Research Article

The prosodic marking of rhetorical questions in Standard Chinese

Katharina Zahner-Ritter a,b,*, Yiya Chen c, Nicole Dehé b, Bettina Braun b

a University of Trier, Phonetics, Universitätsring 15, 54296 Trier, Germany
b University of Konstanz, Department of Linguistics, P.O. Box 186, 78457 Konstanz, Germany
Leiden University Center for Linguistics, Postbus 9515, 2300 RA Leiden, the Netherlands

A R T I C L E  I N F O

Article history:
Received 12 January 2021
Received in revised form 1 September 2022
Accepted 7 September 2022

Keywords:
Lexical tone
Prosody
Intonation
Rhetorical questions
Standard Chinese

A B S T R A C T

The present study investigates the prosody of information-seeking (ISQs) and rhetorical questions (RQs) in Standard Chinese, in polar and wh-questions. Like in other languages, ISQs and RQs in Standard Chinese can have the same surface structure, allowing for a direct prosodic comparison between illocution types (ISQ vs RQ). Since Standard Chinese has lexical tone, the use of f0 as a cue to illocution type may be restricted. We investigate the prosodic differences between ISQs and RQs as well as the interplay of prosodic cues to RQs. In terms of f0, results showed that RQs were lower in f0, with the f0 range on the first word being expanded followed by f0 compression. RQs were further longer in duration and more often realized with non-modal voice quality (glottalized voice) as compared to ISQs. These prosodic cues were largely manipulated in tandem (illocutionary pairs with larger durational differences also showed larger differences in mean f0; voice quality, in turn, seemed to be an additional cue). We suggest three possible explanations (assertive force, focus, speaker attitude) that unite the present findings on RQs in Standard Chinese with the findings on RQs in other, non-tonal languages.

1. Introduction

Questions such as Who eats lemons? may serve more than one function in discourse, two of which are of interest here. First, as an information-seeking question (henceforth ISQ), the interrogative aims at requesting information from the interlocutor. Second, as a rhetorical question (henceforth RQ), it serves to make a point, here to convey that nobody eats lemons (Biezma & Rawlins, 2017; Caponigro & Sprouse, 2007; Han, 2002 on ISQ vs RQ). The present paper investigates the prosodic differences between string-identical ISQs and RQs in Standard Chinese, as well as the interplay between or combination of the cues to RQs. We compared the prosodic realizations in two question types, namely polar questions (1) and constituent (henceforth wh-) questions (2). Understanding the prosodic realization of RQs in Standard Chinese will help us gain a broader and generalizable knowledge of question types across languages and their realizations.

As (1) shows, in Standard Chinese, polar questions typically use particles such as 吗 ma, 么 me, and 嘛 ma that mostly occur in sentence-final position (Chao, 1968; Liang, 2014). The sentence-final particle 么 me, for instance, can turn a statement...
into a polar question. 么 me may be used in both ISQs and RQs. Regarding wh-questions, Standard Chinese is a wh-in-situ language (Cheng, 1991). That is, wh-elements surface in their syntactic base position. In our study, we only use wh-questions in which the wh-word is the subject and thus precedes verb and object. For both polar and wh-questions, the same string of words can be used to produce a question with an information-seeking or a rhetorical illocution. Clearly, polar questions differ from wh-questions with respect to their semantics and syntax (Gronenjijk & Stokhof, 1984, pp. 1744–1747; Krifka, 2011), cf. (1) and (2). Also, in our experiment, the sentence-final particle 么 me occurs in polar questions only, leading to differences in the number of syllables across question type. Crucially, the main comparison of our study is between ISQs and RQs, which essentially are string-identical within each question type (polar vs wh-questions).

Previous research on a variety of different languages has shown that ISQs and RQs differ in a number of prosodic characteristics: f0, duration, and voice quality (Dehé et al., 2022 for a recent overview). These differences hold true for both polar and wh-questions. In regard to the use of f0 for the marking of illocution type, lexical tone languages provide a particularly interesting test case for the distinction between ISQs and RQs, since tone is contrastive at the lexical level and phrase-level f0 is therefore constrained by the canonical form of the lexical tone (e.g., Gussenhoven, 2004). Studying the marking of illocution type in tone languages hence allows us to contribute to our understanding of the use of f0 beyond its primary function of marking lexical tone (cf. Chang, 1975). In the present paper, we focus on Standard Chinese by speakers who were born and grew up in Beijing with Standard Chinese – a variety in which every syllable carries one of four lexical tones, or the neutral tone (Chao, 1930, 1956; Chen, 2016, 2022; Lin, 2007), see below in Section 1.2.

In the remainder of Section 1, we first summarize the main findings on the prosodic differences of ISQs vs RQs for the languages on which experimental studies have been conducted (Section 1.1). Section 1.2 reviews the literature on the use of f0 and other prosodic cues for lexical and non-lexical purposes in Standard Chinese. Based on this background, Section 1.3 outlines the research questions and hypotheses. Section 2 presents the methods of the production experiment, Section 3 its results. Section 4 discusses the results in the context of acoustic cue weighting and other non-lexical uses of prosody in Standard Chinese. From a broader, cross-linguistic context, it also includes findings from typologically different languages to discuss cross-linguistic signals of rhetorical questions (Section 4.1). We finally discuss implications of the interplay between prosodic cues to RQs for the modelling of the relation between prosody and meaning (Section 4.2), and conclude in Section 5.

1.1. Rhetorical questions and how they are signalled

ISQs constitute a directive speech act of requesting information from the addressee (Gronenjijk & Stokhof, 1984; Krifka, 2011), also known as neutral, real, or genuine questions. RQs may share their syntactic surface form with ISQs but attempt to commit the interlocutor to the answer that is presupposed in the RQ (cf. Biezma & Rawlins, 2017; Han, 2002, p. 202, who considers RQs as assertions), e.g., nobody likes lemons in examples (1) and (2). Signals to rhetorical illocution are, among others, shared world-knowledge, e.g., Is the Pope catholic? (Han, 2002, p. 216), syntactic cues such as strong negative polarity items, e.g., “ever” in What has John ever done for Sam? (Han, 2002, p. 202), or lexical cues such as discourse particles, e.g., German schon (Bayer & Obenauer, 2011, p. 454; see also Dehé, Wochner, & Einfeldt, 2022). For Standard Chinese, B. Xu (2013) argues that questions containing 嘗嘗 nándào necessarily have a rhetorical illocution, see (3). Fang (2021) has recently also argued that 嘗嘗 may signal a contradiction between the stated message and an existing assumption, giving rise to a rhetorical illocution.

