Original Paper

Phonetica 2015;72:20–42
DOI: 10.1159/000381723

Received: July 1, 2014
Accepted after revision: March 14, 2015
Published online: June 5, 2015

Phonetic and Visual Cues to Questionhood in French Conversation

Francisco Torreira Emma Valtersson

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Abstract

We investigate the extent to which French polar questions and continuation statements, two types of utterances with similar morphosyntactic and intonational forms but different pragmatic functions, can be distinguished in conversational data based on phonetic and visual bodily information. We show that the two utterance types can be distinguished well over chance level by automatic classification models including several phonetic and visual cues. We also show that a considerable amount of relevant phonetic and visual information is present before the last portion of the utterances, potentially assisting early speech act recognition by addressees. These findings indicate that bottom-up phonetic and visual cues may play an important role during the production and recognition of speech acts alongside top-down contextual information.

1 Introduction

1.1 Theoretical Background

A classic problem in the study of spoken communication is the lack of a one-to-one mapping between the form of an utterance and its pragmatic function (Levinson, 1983, p. 289). For instance, an English utterance such as 'She's coming tonight' spoken with one or more high pitch accents and final falling intonation could function either as a statement or a question depending on top-down factors such as the epistemic and conversational context (Heritage, 2012; Levinson, 2013). In this article, we investigate two sources of variability in face-to-face communication that could provide bottom-up speech act cues: the phonetic realization of the utterance and the bodily behaviour of the speaker. In particular, we assess the amount of phonetic information (i.e., pitch scaling and alignment, final lengthening and final intensity profile, speech rate) and visual information (i.e., eyebrow action, gaze direction, manual gestures of the speaker) that distinguish polar questions from continuation statements

Francisco Torreira and Emma Valtersson contributed equally to this work.

KARGER 125

0031–8388/15/0721–0020
$39.50/0

E-Mail karger@karger.com
www.karger.com/pho

Dr. Francisco Torreira
Language and Cognition Department
Max Planck Institute for Psycholinguistics
Wundtlaan 1, NL–6525 XD Nijmegen (The Netherlands)
E-Mail Francisco.Torreira@mpi.nl

in conversational French. In this language, polar questions do not usually exhibit subject-verb inversion (or any other form of morphosyntactic marking) and are typically produced with final rising intonation. But French speakers also use final rising intonation in statements to signal that their turn is not yet complete. This raises the following question: to what extent can phonetic and visual cues alone discriminate between polar questions and continuation statements in conversational speech? One possibility is that speakers produce bottom-up cues very consistently, allowing for a clear classification between the two utterance types. The alternative is that, even if statistical differences in group means can be observed for some cues, a clear separation between the two groups based on bottom-up cues is nevertheless not possible (i.e., due to substantial overlap between the two categories). This would suggest that phonetic and visual cues would play at most a minor role during the production and recognition of these speech acts, and that their pragmatic encoding and decoding heavily rely on contextual top-down information alone.

A second issue that we explore in this article is the extent to which phonetic and visual cues to questionhood can be found before the final part of an utterance. Given the high frequency of smooth turn transitions in natural conversation (Sacks et al., 1974; Stivers et al., 2009) and the relatively long latencies involved in speech production (600 ms for picture naming, Jescheniak et al., 2003; Indefrey and Levelt, 2004; 1,500 ms for simple sentences, Griffin and Bock, 2000), Levinson (2013) notes that comprehension and production must often overlap during verbal interaction. He proposes that speech act cues appearing early during a turn are particularly useful for addressees, who use such cues to start planning their next-turn responses early on (e.g., an answer to a question). For this reason, the front-loading of speech act cues should be prevalent in conversational speech. Van Heuven and Haan (2002) synthesized a series of Dutch utterances with different slopes of the pitch baseline, different pitch accent ranges, and ending in either rising intonation (for questions) or falling intonation (for statements). A perceptual gating task showed that listeners differentiated questions from statements well before the final intonation movement based on the slope of the pitch baseline, and, to a lesser extent, the scaling of pitch accents. Although van Heuven and Haan (2002) showed that listeners can, in principle, make use of non-final cues to questionhood, the extent to which such cues are present in conversational verbal interactions remains to be seen. In the present study, we will assess the amount of non-final phonetic and visual information distinguishing polar questions from continuation statements taking our data from a corpus of spontaneous conversational French.

1.2 Phonetic Cues to Questionhood

It has often been claimed that questions across languages tend to be produced with some element of high pitch (e.g., Bolinger, 1989; Gussenhoven, 2002, but see Rialland, 2007, for interesting counterexamples in African languages). This high pitch element can be a phonologized tonal target, for instance a high boundary tone located at the end of the utterance, but can also be instantiated as a higher pitch scaling of the intonation contour, or the suppression of F0 declination throughout the utterance (Haan, 2002, for Dutch). Ohala (1984) attributes this bias to what he calls the Frequency Code. For a given amount of energy, smaller larynxes will tend to vibrate at faster rates, resulting in higher pitch. This correlation between larynx size and pitch height can be exploited for the expression of power relations and may indirectly cue differences between

utterances expressing uncertainty (such as questions) and utterances expressing certainty (such as statements). Experimental support for the Frequency Code can be found in Gussenhoven and Chen (2000). In this study, native speakers of Chinese, Dutch and Hungarian were presented with trisyllabic utterances from a fake South Pacific language unknown to them and were asked to judge whether the utterance was a question or a statement. Pitch patterns consisted of a rise-fall and a rise-fall-rise varying, among other things, in the scaling of the pitch peak and the final rise. In agreement with the Frequency Code, listeners selected the stimuli with higher peaks and higher end rises as signaling a question, regardless of their native language.

Although the higher pitch scaling of questions is often taken to be a widespread feature across languages, few studies to our knowledge have found empirical support for its use in conversational speech. Shriberg et al. (1998) set out to discriminate among seven speech acts, including polar and wh-questions, in a large corpus of English telephone conversations using a battery of prosodic features. They found that both polar and wh-questions displayed higher average F0 values than the other speech acts in their coding, and that average F0 helped improve the automatic classification of questions and statements. More recently, Edlund et al. (2012) compared questions to statements in a corpus of spontaneous Swedish, and also observed a higher average pitch value for questions than for statements. Questions have also been found to exhibit higher pitch scaling than statements in numerous read-speech studies on different languages (see Haan, 2002 for Dutch, and references therein for Swedish, English, Cherokee, Hausa, Spanish, Portuguese, Finnish, Arabic, Chinese, Chichewa, Chickasaw, Hungarian, Vietnamese, and Thai).

Studies on several languages have also pointed to tempo differences between questions and statements. Van Heuven and van Zanten (2005) investigated differences in speech rate between read polar questions and statements in Manado Malay, Orkney English and Dutch, and found a faster speaking rate for questions in all three languages, with different distributions of the phenomenon over the utterance. In Manado Malay, the difference was restricted to the right boundary of the prosodic domains, in Orkney English it was more evenly spread over the utterance, and in Dutch it was only found in the middle portion of the utterance. Similarly, Cangemi and D'Imperio (2011) found that questions in a corpus of read Italian sentences tended to exhibit shorter segment durations than statements in the first consonant and last vowel of the utterance. Edlund et al. (2012) also investigated speech rate differences between questions and statements in a corpus of spontaneous Swedish and, in agreement with the read-speech studies above, found that questions exhibited faster rates than statements. These studies indicate that questions tend to be spoken at faster speech rates than statements across different languages, both in read and spontaneous speech, and that there is language-specific variability regarding where the tempo difference is located in the utterance. Van Heuven and van Zanten (2005) speculated on the origins of such tempo differences and pointed to Ohala's (1984) ethological explanations. Faster utterances fit a smaller creature better than a large one, and, indirectly, could cue uncertainty rather than assertion. An increased level of tension during speech production for questions in comparison with statements could also provide the link between faster speech rate and questionhood. Finally, the fact that questions can be regarded as incomplete events, part of larger conversational units, together with the tendency for longer utterances to be spoken faster, may explain why questions tend to exhibit faster speech rates than statements.

1.3 Visual Cues to Questionhood

A number of bodily visual cues have been associated with questionhood in the literature on multimodal communication. Several studies have observed that raised eyebrows tend to co-occur with questions (e.g., Ekman, 1979; Borràs-Comes et al., 2014). However, others have found that questions and non-questions alike exhibit comparable amounts of raised eyebrows (e.g., Purson et al., 1999; Flecha-García, 2010). Another visual cue that has been investigated in relation to questionhood is the direction of the speaker's gaze. In a controlled production task, Borràs-Comes et al. (2014) observed that Dutch and Catalan speakers directed their gaze at their interlocutors more frequently during questions than statements. In a cross-cultural study, Rossano et al. (2009) showed that, during questions, speakers gaze more towards their addressees than vice versa. Along these lines, Stivers and Rossano (2010) later proposed that the direction of speaker gaze is one of several resources for mobilizing response in face-to-face social interaction.

