)

L'objectiu del primer estudi (**capítol 2**) és investigar com es distribuïxen les percepcions d'IFS, CFS i CEQ al llarg d'un contínuum de camp tonal, i si els oients de català empren aquesta distinció en camp tonal per identificar aquests significats. És ben sabut que diferents llengües empren diferents contorns entonatius com a marcadors interrogatius, però en el camp de la fonologia entonativa encara és un tema controvertit si les diferències en camp tonal també s'utilitzen per expressar una distinció categòrica com aquesta. Amb aquest propòsit, vam dur a terme dues tasques experimentals. Primer, vam emprar una tasca d'identificació amb tres possibles opcions de resposta. Així, permetíem la comparació simultània de les tres categories (IFS, CFS i CEQ). En segon lloc, vam emprar una tasca de congruència, la qual fa possible conèixer el grau en què els oients consideren adequat o inadequat l'ús de cadascun d'aquests contorns quan s'inserixen en un determinat context discursiu. En totes dues tasques, l'anàlisi de les respostes d'identificació es complementa amb el de les mesures de temps de reacció, pel fet que aquestes mesures són útils per investigar la categorialitat d'una diferència en entonació. Malgrat que la diferència percebuda entre els dos tipus de declaratives no pot ser explicada exclusivament per diferències de camp tonal, els resultats d'aquest primer estudi mostren un contrast fonològic clar entre 'IFS' i 'CEQ'.

Així, atesos els resultats del capítol 2, el **capítol 3** examina la percepció del contrast entre 'IFS' i 'CEQ' amb una exploració electrofisiològica de l'activitat cerebral. Diferents estudis han indicat que els correlats i tonals que marquen distincions lèxiques poden ser representades en la memòria sensorial auditiva preatencional emprant el potencial evocat cerebral (ERP, *event-related potential*) de la negativitat de disparitat (MMN, *mismatch negativity*). En aquest estudi examinem si els contrastos intracategorials i intercategorials entre IFS i CEQ d'una llengua entonativa com el català també provoquen patrons d'activitat neurofisiològica diferents, la qual cosa indica la codificació automàtica d'aquests contrastos entonatius en el còrtex auditiu. A més, aquest resultat evidencia que el processament cerebral dels contrastos entonatius funciona de manera similar al dels contrastos segmentals.

Com que les declaratives i interrogatives es produïen en la comunicació cara a cara, poden anar associades amb uns gestos facials determinats, com ara moviments del cap i de les celles. En el nostre tercer estudi (**capítol 4**) analitzem una altra qüestió irresolta en el camp de la prosòdia audiovisual: com interactuen els indicis acústics i visuals en la percepció d'aquesta diferència pragmàtica. Encara que la majoria d'estudis sobre prosòdia audiovisual han descrit un mode complementari de processament en el qual la visió proporciona una informació feble i relativament redundant si es compara amb la que proporcionen els indicis acústics, d'altres treballs troben el patró invers. En aquest capítol prenem en consideració el camp tonal i els gestos facials en la distinció entre 'CFS' i 'CEQ'. Vam sintetitzar diferents realitzacions acústiques i gestuals d'aquests dos significats mitjançant una tècnica de transformació digital d'imatges. Després, els participants van realitzar dues tasques d'identificació multimodal en què se'ls presentaven combinacions congruents i incongruents dels materials audiovisuals i se'ls demanava que els classifiquessen com a realitzacions possibles de 'CFS' i 'CEQ'.

En el nostre darrer estudi (**capítol 5**) aprofundim en l'anàlisi de la percepció audiovisual de la interrogativitat, però aquest cop comparant declaratives de focus informatiu (IFS) i interrogatives de cerca d'informació (ISQ), les quals representen els tipus neutres d'ambdós significats pragmàtics. Vam utilitzar un entorn natural per obtenir una sèrie de declaratives i interrogatives, el joc *Qui és qui?*. Basats en aquests materials, els participants van realitzar unes tasques d'identificació unimodal i multimodal (aquest cop només emprant combinacions audiovisuals congruents). Aquesta metodologia ens permet conèixer com es vehicula la interrogativitat tant en la producció com en la percepció de la parla. Aquesta investigació compara les estratègies emprades per part de parlants de català i de neerlandès. Mentre ambdues llengües empren l'entonació per al marcatge de la interrogativitat, el neerlandès també compta amb la inversió sintàctica per a tal propòsit, raó per la qual comparem neerlandès i català en aquest estudi. Aquesta tasca ens permet avaluar si els participants d'ambdues llengües poden diferenciar declaratives i interrogatives neutres unimodalment i multimodalment, així com identificar quins són els elements acústics i gestuals més emprats per marcar aquesta distinció en la producció i en la percepció (la inversió sintàctica quan està disponible, contorns entonatius ascendents, presència de mirada, aixecament de celles), i si aquestes estratègies interactuen en el procés d'identificació d'una oració com a interrogativa que fan els participants.