(3) 
Nándào shuí bang-guò ní ma?
Nandao who help-EXP you PRT


Given that these cues are optional, an interrogative may be ambiguous between ISQ or RQ meaning in Standard Chinese. Previous work has shown that prosodic cues can distinguish string-identical ISQs and RQs in production, particularly in regard to three prosodic dimensions, i.e., f0, duration, and voice quality (Dehé et al., 2022, for overview), but most of these studies focused on intonation languages, for which the following has been found:

- RQs are longer (or realized with a slower speaking rate) as compared to ISQs in a variety of different languages – including German (Braun et al., 2019; Braun, Einfeldt, Esposito, & Dehé, 2020), English (Dehé & Braun, 2020b), Icelandic (Dehé, Braun, & Wochner, 2018; Dehé & Wochner, 2022), French (Beyssade & Delais-Roussarie, 2018), Italian (Sorianello, 2018, 2019), and Estonian (Sahkai, Asu, & Lippus, 2022).
- For German, English, and Icelandic, more instances of breathy voice have been found for RQs as compared to ISQs (Braun et al., 2019; Dehé & Braun, 2020b; Dehé & Wochner, 2022), while for Estonian, more instances of glottalized voice have been observed for RQs compared to ISQs (Sahkai et al., 2022).
- RQs also differ from ISQs in the position of the nuclear pitch accent (Dehé & Braun, 2020b for English; Sahkai et al., 2022 for Estonian), the type of pitch accent (Beyssade & Delais-Roussarie, 2022 for French; Braun et al., 2019 for German; Dehé & Braun, 2020a for Icelandic; 2020b for English; Sorianello, 2019 for Italian), and the types of final edge tones (Beyssade & Delais-Roussarie, 2022 for French; Braun et al., 2019 for German; Dehé & Braun, 2020b for English; Sahkai et al., 2022 for Estonian; Sorianello, 2019 for Italian).

In Japanese, a pitch accent language, RQs have been shown to be longer and lower in overall f0 than ISQs; initial lowering is furthermore a strong perceptual indicator for RQs in Japanese (Miura & Hara, 1995).

Research on the prosody of RQs has only recently included tone languages – but has so far been restricted to one question type: wh-questions (Lo & Kiss, 2020; Lo, Kiss, & Tulling, 2019b). Tone languages are particularly interesting since they pose questions for the interaction between lexical tone and post-lexical intonation in the marking of illocution type (Chen, 2022 for overview). For Cantonese, a tone language with six lexical tones (Zhang, Duanmu, & Chen, 2021 for overview),
Lo, Kiss, and Tulling (2019a) investigated the sentence-final particles in wh-questions and found them to be longer and lower in RQs than in ISQs. Lo and Kiss (2020) furthermore studied wh-questions in Mandarin Chinese and found wh-RQs to be overall longer than string-identical ISQs, except for the sentence-final particle. Moreover, the sentence-final particle was lower in \( f_0 \) in RQs as compared to ISQs, and more often realized with glottalized voice; in turn, the wh-word was higher in RQs as compared to ISQs. No \( f_0 \) differences have been reported for the middle part of the sentence.

Hence, lexical tone languages distinguish different illocution types in wh-questions using duration, \( f_0 \), and – for sentence-final particles – voice quality. So far, nothing is known on whether this prosodic marking generalizes to polar questions, whether it is limited to the realization of the wh-word and the sentence-final particle, and whether speakers use the different kinds of cues to RQs in a compensatory manner (trading relation) or in tandem (cf. Schertl & Clare, 2019). For tone languages, in particular, it might be the case that adjustments in \( f_0 \) (given its primary function of marking lexical tone) are limited and get compensated by larger differences in other prosodic cues. Such limited adjustments in \( f_0 \) have been observed by Chen and Gussenhoven (2008) for the realization of different levels of emphasis. In particular, when speakers of Standard Chinese were encouraged to produce different levels of emphasis, they tended to lengthen more without further modification of the \( f_0 \) range. In the present study, we test this possibility by examining the interplay between different prosodic cues to RQs. As we will briefly summarize in the following section, beyond its lexical function, \( f_0 \) also serves post-lexical functions in Standard Chinese (Xu, 2019; Zhang et al., 2021, for overviews).

1.2. Lexical and non-lexical functions of \( f_0 \) and other prosodic cues in Standard Chinese

Standard Chinese is a tone language in which every syllable carries one of four lexical tones: Tone 1 (T55, high-level, ě), Tone 2 (T35, rising, ě), Tone 3 (T214, low-rising, ě) and Tone 4 (T51, falling ě), or the neutral tone (Chao, 1930, 1956; Chen, 2016, 2022; Lin, 2007). Tone 3 is also frequently associated with creaky voice (e.g., Chao, 1956, p. 53; Kuang, 2017, and references therein, p. 1694). \( f_0 \) hence primarily serves a lexical function in Standard Chinese, such that a change in lexical tone leads to a change in lexical meaning. The canonical shape of the tones is directly evident when tones are produced in isolation, and tones in multi-syllabic phrases are typically influenced by preceding or following tones (e.g., Shen, 1990; Xu, 1997; Xu & Liu, 2006). In addition to the lexical function, \( f_0 \) (and other prosodic cues) are also used to convey post-lexical functions (Xu, 2019; Zhang et al., 2021, for overviews).