A third type of bodily behaviour potentially related to the cueing of questions is the type of manual gestures performed by the speaker. Bavelas et al. (1992) observed that not all manual gestures in face-to-face interaction illustrate the semantics of the linguistic message being conveyed, and distinguished between interactive and topic gestures. Interactive gestures, such as palm-up hand or index finger movements directed towards the addressee, manage the interaction between speaker and addressee, whereas topic gestures (e.g., iconic and metaphoric gestures) are connected to the semantics of the linguistic message. Given that questions usually seek information from the addressee, we hypothesize that interactive gestures directed at the addressee co-occur with questions more often than with other types of utterances.

1.4 Rising Intonation in French

As explained above, final rising intonation in French is found in both polar questions and continuation statements. Several claims regarding the prosodic distinction of these two utterance types have been made in the French intonation literature. Based on introspection, Delattre (1966) claimed that they both have final rising pitch, but also that they are differentiated by the scaling and shape of their final pitch rise, with questions reaching a higher pitch maximum and exhibiting a more concave final pitch contour than continuation statements. Two later studies using spontaneous and read speech materials (Grundstrom, 1973; Rossi et al., 1981) agreed with Delattre's (1966) descriptions regarding the scaling of the final pitch maximum, but did not find evidence for a distinction in terms of contour shape. In their recent autosegmental description of French intonational phonology, Delais-Roussarie et al. (in press) describe a high-rising H* H% nuclear contour that can be used both for continuation statements and polar questions. They also mention a low-rising L* H% nuclear contour, which, in their read speech data, occurs in polar questions with morphosyntactic marking. Regarding non-pitch cues, Rossi et al. (1981) found a more pronounced final intensity drop and longer final vowels for questions than for continuation statements in a phonetic perceptual experiment. Longer final lengthening in French questions was also found in Smith (2002) in a corpus of read sentences, both for questions exhibiting final rising and final falling intonation. The findings of Rossi et al. (1981) and Smith (2002) therefore contrast with those reporting shorter, not longer, durations for questions (van Heuven and van Zanten, 2005; Cangemi and D'Imperio, 2011; Edlund et al., 2012).

In a preliminary investigation focusing on the final intonation rise of polar questions and continuation statements in conversational French (Valtersson and Torreira, 2014), we observed, in agreement with previous studies, that the final intonation rise for questions exhibited higher scaling on average than the rise for continuation statements. We also found, in agreement with Rossi et al. (1981), a larger final intensity drop for questions. But, contrary to Rossi et al. (1981) and Smith (2002), we also observed that the final vowels of questions tend to be shorter, not longer, than those of continuation statements. We also investigated possible differences in the shape of the rise by fitting the pitch contours in the utterance-final vowels with quadratic polynomials (cf. Andruski and Costello, 2004; Torreira, 2007). In disagreement with Delattre's (1966) proposal, we found no systematic differences between the two kinds of utterances. Because of this, together with a substantial amount of overlap in all the phonetic measures for which we found differences, we proposed, along Di Cristo (1998) and autosegmental-metrical analyses (Post, 2002; Delais-Roussarie et al., in press), that both polar questions and continuations in French make use of one rising intonation pattern (H* H% in autosegmental terms), and that its phonetic realization is conditioned by interactive and communicative factors. For instance, the longer final syllables in continuations could assist the speaker as a buffer in the planning of the upcoming utterance, and the higher scaling of questions could be attributed to the use of Ohala's (1984) Frequency Code.

The fact that we found statistical differences between the phonetic realizations of polar questions and continuation statements does not mean that ambiguous and even misleading cases were rare in our data. Regarding pitch, for instance, we observed numerous questions with a shallow final rise and continuation statements with markedly higher final pitch values. This is illustrated in figures 1 and 2, which show waveforms, spectrograms, and pitch tracks of two utterances extracted from the Nijmegen Corpus of Casual French (NCCFr, Torreira et al., 2010). The polar question in figure 1 exhibits a final intonation rise reaching a maximum of 420 Hz, roughly 10 semitones above the speaker's median pitch (previously calculated on an excerpt of 10 min of conversation), whereas the continuation statement in figure 2 ends considerably higher, roughly 15 semitones relative to the speaker's median pitch. It can also be seen in these figures that the initial portions of the contours do not differ much in terms of their initial scaling or declination. Such departures from statistical trends are common enough in our data to make us wonder whether a good discrimination between questions and statements can be obtained from bottom-up cues alone.

In the present study, we extend our previous research by examining the phonetic realization of polar questions and continuation statements before and during their final intonation rises, and by also investigating the bodily behaviour of the speaker. We first use regression modeling in order to identify differences in group means and probability of occurrence for several potentially relevant cues. After identifying statistically relevant cues, we assess how well they serve to discriminate between polar questions and continuation statements using logistic regression in a leave-one-out cross-validation procedure. In order to address Levinson's (2013) front-loading hypothesis, according to which speech act cues should be present early in the utterance, we assess the discriminative power of non-final cues relative to a full model including all cues, final and non-final.

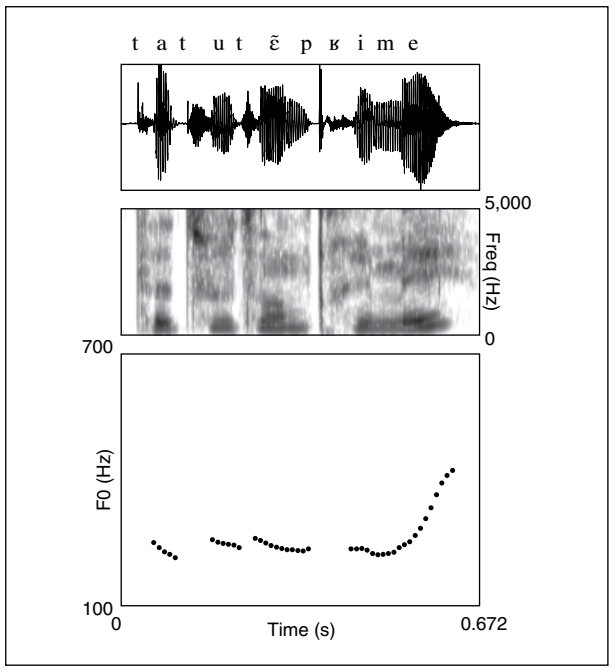

Fig. 1. Waveform, spectrogram, and pitch track for the polar question *T'as tout imprimé* 'Have you printed everything?' spoken by female speaker F05L in the NCCFr.

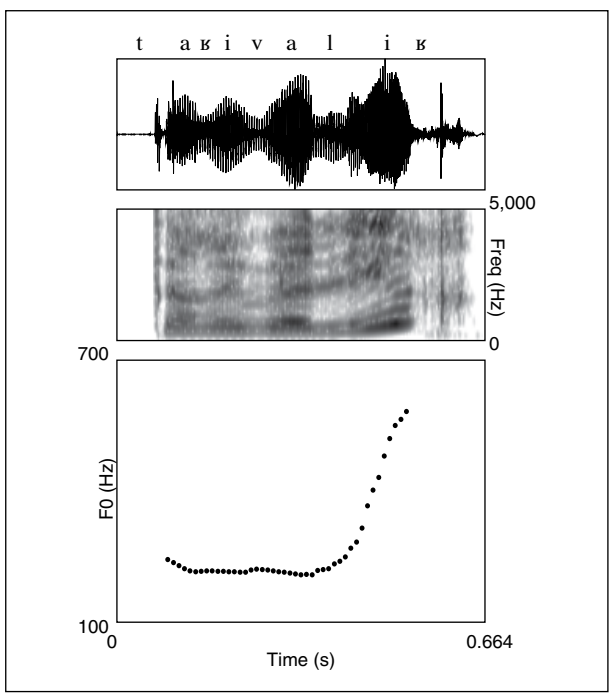

Fig. 2. Waveform, spectrogram, and pitch track for the continuation statement *T'arrives à lire* 'You can read (it)' spoken by female speaker F06R in the NCCFr. The utterance was produced as part of the larger turn *Parce que tu tu lis le truc, t'arrives à lire, mais euh t'as pas la compréhension* 'Because you you read the thing, you can read (it), but uhm you don't get the meaning'.