Una característica que cal subratllar de la nostra metodologia és l'enfocament multimodal de l'estudi de la interrogativitat. Molts estudis tradicionals han obviat el component no verbal de la distinció declarativa/interrogativa i s'han centrat principalment en els marcatges sintàctic, morfològic i entonatiu. Hi ha també poca recerca que tinga en compte més d'una estratègia alhora i que explique la seua interacció potencial com a propietats mobilitzadores de resposta (v. Stivers & Rossano 2010).

Una segona característica que cal emfasitzar és la varietat de metodologies experimentals utilitzades al llarg de la tesi, que tenia l’objectiu d’assegurar la ‘validesa ecològica’ dels resultats. Pel que fa als experiments de producció, hem recollit dades a través de Tests de Compleció de Discurs, àmpliament emprats en la recerca en pragmàtica (Kasper & Dahl 1991, Cohen 1996, Billmyer & Varghese 2000, Golato 2006, Nurani 2009) i jocs, com el *Qui és qui?*, específicament adaptats per obtenir produccions espontànies de determinades categories discursives (Ahmad et al. 2011). Pel que fa als experiments de percepció, hem emprat diferents proves conductuals, d’identificació i de congruència (unimodals o multimodals, binomials o multinomials), de les quals hem analitzat tant les respostes com els temps de reacció, i hem realitzat una exploració electrofisiològica a través de potencials evocats cerebrals mitjançant un paradigma de disparitat (Näätänen 2001).

Appendix 2

Discussió general i conclusions en català

6.1. L'estatus fonològic del camp tonal

Un dels objectius principals de la tesi era descriure el paper del camp tonal en l'expressió de la interrogativitat. En català, la mateixa seqüència de tons baixos i alts en el si d'una configuració nuclear tonal pot expressar tres significats pragmàtics diferents depenent de les característiques del seu camp tonal: una declarativa de focus informatiu (IFS), una declarativa de focus contrastiu (CFS), i una interrogativa antiexpectativa (CEQ). Atès aquest triple contrast, s'han dut a terme diferents experiments per esbrinar si la diferència existent entre aquests tres significats es vehicula de manera categòrica mitjançant el camp tonal.

La investigació del paper del camp tonal s'emmarca en el model Mètric i Autosegmental de l'anàlisi de la prosòdia, que pren com a assumpció central que només es necessiten dos tons, baix (L) i alt (H), per distingir categories entonatives en una llengua com l'anglès. En aquest sentit, el paper del camp tonal ha estat sovint relegat a l'expressió de diferències d'èmfasi o prominència (Pierrehumbert 1980, Beckman & Pierrehumbert 1986). Tanmateix, diferents treballs sobre llengües romàniques i germàniques han demostrat que la variació en camp tonal pot expressar diferències categòriques de significat (Hirschberg & Ward 1992, Ladd & Morton 1997, Savino & Grice 2011, Vanrell 2011), i alguns autors han suggerit que l'enfocament mètric i autosegmental ha de marcar explícitament aquesta propietat en els sistemes de transcripció prosòdica fonològica (Ladd 1994, Face 2011). Els capítols 2 i 3 s'han dedicat a investigar el paper fonològic del camp tonal en català.

El capítol 2 ha presentat dos experiments conductuals en què els participants havien de decidir entre tres possibles respostes

(IFS, CFS, CEQ) quan se'ls presentava amb una sèrie d'estímuls que pertanyien a un contínuum acústic de camp tonal. Hem analitzat les respostes i els temps de reacció d'aquests dos experiments. En el primer experiment, els participants havien d'identificar quin significat atribuïen a cada estímul presentat aïlladament; en el segon, havien d'avaluar el grau de congruència o adequació percebut per a cada estímul quan es presentava en un context discursiu típic per a cadascun dels tres significats possibles. En ambdós experiments, els participants van associar IFS i CEQ amb els extrems inicial i final del contínuum de camp tonal, respectivament, mentre que el CFS va ser associat menys nítidament a un camp tonal específic i va ser percebut de manera semblant a l'IFS.