At the post-lexical level, \( f_0 \) is used to mark information structure, speech acts or affective states (Chen, 2022; Xu, 2019; Zhang et al., 2021, for overviews). Regarding information structure, tones on focused words are typically realized with a greater \( f_0 \) range than words in non-focal position (Jin, 1996; Liu & Xu, 2005; Xu, 1999; Chen and Braun, 2006) and with longer durations (Chen, 2006; Chen & Gussenhoven, 2008; Jin, 1996; Xu, 1999). Tonal and segmental contrasts are hyperarticulated with respect to their distinctive characteristics (Chen, 2008; Chen & Gussenhoven, 2008). Generally speaking, focus-induced \( f_0 \) adjustments have been reported to apply to the whole utterance, with the focused element being expanded in \( f_0 \) and the region thereafter being compressed,4 a mechanism termed post-focal compression (Gårding, Zhang, & Svantesson, 1983; Jin, 1996; Xu, 1999; Xu & Xu, 2005; Zhang et al., 2021, for overview).

Beyond focus, prosodic cues are used to convey emotions or affective states (Li, Fang, & Dang, 2011; Liu & Pell, 2012; Yuan, Shen, & Chen, 2002). “Disgust,” for instance, is associated with a lowering in \( f_0 \) (Li et al., 2011; Liu & Pell, 2012), a slower speech rate (Liu & Pell, 2012), and low harmonics-to-noise ratio values (Liu & Pell, 2012), hinting to the use of non-modal voice quality for this emotion (Keating, Garellek, & Kreiman, 2015).

Importantly, prosody, in particular \( f_0 \) and duration, also marks speech acts, such as the difference between statements and questions in string-identical utterances, see (4).

<table>
<thead>
<tr>
<th></th>
<th>有人(yóurén)</th>
<th>吃(chī)</th>
<th>柠檬(níngměng)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anyone/Somebody</td>
<td>eat</td>
<td>lemon</td>
<td></td>
</tr>
</tbody>
</table>

'Somebody eats lemons.' or 'Does anyone eat lemons?'

Specifically, polar-ISQs are globally produced with higher overall \( f_0 \) than string-identical declaratives (Lee, 2005; Liu & Xu, 2005; Yuan, 2006), with the difference in \( f_0 \) becoming larger towards the end of the utterance (Yuan, 2006). Contrary to Cantonese, where questions end in a final rise irrespective of the lexical tone, Standard Chinese is more faithful to the shape of the lexical contour at the end of the utterance. This means, for instance, that the falling Tone 4 is still a falling tone in questions, but the range of the fall is reduced; conversely, Tone 2 is rising also in questions, but compared to declaratives, it is realized with an enhanced \( f_0 \) range (Chen, 2022; Zhang et al., 2021). In terms of duration, except for the last syllable, syllables have been shown to be shorter in polar questions than in declaratives (Yuan, 2006).

Prosody also distinguishes between wh-questions and string-identical declaratives. In wh-questions, shénme is the wh-pronoun ‘what’, while in declaratives, together with the licensor dian ‘(a little), it is an indefinite / existential, meaning ‘a little bit of something’ (cf. Yang, 2018). In production, \( f_0 \) and other prosodic cues distinguish between the two readings: For example, in sentences containing shénme, wh-ISQs exhibit higher \( f_0 \) compared to string-identical declaratives, mostly towards the end of the utterance (Liu, Li, & Jia, 2016; Yang, 2018; Yang, Gryllia, & Cheng, 2020). Yang (2018) further shows an increased \( f_0 \) range in shénme ‘what’ for questions. Additionally, utterance and word durations are shorter in wh-ISQs than in declaratives, with an exception of shénme ‘what’.

---

3 The numbers (e.g., 51 for the falling Tone 4) indicate the pitch levels involved in the tonal movement, with 1 being at the low end and 5 at the high end; diacritics (e.g., ě) are placed on vowel (here ‘ě’) and indicate the direction of the tonal movement, see International Phonetic Alphabet for all diacritics, see https://www.internationalphoneticsassociation.org/content/full-ipa-chart (last access: 18 November 2021).

4 The \( f_0 \) expansion and compression effects of focus are also constrained by the tonal context (e.g., Chen, 2010) and prosodic structure of the focused element (see review in Chen, Lee, & Pan, 2016).
for which the pattern was reversed (Yang, 2018); see also Yang et al. (2020).

Taken together, Standard Chinese uses f0 modifications to express post-lexical functions on top of lexical tone – a phenomenon that has been referred to as the “multiplexing of the f0 channel (Zhang et al., 2021, p. 9, chapter 24.5). In addition to f0, other prosodic parameters such as duration and voice quality also serve post-lexical functions in Standard Chinese, particularly with respect to the marking of focus, sentence type and attitudes. One of the remaining questions is how Standard Chinese employs f0 and other prosodic cues to differentiate between string-identical ISQs and RQs.

1.3. Research questions and hypotheses

The present study investigates the prosodic marking of RQs as compared to ISQs in Standard Chinese, using prompted productions of target questions (polar and wh-questions). We used an experimental paradigm that has been employed for other languages (Braun et al., 2019; Dehé & Braun, 2020a, 2020b) and adapted it to Standard Chinese: Chinese participants read short contexts (which described different situations) followed by target interrogatives. The contexts were created such that they either triggered an ISQ or an RQ illocution. Target questions were string-identical in both readings (ISQ and RQ). Our two main Research Questions (Qs) are the following:

- **Q1:** Do string-identical ISQs and RQs in Standard Chinese prosodically differ from each other, both in polar and in wh-questions, and if so, what are the prosodic cues?
- **Q2:** Are prosodic cues that distinguish illocution type used in a compensatory manner or are they modified in tandem?

With respect to Q1, we analysed f0, duration, and voice quality, which appear to be the main cues cross-linguistically (Dehé et al., 2022). Based on previous work (see 1.1), we put forward the following hypotheses:

- **H1a:** RQs will be overall lower in their f0 trajectory than ISQs.
- **H1b:** RQs will be longer in their duration than ISQs, for all words.
- **H1c:** RQs will more often be realized with non-modal voice quality than ISQs.