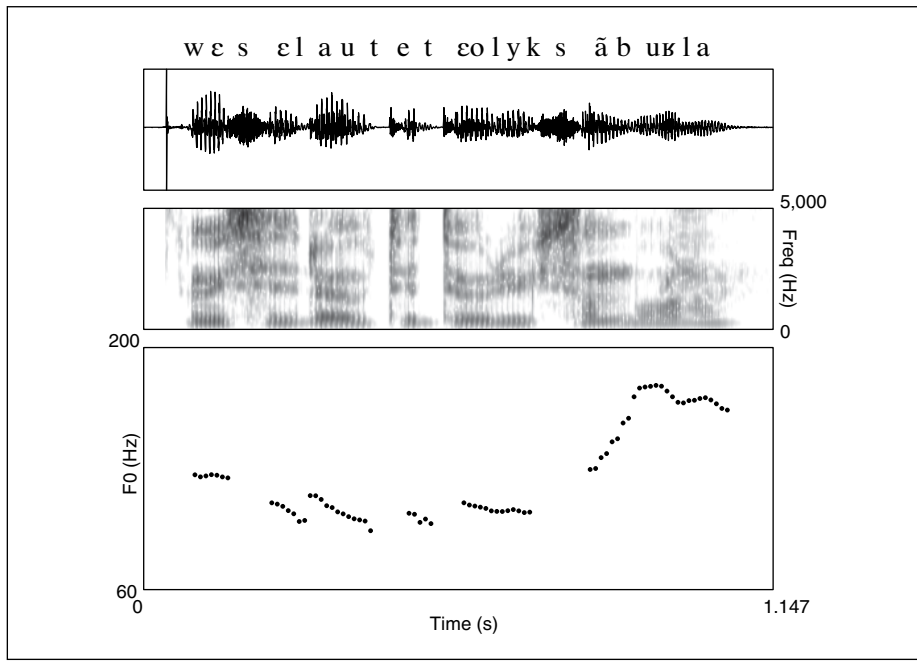

Fig. 3. Waveform, spectrogram, and pitch track of an utterance ending in a final high pitch plateau: *Ouais, c'est là où t'étais au Luxembourg là?* 'Yeah, is that where you were in Luxembourg (over) there?', spoken by male speaker M22R in the NCCFr.

2 Methods

2.1 Materials and Data Extraction

Our data come from the NCCFr, which consists of 23 casual conversations among dyads and triads of friends recorded with head-mounted microphones and filmed with a video camera in a sound-attenuated room. Only the initial dyadic part of each conversation, which had a duration ranging from 4 to 40 min, was used, providing a total of approximately 7.5 h of spontaneous conversation. Twelve dyads were composed of female speakers, and 11 of male speakers. All speakers originated from Central or Northern France and were mostly university students between the ages of 18 and 27. A detailed description of the NCCFr can be found in Torreira et al. (2010).

The two authors independently inspected all conversations in the data and identified cases of pre-pausal rising intonation using Elan and Praat software (Sloetjes and Wittenburg, 2008; Boersma and Weenink, 2014). The reason for only considering utterances in pre-pausal position was to control for possible contextual effects on our phonetic dependent variables (e.g., tonal and segmental anticipatory coarticulation). Figures 1 and 2 illustrate two utterances with pre-pausal rising intonation included in our dataset. It should be noted that we only selected cases of final rising intonation and discarded utterances with a final high or mid pitch plateau. Utterances ending in a high pitch plateau typically occurred in cases where the speaker added an increment of one or more words, such as a tag or the word *là* 'there', to an otherwise complete utterance (Jun and Fougeron, 2000). Utterances with a mid pitch plateau, which were also characterized by a significant amount of final lengthening, were found mostly in enumerations. Figures 3 and 4 illustrate these two cases.

After the initial inspection of the data, each author's annotations were compared so that cases of disagreement could be identified. Most disagreements were valid cases of pre-pausal rising intonation that had escaped the attention of one of the authors and were included in the dataset after a brief

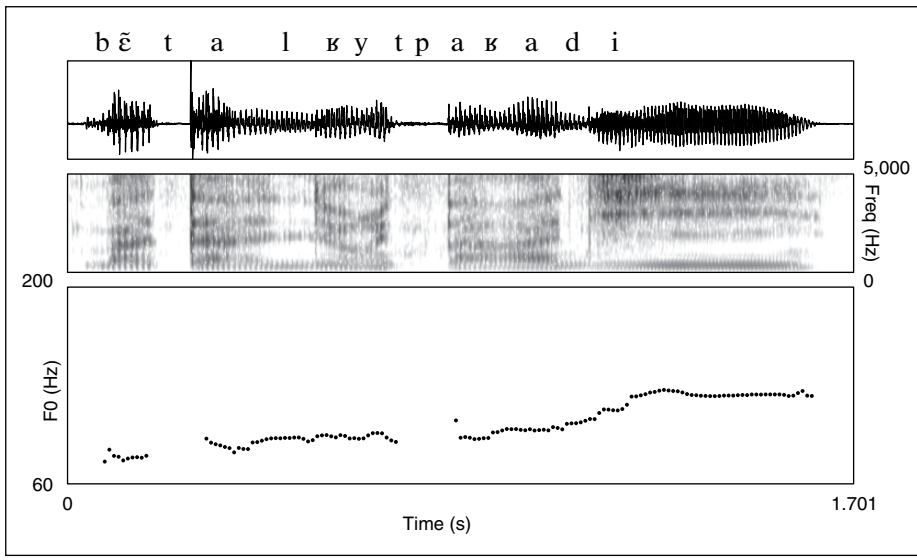

Fig. 4. Waveform, spectrogram, and pitch track of an utterance ending in a mid pitch plateau: *Ben t'as le Rue de Paradis…* 'So there's the (one on) Rue de Paradis…' spoken by male speaker M19R in the NCCFr.

discussion. In a small number of cases for which a disagreement could not be resolved, we excluded the token from the data. These typically involved a final rise according to one of the authors, and a final rise-fall (the *intonation d'implication* pattern described in Delattre, 1966) according to the other author. Finally, utterances with pre-pausal rising intonation containing hesitations, speech errors or laughter were also excluded from the dataset. The final dataset contained 278 tokens of fluent utterances judged to have pre-pausal rising intonation by both authors.

In order to check the replicability of the initial data selection, the second author repeated the same procedure for four random conversations 18 months after the original identification procedure was completed. Out the 60 continuation statements and 29 polar questions in our initial coding for these four conversations, the second author correctly identified 58 continuation statements (96.7% of the total number of continuations) and 28 polar questions (96.5%).

2.2 Coding and Measurements

2.2.1 Pragmatic Coding

For each utterance, each of the two authors annotated whether it was a polar question or a continuation statement in the context of the conversation. Polar questions typically involved knowledge within the listener's epistemic domain (e.g., 'You went to Paris yesterday?'; Geluykens, 1987; Heritage, 2012) and were followed by a turn transition and an answer. Continuation statements typically involved knowledge within the speaker's epistemic domain (e.g., 'I went to Paris yesterday') and were usually not followed by a turn transition. Although many cases seemed ambiguous out of their conversational context, such ambiguities were easily resolved when the case was interpreted in its context and discussed by the two authors. The dataset contained a total of 105 questions, and 173 continuation statements.

In order to check the reliability of our coding of questions and continuation statements, the second author repeated the coding procedure for four random conversations 18 months after the original coding was completed. Out the 60 continuation statements and 29 polar originally coded in these four conversations, there was only one case of disagreement between the second and original codings.

During the pragmatic annotation of the data, we noticed that questions could accomplish a number of different conversational functions. We observed questions that seemed to seek new information

in a straightforward way, such as the question presented below in example 1 (see fig. 1 for a pitch track; square brackets indicate overlapping speech; the relevant polar question is marked with an arrow):

1. NCCFr 20-11-07_1_06:28
A: *euh putain* ((looking for something in her bag))
uh shit
B:→*t'as tout imprimé?*
have you printed everything?
A: *non* [*j'ai gardé*] *j'ai fait j'ai gardé sur mon truc*
no I have kept I have made I have kept it on my thing
B: [*ouais*]
yeah

We also observed a substantial number of questions that referred to or asked for clarification about an element in the interlocutor's previous turn. These questions, which we will call backward-looking questions, often consisted of a non-sentential phrase without an inflected verb, as illustrated in example 2:

2. NCCFr 20-11-07_1_07:26
A: *je sais pas si on l'aura non plus du coup*
I don't know if we'll have it either then
B:→*vendredi?*
on Friday?
B: *bah pas vendredi là je pense qu' on l' aura pas*
not this Friday I think that we won't have it
[*mais euh*::::::::::]
but uh
A: [*ah ouais plus tard*] *ouais*
oh yeah later yeah

Some of the backward-looking questions were not primarily intended as requests for information, but rather as displays of surprise. This was evidenced explicitly by their semantic content (e.g., *C'est vrai?* 'Is that true?'), or by the context in which the question occurred. For instance, as shown in example 3 below. The question could be immediately preceded by a token of understanding (e.g., *ah* 'oh') by the same speaker, and/or be followed by a detailed explanation instead of a positive or negative answer by the interlocutor:

3. NCCFr 04-12-07_1_02:47
A: *ben moi le le cours de Yves j'ai vraiment pas trop aimé*
So me Yves' class I haven't really liked
A: *le premier*
the first lesson
B: *non mais tu lis enfin il y a des bouquins*
No but you (can) read come on there are books
A: *et demain matin il y a cours?*
And tomorrow morning do we have class?
B: *ouais à dix heures*
yeah at ten
A:→*ah a dix heures?*
oh at ten?
B: *c'est la journée paysage demain il y a cours à de dix*
it's landscape day tomorrow we have class at from ten
heures à midi avec Lujinbul et avec Monsieur Toublan
to noon with Lujinbul and with Mr. Toublan
A: *ah mais je croyais que c'était le matin Toublan*
oh but I thought Toublan was in the morning
et l' après-midi Lujinbul
and Lujinbul in the afternoon