Pel que fa als patrons de temps de reacció, el primer experiment va mostrar un pic clar en la frontera acústica percebuda entre 'CEQ' i els altres dos tipus d'oracions declaratives (IFS i CFS). En canvi, en el segon experiment només es va obtenir un pic de temps de reacció per a 'IFS' i 'CEQ', però no per a 'CFS'. Seguint Chen (2003), si un pic de temps de reacció localitzat a una frontera d'identificació indica la que un contrast és caregorial, no podem defensar que les decisions dels participants sobre l'adequació i la inadequació de les oracions CFS presentades en context estiguen distribuïdes categòricament pel que fa al camp tonal.

Per tant, els resultats del capítol 2 demostren que els contorns que presenten un camp tonal induïxen interpretacions IFS, mentre que els contorns amb un camp tonal més ampli comporten interpretacions CEQ. Pel que fa al CFS, en canvi, es mostra com aquest significat es comporta aproximadament com l'IFS pel que fa als valors del camp tonal. L'experiment de congruència mostra que no hi ha un pic de temps de reacció entre les respostes 'adequat' i 'inadequat' que es van donar per al context de CFS, la qual cosa significa que aquestes dues respostes no estan dividides categòricament per part dels oients catalans, i indica, de retruc, que el paper del camp tonal en el marcatge del CFS és més aviat un fenomen gradual. El comportament semblant a l'IFS i l'absència d'un pic d'RT pot ser, aleshores, interpretat de la manera següent: el camp tonal distingix un CFS d'un IFS de manera gradual. Defensem que la detecció d'una oració com a representant d'un CFS pot estar relacionada en una major mesura amb un procés d'inferència pragmàtica, tal que l'oient entén CFS quan, en una conversa normal, s'ha afegit informació que contrasta amb la informació precedent. Finalment, el parlant pot marcar l'estatus correctiu d'aquella oració amb estratègies morfosintàctiques com la dislocació i la compressió tonal postfocal.

El capítol 3 ha presentat dos experiments que pretenien mostrar que la categorialitat percebuda entre IFS i CEQ —segons els resultats del capítol 2— té un correlat electrofisiològic. Els estudis electrofisiològics previs sobre contrastos fonològics segmentals i contrastos tonals (aquells provinents de llengües tonals) han evidenciat que els contrastos fonològics existents en la llengua nativa provoquen respostes MMN significativament més grans que els mateixos contrastos no nadius (Näätänen et al. 1997, Gandour 1994). Alhora, també han mostrat que els contrastos acústics que traspassen una frontera entre categories comporten respostes MMN més grans que aquells que no creuen aquestes fronteres (Dehaene-Lambertz 1997, Chandrasekaran et al. 2007). Aquests resultats no havien estat obtinguts per als contrastos entonatius fins ara. Doherty et al. (2004) i Leitman et al. (2009) defensaven que el MMN més gran elicitat pels estímuls interrogatius (i no pas pels estímuls declaratius) "podria estar demostrant l'habilitat de les oracions interrogatives de captar automàticament l'atenció fins i tot quan la informació declarativa precedent ha estat ignorada" (Leitman et al. 2009: 289). Fournier et al. (2010) defensava, a part, que el reconeixement de significats discursius mitjançant l'entonació no era necessàriament clar observant el cervell humà.

Primer, els resultats presentats en el capítol 3 repliquen els resultats del capítol 2. En un primer experiment d'identificació, es va trobar una identificació clarament no monotònica del contrast entre IFS i CEQ, així com temps de reacció més ràpids per a la identificació d'exemplars intracategorials que per a exemplars interpretats de manera més ambigua. En el segon experiment del capítol 3, es troba una amplitud mitjana d'MMN més gran per al contrast intercategorial que per als intracategorials. Amb això, se suggerix que els contrastos entonatius de la llengua poden ser codificats automàticament en el còrtex auditiu. A més, els nostres resultats mostren que l'activació d'aquestes representacions entonatives del còrtex auditiu està relacionada amb la percepció i l'actuació subjectiva dels individus (és a dir, que s'obtenia una correlació significativa entre les respostes electrofisiològiques i les mesures conductuals obtingudes en el primer experiment, tant per individus com pel que fa a la mitjana general de les dades). Així, els nostres resultats proporcionen evidència electrofisiològica que els contrastos fonològics entonatius (basats en una diferència de camp tonal) també són codificats en el còrtex auditiu, la qual cosa es relaciona amb un conjunt de resultats empírics que demostren una activació més gran de traces de memòria per a elements lingüístics en el cervell humà.