With respect to Q2, we test whether cues are used in a compensatory manner (trading relation), such that smaller adjustments in f0 (which is the primary cue to signal tone) are compensated by larger differences in other prosodic cues, such as duration or voice quality (cf. Schertz & Clare, 2019). Here, we focus on illocutionary pairs consisting of an ISQ and its corresponding RQ. A trading relation is present between f0 and other prosodic cues when f0 does not differ between an ISQ and its corresponding RQs, but duration and voice quality do. A modification in tandem is present when all cues (f0, duration, and voice quality) differ between an ISQ and its corresponding RQ.

---

2. Methods

The experiment was run in Beijing in spring 2018. The study was approved by the Ethics Committee of the University of Konstanz (Institutional Review Board number: IRB 30/2016).

2.1. Materials

Twenty-two polar and 22 wh-questions were constructed, along with two contexts for each question (one context eliciting an ISQ reading, the other an RQ reading), resulting in 22 context-question quadruplets, see Table 1. The quadruplets were translated from Braun et al. (2019) by a native speaker of Standard Chinese. To compare the results across languages, we aimed at maximal comparability to our work on the prosody of RQs in German, English, and Icelandic (Braun et al., 2019; Dehé & Braun, 2020a, 2020b) – in terms of a) the semantics of the situation (context) and the structure and content of the target interrogatives, and b) in terms of the statistical power (number of items). We made six changes for cultural reasons and took care to balance lexical tone in the final syllable of the noun, which is the final syllable in the sentence in wh-questions and the penultimate in polar questions, which included a sentence-final particle. Polar questions started with yìyùrèn (‘anyone’) and wh-questions with shéi (‘who’), both followed by a verb, and an object consisting of one noun, plus the particle in polar questions, see Table 1 for an example quadruplet and Appendix A for a complete list of target questions of the present experiment. The verbs and objects in the target questions were of different length (verbs were either mono- or bisyllabic; object nouns consisted of one to five syllables, see Appendix A). The ISQ version of the context always contained the sequence ‘you would like to know’, in accordance with the property of ISQs that they seek information. By contrast, the RQ version of the context always contained ‘it is known that’, indicating that the answer is obvious to the interlocutors. Interrogatives were felicitous in both illocution types and contained an object noun that was non-constraining as to one of the readings (e.g., lemons), as verified by a pre-test conducted online. In particular, native speakers of Standard Chinese indicated for all 22 items whether they agreed or disagreed with the proposition in the interrogative. On average, participants agreed in 55.3% of the cases (with a range of 13.3% to 83.3% across individual items). Hence, the propositions in the interrogative sentences were on average ambiguous as to one of the two illocution types.

Polar questions always started with the pronoun subject yìyùrèn ‘anyone’; wh-questions started with the question word shéi ‘who’. Consequently, within question types, the first word (the subject of the sentence) was always the same – and hence always carried the same lexical tones. Verbs and object nouns varied in terms of number of syllables and lexical tone, see Appendix A. The end of the utterance was balanced for lexical tone: For the object noun, lexical tone was distributed such that all four tones occurred in the last syllable of the noun (six times Tone 1, six times Tone 2, four times Tone 3, and six times Tone 4, in both polar and wh-questions). Note that different tones were used for reasons of generalizability and to avoid confounding with respect to the use of voice quality.
and duration (Kuang, 2017; 2018, among others). However, due to semantic constraints we only controlled tone at the beginning and end of the utterance. Furthermore, to have comparable conditions, polar and wh-questions contained the same predications (e.g., eating lemons). The particle 么, me, which is commonly used in polar questions (Ling, 2014), was included in both polar-ISQs and polar-RQs. There was no question particle for wh-questions. Since we focus on the difference between ISQs and RQs, the difference in structure in polar and wh-question is secondary.

Additionally, 34 fillers and their contexts (declaratives with attachment ambiguities, exclamatives, alternative questions and neutral polar questions) were translated from Braun et al. (2019) and used in the present experiment.

### 2.2. Procedure

Two experimental lists were constructed, each containing both question types (polar and wh) and both illocution types (ISQ and RQ). Each list contained half of the polar questions (N = 22; 11 in an ISQ and 11 in an RQ reading) and half of the wh-questions (N = 22; 11 in an ISQ and 11 in an RQ reading) and all 34 fillers. Illocution type (ISQ vs RQ) was manipulated within-subjects, i.e., each participant produced both the ISQ version and the RQ version of each target interrogative. Each list contained only one question type of each illocutionary pair (ISQ and RQ), either the polar or the wh-question. One of the two lists was randomly assigned to each participant. Each participant received a randomized order of items, by the constraint of separating the same question (in the two readings) by at least four other items. Three practice trials preceded 78 trials (44 experimental and 34 fillers). Participants received oral instructions in Standard Chinese by the experimenter, a research assistant who is a native speaker of Standard Chinese. The experiment was controlled in Presentation (Presentation, 2000).

On each trial, participants silently read a context displayed on a computer screen. Upon button press, the target interrogative appeared on the screen and the recording started. Participants were instructed to read each context carefully and to produce the subsequent interrogatives in a way that was suitable in the given context. They were allowed to produce the sentence again, if needed. Upon another button press, a new trial started. Productions were recorded using a headset microphone (Shure SM10A) and digitized onto a computer (44.1 kHz, 16 Bit). The experimenter did not interfere during the experiment. Testing took place in a quiet room and the experiment lasted about 25–30 minutes.

### 2.3. Participants

Ten native speakers of Standard Chinese (all female, average age = 26.5 years; SD = 2.0 years) born and raised in Beijing with Standard Chinese participated in the study. Two additional speakers born and raised elsewhere were excluded from the present analysis in order to minimize potential influence of dialectal variation.

### 2.4. Data preparation and annotation

In total, 440 target interrogatives were produced (44 contexts × 10 participants). Twenty-two interrogatives (3.9%) were excluded from the analysis because of technical errors (N = 2), mispronunciations (N = 8), or pauses / hesitations between the words (N = 12). The final data set (N = 418) comprised 212 polar questions (106 ISQs, 106 RQs) and 206 wh-questions (103 ISQs, 103 RQs).