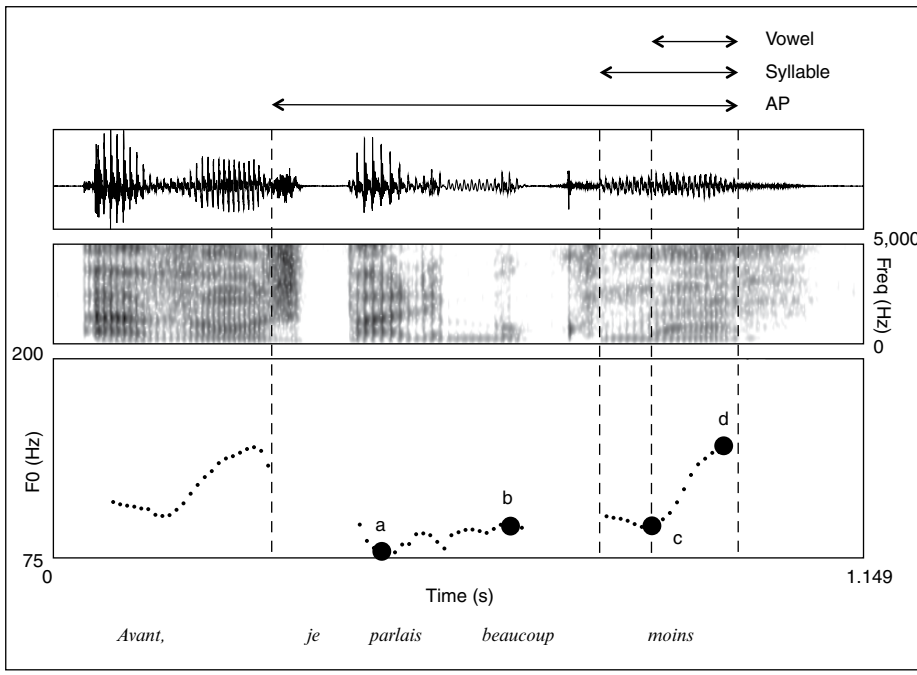

Fig. 5. Illustration of our phonetic measurement scheme. The example shows the continuation statement *Avant, je parlais beaucoup moins, […]* 'In the past, I talked much less, […]' spoken in two APs, of which only the second one was included in the dataset. Points a and b correspond to the pitch values approximately at the centre of the first and penult vowels in the AP containing the pre-pausal intonation rise. Points c and d correspond to the minimum and maximum pitch points in the last vowel of the AP. We also measured the difference in intensity between the intensity peaks in the last two vowels of the AP and estimated the speech rate of the AP up to its final syllable. See text for details.

To estimate whether the main question types in our data correlated with the use of specific phonetic and bodily visual cues, the first author coded each question for (a) any reference to an element in the interlocutor's previous turn, and (b) the presence of non-phonetic and non-gestural elements in the conversational context indicating that the question was intended as a display of surprise. Eighty-two questions referred to the previous turn in the same conversational sequence, whereas 23 opened up new topics of discussion. Non-phonetic and non-gestural evidence of a display of surprise was annotated in 25 out of the 105 questions. In order to check the reliability of this coding, the second author repeated the procedure for four random conversations. In the 18 questions present in these four conversations, there was perfect agreement for question directionality (backward-looking vs. forward-looking) and a good agreement for the coding of surprise display (88.9% of agreement).

2.2.2 Phonetic Measures

Figure 5 illustrates part of our phonetic measurement scheme. For each token in the dataset, we first identified the accentual phrase (AP; Jun and Fougeron, 2000) that contained the final intonation rise. APs in our data could be preceded by either a silent pause, or another AP, and always ended in a final accent. We segmented the beginning and end of the final syllable and vowel in this AP using standard phonetic segmentation procedures. Then, several acoustic pitch measures were taken automatically via a Praat script. All pitch measurements were done in semitones re 100 Hz using the autocorrelation method in Praat set to its default parameters. In order to control for speaker-based variation, we

subtracted from each pitch measurement its speaker's median value, which we had previously calculated from an excerpt of 10 min of conversation. We measured the minimum and maximum pitch values in the final vowel of the AP (illustrated by points c and d in fig. 5), which was always present regardless of the structure of the final syllable (e.g., open or closed, with or without an onset). To investigate differences in F0 scaling before the final rise, we also took pitch measures approximately at the centre of the vowels in the first and penultimate syllables of the AP (points a and b, respectively, in fig. 5) for all APs with three or more syllables (n = 195). In a small number of cases, the penultimate vowel happened to be devoiced in the context of voiceless consonants (cf. Torreira and Ernestus, 2010). In such cases, the pitch measure was taken in the preceding syllable. This is shown in figure 5, in which the vowel /u/ in the word *beaucoup* /boku/ 'much' is highly reduced and devoiced after consonant /k/. To investigate differences in the alignment of the final pitch rise, we also measured the distance from the beginning of the rise, which in figure 5 coincides roughly with point c, and the beginning of the last syllable in 65 tokens with sonorant onset consonants. This measure was not taken in the rest of the data, since the presence of voiceless consonants, significant microprosodic perturbations in non-sonorant voiced consonants, or absence of an onset consonant did not allow for reliable pitch alignment measurements.

APs in French always exhibit one tonal prominence in their final syllable (i.e., a final rise in our case) and may also contain an initial tonal prominence in the form of a high accent located earlier within the AP (Di Cristo, 1998; Jun and Fougeron, 2000; Post, 2002). Because the occurrence of initial high accents could affect our pitch measures, we also annotated their presence in our data. Initial accents were present in 38 of the 278 tokens. The length of the APs in our data ranged from 1 to 15 syllables. Most APs (57%) had from one to four syllables, and only a small proportion (5.4%) had ten or more. Continuation statements tended to be somewhat longer than questions, with an average length of 5.02 syllables (SD = 3.05), versus 3.53 syllables for questions (SD = 2.4).

In order to investigate differences in the final intensity profile of questions and statements, we calculated the final drop in intensity between the intensity peaks of the last two syllables. In a small number of cases in which the penultimate syllable contained a highly reduced and devoiced vowel, the intensity drop was measured between the intensity peak in a preceding, non-devoiced vowel and the final vowel. Finally, we also estimated the speech rate of the AP (excluding its final syllable) by dividing the duration of this interval by the number of syllables contained in it. We excluded the last syllable of the AP from the speech rate measure to exclude variability due to final lengthening. Because of this, 44 APs containing only one syllable could not be assigned a speech rate value.

2.2.3 Bodily Behaviour Measures

To investigate differences between continuation statements and polar questions in the speaker's bodily behaviour, the second author coded eyebrow and head movements, the gaze direction of the speaker, and hand gestures during the AP containing the final intonation rise. One of the 23 video recordings could not be analysed due to a video-audio synchronization problem.

All visible hand movements performed by the speaker excepting self-regulators (e.g., scratching) or other non-communicative bodily movements (e.g., playing with a water glass) were considered as hand gestures. Each hand gesture was labelled as a topic, a beat, or an interactive gesture directed at the addressee following the definitions in Bavelas et al. (1992). Topic gestures are related to the semantics of the linguistic message being conveyed by the speaker and correspond to the iconic, metaphoric and deictic gestures in McNeill (1992). Interactive gestures, on the other hand, included those gestures which were related to the addressee rather than to the semantics of the linguistic message. These included beats and different types of hand gestures oriented towards the addressee such as index-finger pointing, whole-hand pointing, and palm-upward movements. Figure 6 illustrates a case in which the questioner produces a palm-upward gesture oriented toward the recipient. In those cases in which the upper part of the face was visible (e.g., not covered by bangs), we annotated upward ostensible eyebrow movements and frowns. Figure 7 shows a case of visible upward eyebrow movement. We also annotated whether the speaker's gaze was directed towards the addressee at any point within the AP. Finally, we annotated whether the speaker produced a nod or an upward movement with her head.

In order to address the front-loading hypothesis proposed by Levinson (2013), we also annotated whether each of the bodily behaviours above was visible before the last two syllables of the AP, where our final phonetic measures were taken.

To check the reliability of our codings, the first author annotated the visual bodily behaviour variables presented above in four random conversations. The agreement was very good for eyebrow

Color version available online

Fig. 6. Example of a palm-up hand gesture (indicated with an arrow) oriented to the recipient, as produced by speaker F03L.

Color version available online

Fig. 7. Example of raised eyebrows (indicated with an arrow), as produced by speaker M19L.

movements (96.9%), gaze direction (96.9%), hand gestures (92.2%), and head movements (88.9%). The timing annotations agreed in the vast majority of cases for head movement (88.9%), gaze direction (87.5%), and hand gestures (84.4%) and exhibited perfect agreement for eyebrow movements.