Els capítols 2 i 3 mostren que la variació en camp tonal és l'indici principal que els oients de català empren per discriminar entre IFS i CEQ, és a dir, que hi ha un límit al llarg d'un contínuum de camp tonal per damunt del qual interpretem consistentment un significat CEQ. Aquest contrast en camp tonal entre interrogatives i declaratives també s'ha documentat per a altres llengües romàniques (Savino & Grice 2011 per a l'italià de Bari, Roseano et al. 2011 per al friülà, Estebas-Vilaplana & Prieto 2010 per al castellà peninsular, etc.) així com per a llengües no romàniques.

Aquests resultats indiquen que un sistema fiable per a la transcripció prosòdica d'aquestes llengües —almenys per al català— ha de poder assenyalar la distinció entre els patrons IFS (L+H*) i els patrons CEQ (L+¡H*) (Aguilar et al. 2009 per al català). En aquest sentit, i seguint el treball recent de Vanrell (2011), es proposa la inclusió d'un to com [L+¡H*] (amb el diacrític d'augment de l'altura tonal) per expandir l'inventari disponible de contrastos fonològics entre accents tonals. És a dir, que passen a ser tres els tons fonològicament diferents disponibles en el sistema de transcripció entonativa del català: L, H, i ¡H.

6.2. Interacció entre variables prosòdiques i gestuals en el processament oracional

L'objectiu principal dels capítols 4 i 5 era entendre la interacció entre variables acústiques i visuals en la percepció lingüística de la interrogativitat. En el capítol 4, hem investigat la importància del camp tonal dels accents tonals i dels gestos facials en la percepció del contrast entre CFS i CEQ a través d'estímuls multimodals congruents i incongruents.

La qüestió principal que s'havia de respondre en el capítol 4 era en quina mesura les variables gestuals podien ser centrals en la codificació de la distinció lingüísticament rellevant entre CFS i CEQ. En els dos experiments d'identificació inclosos en el capítol, una sèrie de materials audiovisuals congruents i incongruents es va presentar a un grup d'oients nadius de català. L'anàlisi de les seues respostes demostra una clara preferència pels indicis visuals a l'hora de decidir entre les interpretacions CFS i CEQ, mentre que el contrast entonatiu basat en el camp tonal acaba exercint un paper secundari i de reforç. Els resultats també indiquen que, en algunes circumstàncies, els gestos facials poden actuar com a vehiculadors d'interpretació prosòdica i que competixen amb els indicis prosòdics, la qual cosa sembla contradir parcialment els resultats de nombrosos estudis en prosòdia audiovisual que trobaven un efecte merament complementari dels indicis visuals (Krahmer et al. 2002, Swerts & Krahmer 2004, Srinivasan & Massaro 2003, House 2002, Dohen & Lœvenbruck 2009, i d'altres).

Val la pena mencionar que mentre que a nivell segmental s'han observat clars efectes d'integració audiovisual, sobretot des de la publicació de l'estudi de McGurk & MacDonald (1976). Aquest estudi clàssic mostrava que, quan els oients adults d'anglès escoltaven [ba] a l'hora que veien els moviments labials corresponents a [ga], percebien [da] de resultes, una seqüència inicialment inexistent en els materials proporcionats als participants. Tot i això, quan els mateixos subjectes escoltaven o veien els mateixos materials unimodalment, percebien tant [ba] com [ga], respectivament. Els nostres resultats estan relacionats amb l'efecte McGurk en tant que ambdues modalitats, l'acústica i la visual, competixen i interactuen en les decisions dels nostres participants, però diferixen de l'efecte McGurk més 'clàssic' en el fet que no obtenim una categoria intermèdia entre les declaratives i interrogatives contrastades.