All interrogatives were annotated in Praat (Boersma and Weenink, 2016) on three tiers: on the syllabic level, the word level, and with respect to voice quality (at the beginning and end of the utterance), see Fig. 1. Segmental boundaries were manually placed by a native speaker of Standard Chinese based on standard segmentation criteria (Turk, Nakai, & Sugahara, 2006). Pitch tracking errors were manually corrected (first author) by removing erroneous pitch points in the Praat Manipulation editor (Boersma & Weenink, 2016) and saving the modified Manipulation-Object as a wav-file (Pitch overlap-add). The corrected files were used for further processing and analyses. F0 values of the f0 trajectory over time were automatically extracted from the files with corrected f0 using the Praat script ProsodyPro (Y. Xu, 2013). Specifically, we extracted ten measurements (in Hz) from each word.6

### Table 1

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Example context and question quadruplet in an ISQ (left) and RQ reading (right), for a polar (upper panel) and a wh-question (lower panel).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Context for ISQ</strong></td>
<td><strong>Context for RQ</strong></td>
</tr>
<tr>
<td>At a party, you offer cake made with lemons. You would like to know which of the guests like this fruit and whether they would like some or not.</td>
<td>Your aunt offers lemons to her guests. However, it is known that this fruit is too sour to be eaten on its own.</td>
</tr>
<tr>
<td>You say to your guests: 你 (niúrén) 吃 (chī) 柠檬 (níngmèng) 么 (me)? sentence-final particle</td>
<td>You say to your cousin: 你 (niúrén) 吃 (chī) 柠檬 (níngmèng) 么 (me)? sentence-final particle</td>
</tr>
<tr>
<td><strong>Who eats lemons?</strong></td>
<td><strong>Who eats lemons?</strong></td>
</tr>
</tbody>
</table>

6 Note that ProsodyPro uses interpolation for unvoiced portions of the signal. From our perspective, this is entirely unproblematic for our analyses, given that the voiceless sounds in our materials are mostly voiceless fricatives, which have been show to carry f0 information for the human perceptual system ("segmental" intonation, Niebuhr, 2012, 2017). Importantly, voiceless portions are the same in both illocution types (ISQ and corresponding RQ), which is the main interest of our study.
(a) Representative polar question pair

![Polar ISQ](image1)

![Polar RQ](image2)

(b) Representative wh-question pair

![Wh ISQ](image3)

![Wh RQ](image4)

Fig. 1. Representative polar (a) and wh-question pair (b), ISQ top panel, RQ bottom panel. Tier 1 shows the syllable representation in Standard Chinese with the respective information on lexical tone; tier 2 gives the English word-by-word translation; tier 3 indicates the voice quality classification at the first word and the last syllable(s); tier 4 provides the English translation of the question.
Voice quality was annotated by two separate annotators (first author and native Chinese student assistant), based on perceptual classification, as modal, breathy, or glottalized voice quality (Braun et al., 2019; Laver, 1980). To avoid effects of tonal identity on voice quality (Kuang, 2017; 2018, among others), we labelled voice quality only at the beginning and the end of the utterance, where we controlled or balanced for lexical tone:

- Polar questions: one label for yějūrén ‘anyone’; one label for the last syllable of the object noun, and one for the sentence-final particle (i.e., three voice quality labels in total).
- Wh-questions: one label for shéi ‘who’ and one for the last syllable of the object noun (i.e., two voice quality labels in total).

In total, 1048 voice quality labels were set (\(N = 636\) in polar question, i.e., three labels in 212 analysed productions, and \(N = 412\) in wh-questions, i.e., two labels in 206 analysed productions). We checked the reliability of the voice quality labelling based on 39.9% of the voice quality labels (i.e., 418 productions). We checked the reliability of the voice quality labelling based on 39.9% of the voice quality labels (i.e., 418 labels). The two labellers agreed in 98.1% of the cases, \(\kappa = 0.95\), “almost perfect” (Cohen, 1960; Gamer, Lemon, Fellows, & Singh, 2012; Landis & Koch, 1977).

2.5. Statistical analysis

This section gives an overview of the statistical analyses, which were done separately for each question type (polar and wh-questions). We also plot the results separately for each question type, even though the effects of illocution type were independent of question type, i.e., are comparable in both question types. Analysis scripts are available at Mendeley upon publications (https://data.mendeley.com/datasets/49nv-

Research Question 1 (prosodic differences between ISQs and RQs, separately for each question type): The main analysis concerned the global prosodic differences between illocution types (ISQ vs RQ) in polar and wh-questions with respect to the three prosodic parameters: a) the \(f_0\) trajectory over the target question, b) utterance (and word) durations, and c) voice quality at the beginning and end of the target question (cf. Q1, H1a-c).