3 Results

3.1 Phonetic Cues

3.1.1 Pitch Measures

Figures 8 and 9 show the average values of our four pitch height measures for male and female speakers in utterances consisting of more than two syllables (male and female speakers exhibited somewhat different pitch patterns and are therefore shown separately).

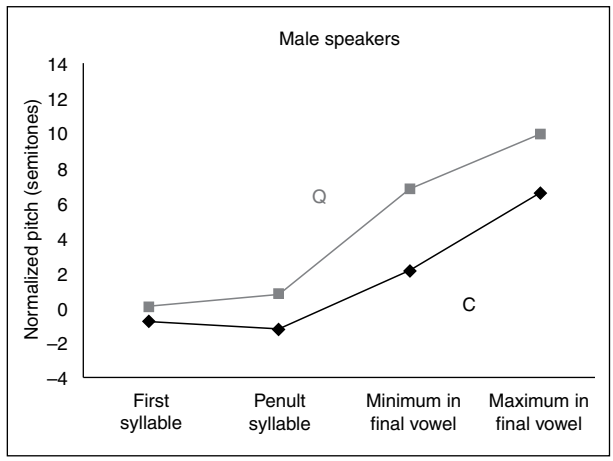

Fig. 8. Speaker-normalized mean pitch values (in semitones) for four pitch measurements in continuation statements (C) and polar questions (Q) of more than two syllables spoken by male speakers (n = 116).

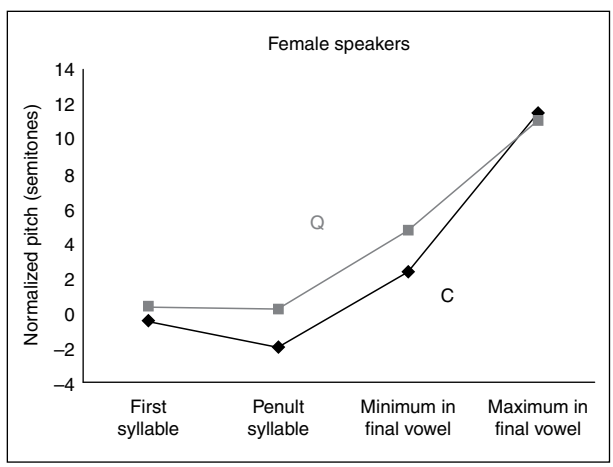

Fig. 9. Speaker-normalized mean pitch values (in semitones) for four pitch measurements in continuation statements (C) and polar questions (Q) of more than two syllables spoken by females (n = 79).

These figures show that, excepting the final pitch maximum at the end of the utterance in the case of female speakers, questions generally exhibited a higher average pitch scaling than continuation statements. Moreover, the difference in final minimum pitch appears to be greater for males than for females. A series of regression models with each of our pitch measures as dependent variable, utterance type and gender as fixed predictors, and speaker as a random factor confirmed the statistical validity of these observations. We obtained statistically significant interactions between utterance type and speaker gender both for final minimum pitch (utterance type: $\beta = 2.31$, $t = 3.31$, $p < 0.0005$; gender × utterance type: $\beta = 2.33$, $t = 2.45$, $p < 0.05$) and final maximum pitch (gender: $\beta = –4.31$, $t = –6.24$, $p < 0.0001$; gender × utterance type: $\beta = 2.47$, $t = 3.04$, $p < 0.005$). Main effects of utterance type were also observed for initial pitch ($\beta = 0.83$, $t = 2.87$, $p < 0.005$) and for the pitch value measured in the penultimate syllable of the utterance ($\beta = 1.87$, $t = 5.66$, $p < 0.0001$). In contrast with our findings for final pitch values, no effect of gender or interaction between gender and utterance type was observed for the non-final pitch measures.

Although figures 8 and 9 show group means for each pitch measure, not pitch differentials within data points, these figures also suggest that continuation statements were more prone to declination before the final rise than polar questions. Looking at utterances with more than two syllables, we could confirm this with a regression model with the pitch differential between the two first pitch points as the response, utterance type and gender as predictors, and speaker as a random factor. The model estimated that the pitch of continuation statements decreases slightly between the first and penultimate syllable with an intercept of –0.79 semitones ($t = –3.52$, $p < 0.0005$), but this downward trend is cancelled for polar questions ($\beta = 1.02$, $t = 3.19$, $p < 0.0005$).

Finally, we checked if the presence of initial pitch accents affected the realization of our pitch measures by adding this information to the models above. Accents were present in both utterance types, although they were somewhat more frequent in continuation statements (n = 29, 16.6%) than in polar questions (n = 9, 8.6%). These models revealed that initial accents raised the scaling of the pitch value in the penultimate syllable of the AP, but not the other pitch values. In particular, an interaction between accent and utterance type indicated that this effect was less marked for continuation statements (accent: $\beta = 0.86$, $t = 2.01$, $p < 0.05$) than for questions (accent × utterance type: $\beta = 1.69$, $t = 2.07$, $p < 0.05$). These findings indicate that the presence of initial high accents in French APs may lead to an undershoot of the dip preceding final rising pitch movements.

We then inspected the alignment of the final pitch rise with respect to the syllable onset. Visual inspection of the distributions of alignment values revealed that the most frequent alignment was close to the syllable onset for both utterance types, with some variability around this point for both groups (i.e., both early and late alignments were observed for both questions and statements). A regression model with the distance from the beginning of the pitch rise to the syllable onset as the response, utterance type as fixed predictor, and speaker as a random factor did not yield a statistical difference ($\beta = –5.9$, $t = –0.63$, $p = 0.53$). We therefore conclude that the alignment of the beginning of the final pitch rise relative to the onset of the last syllable does not differ systematically between questions and continuation statements.

3.1.2 Durational and Intensity Measures

We then inspected the durational and intensity properties of polar questions and continuation statements. A regression model with speech rate as the response variable and utterance type as the main predictor indicated that polar questions tended to be spoken slightly faster than continuations ($\beta = 1.22$, $t = 4.67$, $p < 0.0001$). The duration of the final vowel tended to be slightly shorter in polar questions than in continuation statements ($\beta = –0.013$, $t = –2.59$, $p < 0.01$). In order to check if this effect was a difference in final lengthening, or if it was due to a more general difference in speech rate affecting the whole AP (including its final syllable), we compared a full logistic regression model predicting utterance type with speech rate and final vowel duration as predictors with two reduced models including speech rate and final vowel duration separately. Interestingly, the full model did not significantly improve the fit of the reduced model with speech rate as only predictor [$\chi(1) = 2.66$, $p = 0.1$]. However, when the full model was compared to the reduced model with duration as only predictor, a statistically improvement in fit was observed [$\chi(1) = 16.68$, $p < 0.001$]. This suggests that the difference in duration found in the last vowel of the AP can be largely

attributed to a more general difference in speech rate affecting the whole AP, rather than to a consistent difference in final lengthening.

We finally inspected whether the final intensity profile of questions and statements in our data differed. In agreement with Rossi et al. (1981), a regression model with intensity drop (in dB) as the dependent variable and utterance type as the main predictor indicated that questions tend to exhibit a more pronounced intensity drop in the last two vowels than continuation statements ($\beta = -2.21$, $t = -3.53$, $p < 0.001$).

3.1.3 Phonetic Cues and Question Type

We then investigated whether the main types of questions in our pragmatic coding (i.e., backward-looking vs. forward-looking, displaying surprise vs. not displaying surprise) differed phonetically from each other, and if so, how the different kinds of questions compared to continuation statements.

To do this, we fitted linear regression models with each of the phonetic cues identified in the statistical tests of the previous subsection as the response variable, the two question type features as main predictors, gender as a covariate in the case of pitch cues, and speaker as a random factor. No statistical differences were found between backward-looking and forward-looking ($-1.5 > t < 1.5$ in all models). Interestingly, however, we did find highly significant differences between questions that displayed surprise and other questions for all of the pitch measures. Questions displaying surprise exhibited higher pitch scaling on average than other questions in all of the four pitch measures (initial pitch: $\beta = 1.62$, $t = 3.48$, $p < 0.0005$; pitch in the penult: $\beta = 1.65$, $t = 3.01$, $p < 0.005$; final minimum pitch: $\beta = 2.03$, $t = 0.01$, $p < 0.05$; final maximum pitch: $\beta = 3.62$, $t = 4.36$, $p < 0.0001$). No interactions at the α level of $p < 0.05$ were observed between gender and the main predictors, although a statistical trend ($t = 1.84$, $p = 0.06$) was found for final maximum pitch, with males producing a clearer distinction between the two types of questions than females.