Un altre resultat interessant dels dos experiments del capítol 4 és que el paper de la informació acústica és més fort quan la informació visual és particularment ambigua, cosa que suggerix un patró d'integració audiovisual. Això significa que quan als participants se'ls mostraven exemplars no gaire clars de gestos CFS i CEQ, la seua dependència en la informació acústica augmentava. Un altre estudi complementari, que emprava materials sintètics i que comparava la percepció del contrast entre IFS i CEQ en català (Borràs-Comes et al. 2011), proporciona evidència addicional per al patró observat ací. En aquell estudi, la dependència en els indicis acústics augmentava generalment quan aquests es presentaven simultàniament amb una configuració facial IFS, i decreixia quan es presentaven amb una configuració facial CEQ. Com seria d'esperar, com que l'IFS és un tipus neutre de declarativa i la CEQ és un tipus marcat d'interrogativa, els participants depenien més fortament dels gestos facials de les CEQ que dels gestos pràcticament inexistents de les IFS.

Per un altre cantó, quan les propietats gestuals i entonatives són igual de prominents, els oients tendixen a basar les seues resposts tant en els senyals acústics com en els visuals d'una manera més equilibrada. El suport per a aquesta explicació ve de la distinció entre IFS i CFS en català central, també mitjançant l'ús d'avatars (Prieto et al. 2011). La diferència trobada entre IFS i CFS es basava tant en una activació gradual del camp tonal com en la força de l'activació de dos gestos concrets: l'avançament del cap i l'aixecament de les celles. Com que ambdues modalitats mostraven una distinció gradual i igualment prominent pel que fa al contrast lingüístic estudiat (IFS vs. CFS), es va trobar un ús equilibrat de les variables acústiques i visuals en la identificació de les dues categories (si ens centrem en els correlats gestuals, amb el moviment del cap representant un correlat més clar que l'elevació de les celles per a la identificació de CFS).

Aquesta interpretació compensatòria està lligada als resultats de Crespo-Sendra (2011) sobre la percepció audiovisual del contrast entre IFS i CEQ en català valencià i neerlandès. Mentre que les característiques facials d'ambdós significats són similars als que proporcionats en aquesta tesi, hi havia una clara diferència entre les dues llengües pel que fa al marcatge entonatiu: mentre que el català valencià marca la distinció entre els dos tipus d'interrogatives amb una diferència d'altura tonal aplicada a la mateixa configuració tonal ascendent (L* H%; transcrita L* HH% segons el sistema Cat_ToBI i L*H H% segons el sistema ToDI), el neerlandès empra dos contorns clarament diferenciats per distingir entre els dos significats (L* H% per a ISQ, i L+H* LH% per a CEQ). Quan els dos grups de parlants van ser presentats amb combinacions congruents i incongruents d'aquells materials audiovisuals, els parlants de català valencià depenien significativament més en les variables visuals, mentre que els parlants de neerlandès mostraven un efecte més equilibrat entre les dues variables.

6.3. El rol de les variables verbals i no verbals en la detecció de les interrogatives

En el capítol 5 s'ha explorat la importància relativa de diferents tipus de tons de frontera, de la mirada i de l'aixecament de les celles en la percepció del contrast entre IFS i ISQ en dos tipus de llengües: el neerlandès, que exhibix una estratègia sintàctica per al marcatge interrogatiu (la inversió subjecte/verb), i el català, que no compta amb aquesta estratègia. Els resultats del nostre experiment de percepció mostren que tant els participants neerlandesos com els catalans poden identificar interrogatives i declaratives per sobre del nivell d'atzar en totes les condicions de presentació. Més concretament, mostren una dependència més gran en la informació acústica, però també una millor precisió en les respostes d'identificació quan la informació visual s'afegix a l'acústica.

Aquest patró de resultats està parcialment en contradicció amb el proporcionat en el capítol 4. Quan els participants havien de distingir entre IFS i ISQ —tipus no marcats de declaratives i d'interrogatives, respecticament— mostraven una dependència més gran en la informació acústica que en la informació visual (encara que una presentació només visual dels materials també comportava una identificació significativament bona). En línia amb el que s'ha mencionat abans, suggerim que aquests resultats s'expliquen per les característiques específiques dels indicis acústics i visuals analitzats en ambdós capítols.

Pel que fa al contrast entre CFS i CEQ (capítol 4), la informació visual dels dos patrons facials era molt diferent l'una de l'altra, ambdues caracteritzades per moviments prominents del cap i de les celles (avançament o endarreriment del cap i aixecament o frunziment de les celles), i les característiques acústiques —encara que representen in contrast fonològic en la fonologia entonativa del català (v. capítols 2 i 3)— estan basades en una única diferència en el camp tonal del contorn entonatiu. En canvi, en el contrast entre IFS i ISQ (capítol 5), la informació visual que caracteritza aquesta diferència és perceptivament menys prominent i està determinada únicament per la presència o absència d'una sola propietat, la mirada, el rol de la qual millorava si se li afegia un moviment ascendent de celles; la informació acústica, però, estava basada en una de les dicotomies més interlingüístiques de les llengües entonatives pel que fa al marcatge interrogatiu, la distinció ascendent vs. descendent en el marc del to de frontera, i fins i tot en diferències sintàctiques quan estaven disponibles.