To investigate differences in the \(f_0\) trajectory over the target question between ISQs and RQs (cf. H1a), general additive mixed modelling was applied (GAMMs, Wieling, 2018; Wood, 2006, 2017), separately for question type (polar and wh-questions). We extracted ten \(f_0\) values for each word in a question, and the resulting time-normalized \(f_0\) contours were compared across illocution type (ISQ vs RQ). GAMMs were chosen for the analysis of the \(f_0\) trajectory as they represent an optimal way for the analysis of time-varying data with non-linear relationships and auto-correlation (Baayen, van Rij, de Cat, & Wood, 2018; Wieling, 2018; for a comparison of intonation contrats using GAMM, see Zahner-Ritter, Einfeldt, et al., 2022; Zahner-Ritter, Zhao, Einfeldt, & Braun, 2022). In brief, GAMMs model non-linear dependencies in \(f_0\) and illocution type over time via smooth functions. These smooth functions include a pre-specified number of base functions of different shapes, e.g., linear and parabolic functions of different complexity (e.g., Wieling, 2018). Fixed effects are modelled in the same way as in linear mixed effect regression models. In addition, GAMMs also model non-linear effects over time. The visualization of the predicted differences gives the time period in which two contours differ as a function of illocution type. For model fitting of the GAMMs, we used the R package mgcv (Wood, 2011, 2017); the package itsadug was used to plot the model results (van Rij, Wieling, Baayen, & van Rijn, 2017). The response variable was the \(f_0\) value (in Hz) at different time points (10 measurements per word). One model was fitted for polar, one for wh-questions. The models included illocution type as a parametric effect (fixed effect), along with a factor smooth for the interaction of illocution type over (normalized) time, \(s(\text{Normtime}, \text{by } = \text{ illocution type})\), using the thin-plate regression spline (‘tp’). We modelled separate smooths for subjects and items to account for the experimental structure. The model including the smooth term that captured the interaction of illocution type over time was subsequently compared to a simpler model without the smooth term, using the function CompareML(). This comparison tested whether the inclusion of this term significantly improved the fit of the model in terms of Maximum Likelihood (see Porretta, Tucker, & Järvič, 2016). All models were corrected for autocorrelation in the data using a correlation parameter, determined by the acf_resid() function from the package itsadug (van Rij et al., 2017). Model fits were checked using the gam.check() function and the number of base functions (\(k\)) was adjusted if necessary. Also, best-fitting models were re-run with the scaled \(t\) distribution (family = “scat”), closely following the suggestion in van Rij et al. (2019, p. 17), in order to account for tailed residuals.

Utterance (and word) durations (cf. H1b) were statistically analysed using linear mixed-effects regression models (lmer, Baayen, Davidson, & Bates, 2008). voice quality labels (cf. H1c) were analysed using logistc mixed effects regression models (glmer), coding glottalized voice as 1 and modal voice as 0 (breathy voice did not occur in the data). For glmers, we used the “bobyqa”-optimizer (Powell, 2009) in the glmerControl function in order to reduce convergence issues. Otherwise, the modelling procedure was the same for the continuous (duration, H1b) and categorical data (voice quality, H1c): Levels in all categorical variables were dummy coded, i.e., each individual level is compared against the reference level (intercept). Participants and items were entered as crossed random factors (Baayen et al., 2008). Random slopes were added and retained if they improved the fit of the model (Matuschek et al., 2017) – based on model comparisons with the anova() function that compares LogLikelihoods. P-values were obtained using the Satterthwaite approximation implemented in the R package lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017). They were adjusted based on the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) to counteract an increase in type-I error rate. We report adjusted p-values in the results part (\(p_{adj}\)) and assume a standard significance level of 0.05.

Research Question 2 (interplay between cues to RQs): To test whether mean \(f_0\) and utterance duration compensate for each other, we correlated the difference in mean \(f_0\) (\(f_0\)) with

\[\text{smooth functions include a pre-specified number of base functions of different shapes, e.g., linear and parabolic functions of different complexity (e.g., Wieling, 2018). Fixed effects are modelled in the same way as in linear mixed effect regression models. In addition, GAMMs also model non-linear effects over time. The visualization of the predicted differences gives the time period in which two contours differ as a function of illocution type. For model fitting of the GAMMs, we used the R package mgcv (Wood, 2011, 2017); the package itsadug was used to plot the model results (van Rij, Wieling, Baayen, & van Rijn, 2017). The response variable was the } f_0 \text{ value (in Hz) at different time points (10 measurements per word). One model was fitted for polar, one for wh-questions. The models included illocution type as a parametric effect (fixed effect), along with a factor smooth for the interaction of illocution type over (normalized) time, } s(\text{Normtime}, \text{by } = \text{ illocution type}), \text{ using the thin-plate regression spline (’tp’). We modelled separate smooths for subjects and items to account for the experimental structure. The model including the smooth term that captured the interaction of illocution type over time was subsequently compared to a simpler model without the smooth term, using the function CompareML(). This comparison tested whether the inclusion of this term significantly improved the fit of the model in terms of Maximum Likelihood (see Porretta, Tucker, & Järvič, 2016). All models were corrected for autocorrelation in the data using a correlation parameter, determined by the acf_resid() function from the package itsadug (van Rij et al., 2017). Model fits were checked using the gam.check() function and the number of base functions (} k \text{) was adjusted if necessary. Also, best-fitting models were re-run with the scaled } t \text{ distribution (family = “scat”), closely following the suggestion in van Rij et al. (2019, p. 17), in order to account for tailed residuals.} \]

\[\text{Utterance (and word) durations (cf. H1b) were statistically analysed using linear mixed-effects regression models (lmer, Baayen, Davidson, & Bates, 2008). voice quality labels (cf. H1c) were analysed using logistc mixed effects regression models (glmer), coding glottalized voice as 1 and modal voice as 0 (breathy voice did not occur in the data). For glmers, we used the “bobyqa”-optimizer (Powell, 2009) in the glmerControl function in order to reduce convergence issues. Otherwise, the modelling procedure was the same for the continuous (duration, H1b) and categorical data (voice quality, H1c): Levels in all categorical variables were dummy coded, i.e., each individual level is compared against the reference level (intercept). Participants and items were entered as crossed random factors (Baayen et al., 2008). Random slopes were added and retained if they improved the fit of the model (Matuschek et al., 2017) – based on model comparisons with the anova() function that compares LogLikelihoods. P-values were obtained using the Satterthwaite approximation implemented in the R package lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017). They were adjusted based on the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) to counteract an increase in type-I error rate. We report adjusted p-values in the results part (} p_{adj}\text{) and assume a standard significance level of 0.05.} \]

\[\text{Research Question 2 (interplay between cues to RQs): To test whether mean } f_0 \text{ and utterance duration compensate for each other, we correlated the difference in mean } f_0 \text{ (} \Delta f_0 \text{) with} \]