Since questions displaying surprise appeared to be produced with higher pitch than other questions, we wondered whether the pitch differences between questions and continuation statements identified in section 3.1.1 was entirely driven by them. To investigate this, we ran the regression models on pitch values presented above in subsection 3.1.1 after excluding all questions displaying surprise. Questions not displaying surprise were still higher in pitch than continuation statements in the two first points (initial pitch: $\beta = 0.63$, $t = 2.38$, $p < 0.05$; pitch in the penult: $\beta = 1.51$, $t = 5.5$, $p < 0.0001$). For the pitch values measured in the final syllable, we only found statistically significant differences for male speakers (final minimum pitch: $\beta = 3.97$, $t = 7.12$, $p < 0.0001$; final maximum pitch: $\beta = 1.94$, $t = 3.44$, $p < 0.001$). Female speakers exhibited a statistical trend for final minimum pitch ($\beta = 1.79$, $t = 1.94$, $p = 0.054$) and no statistical difference for final maximum pitch ($p = 0.34$). Figures 10 and 11 illustrate the average pitch scaling of questions displaying surprise, questions not displaying surprise, and continuation statements for male and female speakers. No durational or intensity differences were observed between questions displaying surprise and other questions ($p > 0.2$ in both cases).

3.2 Visual Cues

Table 1 shows the number of utterances in our data containing each of the visual cues in our coding, and the percentage of polar questions for each cue. With the exception of gaze, the majority of utterances in our data were not accompanied by bodily

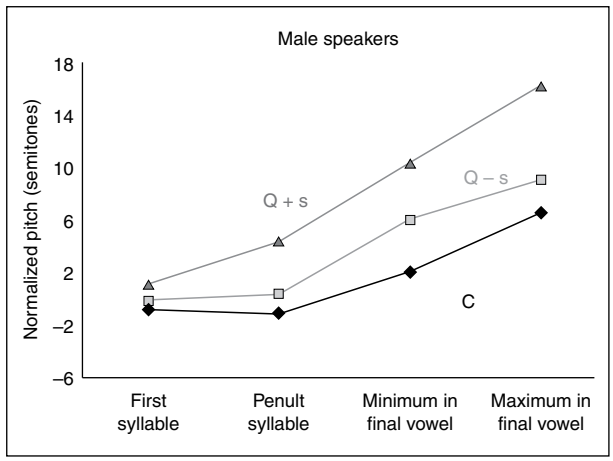

Fig. 10. Speaker-normalized mean pitch values (in semitones) for four pitch measurements in continuation statements (C), polar questions displaying surprise (Q + s), and polar questions not displaying surprise (Q – s) for utterances of more than two syllables spoken by male speakers.

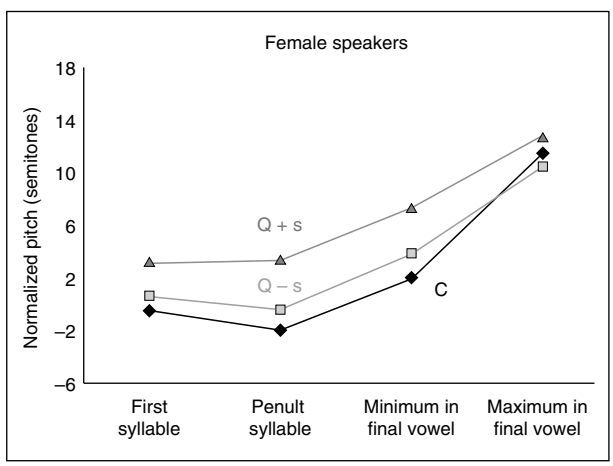

Fig. 11. Speaker-normalized mean pitch values (in semitones) for four pitch measurements in continuation statements (C), polar questions displaying surprise (Q + s), and polar questions not displaying surprise (Q – s) for utterances of more than two syllables spoken by female speakers.

behaviour. Frowns only occurred in 6 cases, nods in 8 cases, upward head movements in 12 cases, and raised eyebrows were observed on 13 occasions. The percentage of polar questions in cases with raised eyebrows was over twice as high as when no eyebrow action was present (69.2 vs. 30.4%). This difference was statistically significant despite the poor statistical power of the test due to the low number of data points ($\beta = 2.2$, $z = 2.88$, $p < 0.005$). The distributions in the case of head movements were very clear. All nods and all upward head movements occurred in questions.

Gaze towards the addressee was present in our data in most of the utterances (in 203 out of 260 cases coded for this variable). As expected from previous studies, we found that utterances which were accompanied by gaze towards the addressee tended to be polar questions more frequently than utterances which were not (42.4 vs. 22.8%). This difference was statistically significant in a regression model with utterance type as the response, gaze as fixed predictor, and speaker as random factor ($\beta = 1.46$, $z = 3.49$, $p < 0.0005$). Regarding manual gestures, we found that topic gestures, which were not infrequent in

Table 1. Number of tokens in our dataset displaying different kinds of eyebrow action, gaze directed at the interlocutor, and different kinds of manual gesture behaviour

		Number of tokens	% of polar questions
Eyebrow action	no eyebrow action	207	30.4
	raised eyebrows	13	69.2
	frown	6	50.0
Head movement	nod	8	100.0
	upward movement	12	100.0
Gaze	absent	57	22.8
	present	203	42.4
Manual gestures	no gesture	200	44.5
	beats	11	0.0
	oriented towards interlocutor	6	83.3
	topic	51	9.8

The percentage of polar questions for each type of bodily behaviour is listed on the right.

our data (n = 51), accompanied polar questions rarely (9.8%). In utterances without manual action, polar questions represented 44.5% of the cases. This difference was statistically significant in a logistic regression model with utterance type as response and manual gesture type (topic vs. no gesture and other gesture types) as fixed predictor, and speaker as a random factor ($\beta = -2.15$, $z = -3.82$, $p < 0.0005$). Moreover, we found that the 11 cases of beats observed in the data occurred exclusively with continuation statements, and that 5 out the 6 utterances featuring an interactive gesture oriented towards the interlocutor (e.g., finger-pointing gestures) occurred in polar questions. Despite the small number of observations of gestures oriented towards the interlocutor, this bias was statistically significant in a mixed-effects regression model with utterance type as response, gesture type (oriented towards interlocutor vs. no gesture and other gesture types) as fixed predictor, and speaker as random factor ($\beta = 5.87$, $z = 3.56$, $p < 0.0005$).

We then investigated whether specific visual bodily cues were statistically associated with specific question types. We fitted logistic regression models with each of the question types (i.e., backward-looking vs. forward-looking, displaying surprise vs. not displaying surprise) as the response variable, the presence of the different visual cues as main predictors, and speaker as a random factor. Only one model yielded a statistically significant effect. Raised eyebrows occurred significantly more frequently in questions displaying surprise than in other questions (26.3 vs. 7.8%, respectively; $\beta = 1.47$, $z = 2.01$, $p < 0.05$). There was also a statistical trend ($p = 0.07$) for presence of gaze towards the addressee, which tended to occur more often in backward-looking than in forward-looking questions (96.7 vs. 77.1%).

3.3 Automatic Classification

The previous sections have reported a series of statistical differences between polar questions and continuation statements in terms of their phonetic realization and the bodily behaviour of the speaker. These statistical differences in group means and

probability of occurrence reflect systematic trends in the data, but are not informative about the degree of separation between the two speech act categories. Because of this, we do not know to what extent the observed differences could be useful for discrimination. In the present section, we assess the degree to which polar questions and statements can be accurately classified on the basis of the identified phonetic and visual cues by using a leave-one-out cross-validation procedure. Our procedure simulated predicting the speech act type of new unseen utterances in the following way: utterance type (polar question vs. continuation statement) was predicted for each token in the dataset by means of logistic regression models trained on the rest of the dataset, yielding a percentage of correct classification for each different model. We used different regression models including the phonetic and visual cues identified in the previous subsection and compared their accuracy. Gender was added to models that included final minimum and maximum pitch, since these variables were found to interact in our initial regression models. For the same reason, the presence of initial accents was added to models including the pitch in the penultimate syllable of the AP.

The cross-validation procedure was not run on the complete dataset for two reasons. First, the complete dataset contained utterances for which some phonetic and visual variables were left undefined (e.g., non-final pitch measures in utterances of less than three syllables, speech rate and intensity drop in one-syllable utterances, eyebrow action when eyebrows were not visible). Because of this, we only used a subset of the complete dataset in which all tokens were coded for all phonetic and visual variables. This subset was composed of 187 utterances, of which 47 were questions. Second, we wanted to control for the fact that continuation statements were almost twice as frequent as polar questions in our data. To do this, we ran 100 leave-one-out cross-validations for each model on datasets including *all* available polar questions and a *random sample* of continuation statements of a similar size and calculated the average accuracy of the 100 classifications. By doing this, we set the chance level of correct classification at 50% regardless of the prior distribution of polar questions and continuation statements in the data, and can consider any increase in accuracy as being exclusively due to the features in the model. Simulations were therefore run 100 times for each model on a subset of 94 tokens coded for all the relevant phonetic and visual variables, of which 47 were all the available polar questions and 47 randomly sampled continuation statements.