En aquest sentit, es pot defensar que la diferència trobada en el **pes** perceptiu de la informació auditiva i visual en aquests dos capítols està especialment lligada a la prominència perceptiva expressada per aquests indicis. Per exemple, la diferència entre dos tipus de contorns descendents (fins i tot si mostren una diferència pel que fa al camp tonal) serà menys prominent que l'existent entre un contorn descendent i un d'ascendent. Altrament, la diferència que hi ha entre tenir les celles aixecades o frunzides serà més prominent que la diferència entre unes celles aixecades i la seua configuració per defecte.

El que és especialment interessant del capítol 5 és que, en el si d'una única modalitat, també es troben efectes d'interacció com els que s'han trobat trobats entre els indicis acústics i visuals. Encara que en els nostres materials en neerlandès no hi havia cap ISQ produïda sense inversió subjecte/verb, els participants neerlandesos van classificar les oracions SV produïdes amb una entonació final ascendent igualment com a exemplars d'interrogatives. Aquesta preferència per les interpretacions interrogatives, en canvi, no es produïa si l'entonació ascendent havia estat aplicada a una estructura (sintàcticament marcada) VS, cosa que suggerix un pes **jeràrquic** de les variables que tenim disponibles en la nostra detecció de la interrogativitat. Finalment, obteníem el mateix resultat si comparàvem la mirada amb l'entonació, tant en neerlandès com en català; és a dir, que la presència de mirada augmentava significativament les interpretacions interrogatives dels enunciats, però només si anava acompanyada d'un contorn descendent.

Aquest resultat està en relació amb investigacions recents sobre el paper dels indicis verbals i no verbals com a elements mobilitzadors de resposta realitzats a través de l'anàlisi de corpus de parla espontània. Stivers & Rossano (2010) concloïen que "una petició (o un oferiment o una petició d'informació) és alta pel que fa a la rellevància que té proporcionar-hi una resposta, però que una petició dissenyada ‘directament’ (p. ex., amb morfosintaxi i/o prosòdia interrogatives) serà encara més alta [pel que fa a la rellevància de donar-hi una resposta]. Similarment, una valoració (o un avís o un anunci) serà baix pel que fa a la rellevància d’una resposta. Tanmateix, si se l’acompanya d’uns quants elements mobilitzadors de resposta, això incrementarà la rellevància d’una resposta per a tal acció” (Stivers & Rossano 2010: 27–28).

Aquest principi de rellevància de resposta pren en consideració el rol de la “morfosintaxi i/o prosòdia interrogatives” en el lloc més alt de la jerarquia, però també pren en consideració l’efecte **incremental** d’altres indicis disponibles. Stivers & Rossano (2010) trobaven tant per a l’anglès com per a l’italià que no hi havia cap propietat que estigués sempre present en tots els casos en què s’havia obtingut una resposta, de manera que concloïen que no hi ha cap element que siga intrínsec a l’acte de demanar informació. A més, concloïen que l’ús d’un major nombre d’elements mobilitzadors de resposta incrementava la rellevància d’una resposta per a una determinada acció. De fet, quan nosaltres analitzem els resultats de percepció AV del capítol 5, trobem una correlació positiva entre la concentració d’indicis interrogatius en una oració i l’avaluació d’aquesta oració com a interrogativa, per a totes dues llengües. Aquest patró de resultats suggerix —almenys si tenim en compte les dades del neerlandès i del català— la jerarquia següent dels diferents indicis per al marcatge interrogatiu SINTAXI > ENTONACIÓ > MIRADA (CELLES).

En resum, aquesta tesi ha proporcionat resultats que són rellevants per a la qüestió de la interacció entre indicis auditius i facials en la percepció d’una oració com a declarativa o interrogativa, fet que suggerix, en darrera instància, la relació amb conceptes com *pes jeràrquic*. Els resultats presentats ací permeten un millor coneixement de la comunicació humana i del paper que exercixen els gestos facials i les propietats entonatives dins d’aquest sistema, especialment el camp tonal.