\[\text{For the disyllabic pronominal subject yějūrén, we assigned the respective label (”glottal” or ”breathy”) if one of the two syllables was non-modal; otherwise, the label ”modal” was assigned.} \]
the difference in intertemporal duration (Δduration) for each illocutionary pair (ISQ and corresponding RQ by the same speaker). A negative correlation was taken to indicate compensation between cues, i.e., the larger the difference across illocution types in one cue, the smaller the difference in the other. A positive correlation was taken to suggest that the cues are modified in tandem, i.e., the larger the difference in one cue, the larger the difference also for the other cue. Each illocutionary pair was further coded with respect to whether or not it was marked for voice quality. We assigned ‘yes’ if and only if one of the positions labelled for voice quality (first word, last syllable in object noun, plus sentence-final particle in polar questions) had modal voice in the ISQ version and glottalized voice in the corresponding RQ and none of the positions in the respective illocutionary pair showed the reversed contrast (i.e., glottalized in ISQ and modal in RQ). In all other cases, we assigned ‘no’. We subsequently checked the distribution of voice quality marking (‘Yes’) across illocutionary pairs. If voice quality marking compensates other cues (f0 and duration), we expect voice quality distinctions to occur for illocutionary pairs with weak marking of f0 and duration (small differences across illocution types).

3. Results

In this section, we will first provide the analyses in response to the first research question (prosodic differences between ISQs and RQs in polar and wh-questions), presenting the acoustic comparison for the f0 trajectory of the target interrogatives (H1a, Section 3.1.1), (word) durations (H1b, Section 3.1.2), and the use of voice quality (H1c, Section 3.1.3). We will then turn to the interplay between these cues to RQs in response to the second research question (Section 3.2).

3.1. Prosodic differences between ISQs and RQs in polar and wh-questions: F0, duration, and voice quality

3.1.1. Global f0 trajectory

Fig. 2 provides a visualization of the time-normalized f0 trajectory for the entire target interrogatives (ISQ vs RQ) to reveal global trends in f0, for polar (A) and wh-questions (B).

Note that Fig. 2 averages the f0 contours over each word, irrespective of lengths and tones of this word. However, since we are comparing sentence pairs (i.e., the same sentences by the same speakers in the two illocution types), we can nevertheless analyse and interpret these average contours. To statistically corroborate the differences in the f0 trajectory between ISQs and RQs, we used GAMMs. The factor smooths for the interaction of illocution type over (normalized) time were necessary for both polar questions ($\chi^2(2.00) = 201.2$, $p_{padj} < 0.0001$) and wh-questions ($\chi^2(2.00) = 52.6$, $p_{padj} < 0.0001$), indicating that the impact of illocution type differed over the course of the utterance.

For both polar and wh-questions, the final GAMM included illocution type as a parametric effect (fixed effect), along with a factor smooth for the interaction of illocution type over (normalized) time, s(Normtime, by = illocution_type) and a smooth for subjects and items (random slopes). The final model accounted for 69.6% of the deviance in polar questions and 65.4% in wh-questions. The final model was corrected for auto-correlation as well as re-run with the scat-linking function.\(^8\) Given that we can interpret the GAMM results more intuitively with visualizations, we present the visualized model output in Fig. 3. The summary table of the final model can be found in the supplementary analysis script. Fig. 3 shows the predicted difference in f0 (predicted f0 values in RQ condition minus ISQ condition). The left panel shows the predicted difference curve for polar questions; the right panel for wh-questions. The predicted difference curves show when in time ISQs and RQs differ significantly from each other (values below 0 indicate that RQs are lower, values above 0 the reverse; for the period(s) when the 95% confidence interval of the difference curve does not include the horizontal line at zero, the difference is significant, as indicated by the vertical red lines that highlight significant periods).

The f0 difference curves in Fig. 3 reveal that RQs are lower for most parts of the question (except for parts of the verb). The analyses further reveal a larger f0 range for RQs on the first word both for yōurēn ‘anyone’ in polar questions and shéi ‘who’ in wh-questions. The increased f0 range seems to be due to a lowering of the low tonal target (around Normtime 5 for polar and Normtime 7 for wh-questions, see Fig. 2), resulting in a dip of the contour for RQs. There is no difference across illocution type for almost half of the verb in polar questions (around Normtime 10–15) and for almost one third of the verb in wh-questions (around Normtime 11–13). From the verb onwards, the f0 trajectory between ISQs and RQs diverges more and more as the interrogative unfolds.

To summarize the findings regarding f0, Standard Chinese RQs show a lower f0 trajectory than ISQs, which holds for both polar and wh-questions. The f0 trajectory in RQs is characterized by a larger f0 range for the subject word (yōurēn ‘anyone’ in polar and shéi ‘who’ in wh-questions), due to a lowering of the low target in RQs. From the verb onwards, the f0 trajectory diverges, with RQs having lower f0 values than ISQs. Hence, along with the lexical function of f0 in Standard Chinese, speakers also use f0 to mark a difference in illocution type, i.e., the difference between ISQs and RQs. We now turn to prosodic cues other than f0, which are duration and voice quality.

3.1.2. Duration

We first tested whether the global utterance duration differed as a function of illocution type: Polar-RQs were 163 ms longer than polar-ISQs (1498 ms vs 1335 ms; $\beta = 0.165$; SE = 0.03, df = 20.50, $t = 5.42$, $p_{padj} < 0.0001$) and wh-RQs were 166 ms longer than wh-ISQs (1280 ms vs 1114 ms; $\beta = 0.172$; SE = 0.02, df = 175.2, $t = 7.75$, $p_{padj} < 0.0001$). Since illocution type and question type did not interact ($p_{adj} = 0.91$), we assume that durational differences between ISQs and RQs hold independently of question type. For exploratory purposes, we assessed whether the longer duration in RQs was carried by a specific part in the utterance (Lo & Kiss, 2020; or whether the lengthening applied globally). To this end, we tested for an interaction between illocution type and word in an omnibus model that combined polar and wh-questions.