Figure 12 displays percentages of correct classification for 12 different models including phonetic and visual cues. As expected from our previous regression analyses, all models yielded rates of correct classification above chance level. The model including all phonetic cues achieved a correct classification of 76.8%. Phonetic cues alone therefore discriminated correctly over three quarters of the data. The minimum pitch in the last vowel of the AP rise performed almost as well at 74.2%, but the final maximum pitch performed significantly worse (63.4%). Regarding non-final pitch cues, initial pitch yielded 61.8%, and the pitch in the penultimate syllable performed similarly at 62.3%. Speech rate achieved 65% of correct classification, while the final intensity drop yielded a score of 59.6%. Regarding visual cues, we fitted models that incorporated information about manual gestures (a variable with four levels: no gesture vs. topic vs. beats vs. oriented towards interlocutor), head movement (nod or upward movement vs. no movement), and eye gaze, which were all present in a substantial number of tokens in our data. Eye gaze offered a small gain in performance above chance level (54.9%), whereas head movement and manual gestures performed

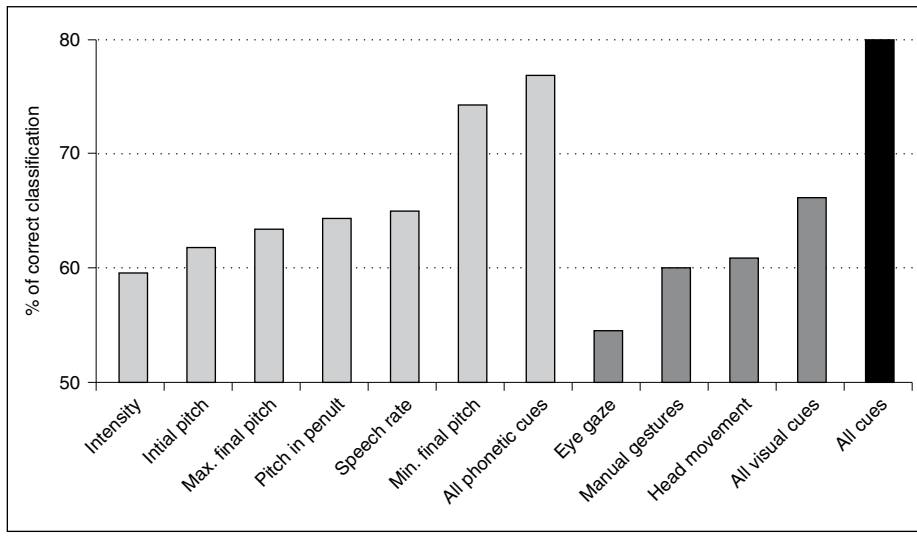

Fig. 12. Automatic classification accuracy for 12 models including different phonetic and visual cues.

significantly better (60.9 and 63.1%, respectively). Combining all visual cues resulted into a moderate increase in accuracy (66.3%). Finally, adding all these visual features to all phonetic features improved the classification accuracy of the phonetic model by 3.8% (80.6%). This suggests that, for purposes of classifying questions and continuation statements, the visual bodily behaviour features in our data were largely redundant with respect to the acoustic features.

We also fitted models including non-final cues only (initial pitch, speech rate, the presence of non-final gaze, head movements, and manual gestures). By non-final cues we mean cues that were present before the last two syllables of the utterance, where the final pitch rise invariably took place and our final intensity drop measure was taken. The model with non-final phonetic cues yielded an accuracy of 72%, and the model with non-final visual cues performed moderately above chance level (63.3%). A model combining both phonetic and visual non-final cues improved the accuracy of the phonetic model slightly, classifying correctly almost three quarters of the data (73.3%). Notice too, that, as reported above, initial pitch alone performed moderately above chance level at 61.8%. These results indicate that a moderate to substantial part of the phonetic and visual information that distinguishes questions from continuation statements in French is located before the final part of the utterance.

Finally, we fitted models to a subset of the data excluding questions displaying surprise, since these questions were more markedly different from continuation statements in pitch scaling than questions not displaying surprise and may have inflated the accuracy scores reported above. The full phonetic model yielded an accuracy of 75.6%, that is, only 1.2% worse than when applied to the dataset including all questions (i.e., displaying surprise and not displaying surprise). In the same way, the full model including both phonetic and visual cues performed only slightly worse than when applied to the full dataset (78.2 vs. 80.6%). This indicates that the predictive

power of the classification models presented earlier in this section is not solely due to the presence in the data of questions displaying surprise with a markedly higher pitch scaling than that of continuation statements.

4 Discussion and Conclusion

The main goal of this study was to estimate the amount of phonetic and visual information that distinguishes French rising polar questions and continuation statements, two types of utterances that often have similar linguistic forms despite having different pragmatic functions. From a broader point of view, we were interested in assessing the extent to which bottom-up phonetic and visual cues could play a role during online speech act production and recognition in face-to-face conversation.

In the phonetic domain, we have observed a number of statistical differences between the two utterance types. The pitch contours of questions tended to start at a slightly higher register than those of continuation statements and were less subject to declination, therefore exhibiting a higher scaling around the start of the final pitch rise. The final rise of questions tended to reach higher than that of continuation statements for male speakers, but not for females. These differences in pitch scaling between questions and continuation statements were more marked in the case of questions displaying surprise. Regarding duration, we observed that the last vowel of polar questions tended to be slightly shorter than that of continuation statements. However, the statistical difference in final vowel duration disappeared when speech rate, which was faster for questions, was taken into account. Because of this, we concluded that French polar questions tend to be spoken faster than continuation statements throughout the whole utterance, not only in their final syllable. Finally, we observed that questions had a more pronounced intensity drop in their final syllable than continuation statements. Except for the lack of differences in final pitch scaling for female speakers, our findings generally agree with previous research on the phonetic differences between the realization of questions and statements, both cross-linguistically and in French.

Regarding visual cues, we observed statistically significant differences in terms of eyebrow action, head movement, gaze behaviour, and manual gesture type. Eyebrow raising, head movements, and interactive gestures oriented towards the interlocutor were not very frequent in our data, but when present, tended to occur in questions rather than statements. On the other hand, beats, which were not frequent either, always occurred in continuation statements. Topic gestures semantically related to the message (e.g., iconic gestures), which were more frequent in our data, were common in continuation statements, but rare in questions. Speakers often directed their gaze at the recipient, both in questions and continuation statements. However, when the gaze of the speaker was not directed at the recipient, the proportion of questions decreased significantly.

The results of the automatic classifications in section 3.3 allow us to make several interesting observations. First, phonetic information, in particular the scaling of the pitch contour around the start of the final intonation rise and, to a lesser extent, the speech rate of the utterance, allows for a reasonably good degree of separation between the two kinds of utterances that we studied, with a classification accuracy over 75%. This is despite the fact that we used conversational speech materials subject to multiple sources of uncontrolled variability. Pitch scaling, for instance, is known to

be related to discourse structure (Swerts and Geluykens, 1994; Swerts, 1997; Shriberg et al., 2000) and to the emotional state of the speaker (Ladd et al., 1985). Similarly, speech rate is very likely to be subject to multiple factors ignored in this study, such as utterance complexity and utterance length (Yuan et al., 2006). The fact that the identified phonetic cues could correctly discriminate between questions and statements in over three quarters of the data in the absence of detailed lexical and contextual information therefore suggests that phonetic cues could play an important role during the online processing of speech acts in verbal interaction. Perception experiments mimicking a conversational setting, or at least offline experiments using conversational speech materials, could investigate whether and how listeners use phonetic cues to identify different utterance types such as the ones examined in this study.

Second, our findings suggest that the amount of visual information associated with questionhood in our data is somewhat smaller than that of phonetic information. Moreover, it appears that visual cues are not complementary to phonetic cues, but rather largely redundant with the speech channel, since they increase the accuracy of the phonetic-only model only slightly despite being moderately informative (they yielded almost two thirds of correct classifications by themselves). Speakers therefore do not appear to produce phonetic and visual cues in a strategic way (by increasing the amount of visual cues when phonetic cues are lacking, and vice versa), but rather produce both in a redundant manner. This is consistent with the view that communicative bodily behaviour is planned and produced closely together with speech in an integrated system (McNeill, 1992; Enfield, 2009).

Third, our findings indicate that a significant amount of bottom-up information associated with utterance type is present in the signal already before the final part of the utterance. This suggests that recipients could in principle make use of such information in order to start planning their next turns (i.e., a response to a question rather than a backchannel utterance) before the final syllables of the interlocutor's turn (van Heuven and Haan, 2002). Phonetic information and bodily behaviour can therefore be considered as potential sources of front-loaded, non-final, speech act cues, along morphosyntactic devices providing early cues to speech act type such as subject-verb inversion and wh-pronouns in English questions (Levinson, 2013).

Finally, despite the good degree of separation between utterance types achieved in our automatic classification, we would like to remind the reader that a substantial number of tokens, almost a fifth of the dataset, were classified wrongly even when all relevant phonetic and visual cues were used. This confirms our initial impressions that phonetically ambiguous and misleading cases such as the examples in figures 1 and 2 are common in conversational French. Since such a level of confusion is usually not observed in the behaviour of conversational participants, our results underline the importance of top-down information in speech act production and recognition (Geluykens, 1987). In opposition to this view, it could be argued nevertheless that our classification could be improved with more fine-grained measures and codings than the ones we used. For instance, our coding of visual cues only took into account the presence versus absence and rough timing of three types of bodily behaviour, but ignored the details of how such behaviours were implemented (e.g., more precise timing and direction of the speaker's gaze, amplitude and velocity of manual gestures). A better classification might also be obtained if the pragmatic coding of utterances was done at a finer level of granularity, for instance by distinguishing more subtypes of questions than we did in our coding (e.g., repair initiators,

follow-up questions, requests for action). The extent to which a better separation between utterance types can be obtained by using more detailed measures and codings is an open question, which could be investigated with more accurate instrumental techniques (e.g. eye trackers, optical markers) and a larger database containing more instances of different subtypes of questions and statements than were available to us in the present study.

In summary, we have shown that a significant amount of phonetic and visual information distinguishes two different pragmatic functions encoded with a similar linguistic form. This should be taken into account by models of speech act production and recognition in verbal interaction. We also note that a moderately sized collection of utterances extracted from a corpus of spontaneous conversation could be used to study the phonetic realization of different action types. Future studies on this topic should therefore consider the option of using uncontrolled conversational data alongside, or instead of, laboratory-controlled reading experiments.

Acknowledgements

This work was made possible thanks to the financial support of the Language and Cognition Department at the Max Planck Institute for Psycholinguistics, Max-Planck-Gesellschaft, and a European Research Council's Advanced Grant (269484 'INTERACT') to Stephen C. Levinson. We would like to thank Marisa Casillas, Binyam Gebrekidan, Seán Roberts, Giovanni Rossi, and the members of the Language and Cognition Department at the Max Planck Institute for Psycholinguistics for useful comments and discussion.

References

Andruski JE, Costello J (2004): Using polynomial equations to model pitch contour shape in lexical tones: an example from Green Mong. J Int Phon Assoc 34:125–140.

Bavelas JB, Chovil N, Lawrie DA, Wade A (1992): Interactive gestures. Discourse Processes 15:469–489.

Boersma P, Weenink D (2014): Praat: Doing Phonetics by Computer; version 5.4.04 (computer program). Amsterdam, Institute of Phonetic Sciences, University of Amsterdam.

Bolinger DW (1989): Intonation and Its Uses. Palo Alto, Stanford University Press.

Borràs-Comes J, Kaland C, Prieto P, Swerts M (2014): Audiovisual correlates of interrogativity: a comparative analysis of Catalan and Dutch. J Nonverbal Behav 38:53–66.

Cangemi F, D'Imperio M (2011): Local speech rate differences between questions and statements in Italian. Proc 17th Int Congr Phon Sci, Hong Kong, pp 392–395.

Delais-Roussarie E, Post B, Avanzi M, Buthke C, Di Cristo A, Feldhausen I, Jun SA, Martin P, Meisenburg T, Rialland A, Sichel-Bazin R, Yoo HY (in press): Intonational phonology of French: developing a ToBI system for French; in Frota S, Prieto P (eds): Intonation in Romance. Oxford, Oxford University Press.

Delattre P (1966): Les dix intonations de base du français. French Rev 40:1–14.

Di Cristo A (1998): Intonation in French; in Hirst DJ, Di Cristo A (eds): Intonation Systems. A Survey of Twenty Languages. Cambridge, Cambridge University Press, pp 195–218.

Edlund J, Strömbergsson S, House D (2012): Telling questions from statements in spoken dialogue systems. Proc SLTC 2012.

Ekman P (1979): About brows: emotional and conversational signals; in von Cranach M, Foppa K, Lepenies W, Ploog D (eds): Human Ethology. Cambridge, Cambridge University Press, pp 169–248.

Enfield NJ (2009): The Anatomy of Meaning. Cambridge, Cambridge University Press.

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

Geluykens R (1987): Intonation and speech act type: an experimental approach to rising intonation in queclaratives. J Pragmatics 11:483–494.

Griffin ZM, Bock K (2000): What the eyes say about speaking. Psychol Sci 11:274–279.

Grundstrom A (1973): L'intonation des questions en français standard. Studia Phonetica 8:19–49.

Gussenhoven C (2002): Intonation and interpretation: phonetics and phonology. Proc Speech Prosody 2002.

Gussenhoven C, Chen A (2000): Universal and language-specific effects in the perception of question intonation. Proc Interspeech 2000.
Haan J (2002): Speaking of questions: an exploration of Dutch question intonation; doct diss Utrecht University.
Heritage J (2012): Epistemics in action: action formation and territories of knowledge. Res Lang Soc Interact 45:1–29.
Indefrey P, Levelt W (2004): The spatial and temporal signatures of word production components. Cognition 92:101–144.
Jescheniak JD, Schriefers H, Hantsch A (2003): Utterance format affects phonological priming in the picture-word task: implications for models of phonological encoding in speech production. J Exp Psychol Hum Percept Perform 29:441–454.
Jun SA, Fougeron C (2000): A phonological model of French intonation; in Botinis A (eds): Intonation: Analysis, Modeling and Technology. Dordrecht, Kluwer, pp 209–242.
Ladd DR, Silverman KE, Tolkmitt F, Bergmann G, Scherer KR (1985): Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. J Acoust Soc Am 78:435–444.
Levinson SC (1983): Pragmatics. Cambridge, Cambridge University Press.
Levinson SC (2013): Action formation and ascription; in Stivers T, Sidnell J (eds): Handbook of Conversation Analysis. Hoboken, Wiley-Blackwell, pp 103–130.
McNeill D (1992): Hand and Mind. Chicago, University of Chicago Press.
Ohala JJ (1984): An ethological perspective on common cross-language utilization of F0 of voice. Phonetica 41:1–16.
Post B (2002): French tonal structures. Proc Speech Prosody 2002, Aix-en-Provence.
Purson A, Santi S, Bertrand R, Guaïtella I, Boyer J, Cavé C (1999): The relationships between voice and gesture: eyebrows movements and questioning. Proc Eur Conf Speech Commun and Technol, Budapest.
Rialland A (2007): Question prosody: an African perspective; in Riad T, Gussenhoven C (eds): Tones and Tunes: Typological Studies in Word and Sentence Prosody. Berlin, Mouton De Gruyter, pp 35–63.
Rossano F, Brown P, Levinson SC (2009): Gaze, questioning and culture; in Sidnell J (ed): Conversation Analysis: Comparative Perspectives. Cambridge, Cambridge University Press, pp 187–249.
Rossi M, Hirst D, Di Cristo A (1981): Continuation and question; in Rossi M, Di Cristo A, Hirst D, Martin P, Nishinuma Y (eds): L'intonation: de l'acoustique à la sémantique. Paris, Klincksieck, pp 149–177.
Sacks H, Schegloff EA, Jefferson G (1974): A simplest systematics for the organization of turn-taking for conversation. Language 50:696–735.
Shriberg E, Stolcke A, Hakkani-Tür D, Tür G (2000): Prosody-based automatic segmentation of speech into sentences and topics. Speech Commun 32:127–154.
Sloetjes H, Wittenburg P (2008): Annotation by category – ELAN and ISO DCR. Proc 6th Int Conf on Lang Resources and Evaluation.
Smith CL (2002): Prosodic finality and sentence type in French. Lang Speech 45:141–178.
Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, Heinemann T, Hoymann G, Rossano F, de Ruiter JP, Yoon KE, Levinson SC (2009): Universals and cultural variation in turn-taking in conversation. Proc Natl Acad Sci U S A 106:10587–10592.
Stivers T, Rossano F (2010): Mobilizing response. Res Lang Soc Interact 43:3–31.
Swerts M (1997): Prosodic features at discourse boundaries of different strength. J Acoust Soc Am 101:514–521.
Swerts M, Geluykens R (1994): Prosody as a marker of information flow in spoken discourse. Lang Speech 37:21–43.
Torreira F (2007): Tonal realization of syllabic affiliation in Spanish. Proc ICPhS XVI, Saarbrücken, pp 6–10.
Torreira F, Adda-Decker M, Ernestus M (2010): The Nijmegen Corpus of Casual French. Speech Commun 52:201–212.
Torreira F, Ernestus M (2010): Phrase-medial vowel devoicing in spontaneous French. 11th Annu Conf Int Speech Commun Assoc, Interspeech, pp 2006–2009.
Valtersson E, Torreira F (2014): Rising intonation in spontaneous French: how well can continuation statements and polar questions be distinguished? 7th Int Conf on Speech Prosody, Speech Prosody 7, pp 785–789.
van Heuven V, Haan J (2002): Temporal distribution of interrogativity markers in Dutch: a perceptual study; in Gussenhoven C, Warner N (eds): Papers in Laboratory Phonology. Berlin, Mouton de Gruyter, vol 7, pp 61–86.
van Heuven VJ, van Zanten E (2005): Speech rate as a secondary prosodic characteristic of polarity questions in three languages. Lang Speech 47:87–99.
Yuan J, Liberman M, Cieri C (2006): Towards an integrated understanding of speaking rate in conversation. Proc Interspeech 2006.