---

\(^8\) The R syntax of the final GAMM was the following: f0.gamb6_acf_scat_polar=bam(f0 ~ illocution_type+s(Normtime, by=illocution_type, bs = ’p’, k = 15) +s(Normtime, vp, bs=f’s’, m = 1) + s(Normtime, item, bs=f’s’, m = 1), data=data, rho = rhoval, AR.start = data $start_event, method="fREML", discrete=T, family="scat". 
The variable word included the individual words in the target questions (yōurēn or shēi, verb, object noun). The interaction between illocution type and word was significant ($\chi^2 = 9.03, df = 2, p_{adj} < 0.05$). Hence, the durational differences between ISQs and RQs affected the individual parts of the utterance differently, see Fig. 4. In particular, in both question types, the relative difference across illocution type was strongest for the first word, with yōurēn ‘anyone’ being 25% longer in polar-RQs than in polar-ISQs, and shēi ‘who’ being 28% longer in wh-RQs than wh-ISQs. The proportional difference between illocution types in the other words for RQs was less than 12%: verb: 9% in both polar and wh-questions, object noun: 10% in polar and 12% in wh-questions, sentence-final particle me in polar question: 5%. All words except for the sentence-final particle ($p_{adj} = 0.26$) were longer in RQs than in ISQs (all other $p_{adj} < 0.05$). To sum up, RQs are generally longer than ISQs, with the largest difference between illocution types occurring for the pronominal subject yōurēn in polar questions and the wh-word shēi in wh-questions.

3.1.3. Voice quality

The majority of instances at the different measure points for voice quality showed modal voice (83% of the overall labels were modal voice, 17% were glottalized voice; breathy voice...
did not occur at all in our data). Fig. 5 shows the proportion of glottalized voice for polar (A, left) and wh-questions (B, right) for ISQs and RQs – at the different measurement points across the question (first word, final syllable in object noun, and sentence-final particle for polar questions); the respective other realizations are modal voice. Note that proportions relate to the total number of instances in each condition (bar), i.e., \( N = 106 \) polar-ISQs, \( N = 106 \) polar-RQs, \( N = 103 \) wh-ISQs, \( N = 103 \) wh-RQs. Position (i.e., first word, final syllable in object noun, and sentence-final particle in polar question) and illocution type did not interact (\( p_{\text{adj}} = 0.30 \)), as revealed in a combined model for polar and wh-questions, suggesting no evidence to assume that the distinction in voice quality between ISQs and RQs was different for the different positions in the utterance. There was no interaction between illocution type and question type (\( p_{\text{adj}} = 0.93 \)), corroborating that the difference between ISQs and RQs in terms of voice quality marking was comparable in both question types. For both polar and wh-questions, there are more instances of glottalized voice in RQs than in ISQs (polar questions: 21.4% vs 9.4%, \( \beta = 1.35 \), \( SE = 0.28 \), \( z = 4.9 \), \( p_{\text{adj}} < 0.0001 \); wh-questions: 25.7% vs 12.1%, \( \beta = 1.59 \), \( SE = 0.37 \), \( z = 4.3 \), \( p_{\text{adj}} < 0.0001 \)).

Taken together, Standard Chinese RQs showed an overall higher number of instances of glottalized voice than ISQs, independently of question type and position. From previous research, we know that other factors also influence the occurrence of glottalized voice beyond illocution type, e.g., mean \( f_0 \), lexical tone, and position (Chen & Gussenhoven, 2008; Kuang, 2017; Shih, 1997, 2000; Xu, 1999). In a first attempt to quantify the strength of the predictor illocution type for the occurrence of glottalized voice – in relation to these other factors – we used a random forest model. This model was trained to predict the occurrence of glottalized voice in the final syllable of the object noun (the syllable in which all four tones occurred, four times Tone 3, six times all other tones). Depending on question type, this syllable was in utterance-final position (in wh-questions) or in the penultimate position (in polar questions due to the sentence-final particle). The model revealed lexical tone to be the most important predictor for voice quality in the last syllable of the noun, followed in importance by speaker, illocution type (ISQ vs RQ), and position, see Appendix B for details. This model puts the strength of the factor illocution type into perspective. Specifically, ISQs and RQs are distinguished by voice quality marking in Standard Chinese, but the occurrence of glottalization is intertwined with lexical tone and also depends on the speaker.

3.2. Interplay between individual cues to RQs

The acoustic analyses in Section 3.1 showed that RQs are realized with an overall lower \( f_0 \) trajectory (mainly due to a lowering of the low tones in the first word and a compressed \( f_0 \) range towards the end of the utterance), longer durations (overall, but especially for the first word), and more instances of glottalized voice (both at the beginning and the end of the utterance). These acoustic comparisons between illocution type (ISQ vs RQ) identify individual cues to RQs, but they do not provide information about the interplay between these cues, i.e., whether these cues are modified together or whether one cue might substitute the other. To pursue Research Question 2, which addresses exactly this interplay between the cues, we analysed whether speakers used duration and mean \( f_0 \) in tandem for each illocutionary pair (ISQ and corresponding RQ, \( N = 200 \)). This was operationalized by calculating the difference in mean \( f_0 \) and in utterance duration between the RQ and ISQ production of each pair (\( \Delta \text{duration} \) and \( \Delta f_0 \), henceforth). Each illocutionary pair was further coded with respect to whether or not it was marked for voice quality. Recall that ‘yes’ was assigned if and only if one of the positions labelled for voice quality (first word, last syllable in object noun, plus sentence-final particle in polar questions) had modal voice in the ISQ version and glottalized voice in the corresponding RQ and none of the positions in the respective illocutionary pair showed the reversed contrast (i.e., glottalized in the ISQ and modal voice in the RQ version). Otherwise, this illocutionary pair was coded as ‘no’. Fig. 6 shows \( \Delta f_0 \) (in Hz, ISQ minus RQ) against \( \Delta \text{duration} \) (in seconds, RQ minus ISQ). Each dot represents one illocutionary pair. Illocutionary pairs that are marked by voice quality are plotted in orange triangles (voice quality marking: yes), those that are not have black dots. Regression lines are shown for illocutionary pairs with (orange solid line) and without voice quality marking (black dotted line); the standard error of the regression line is shown in grey shading.

Overall, there was a moderate positive correlation between \( \Delta f_0 \) and \( \Delta \text{duration} \) for polar questions (\( r = 0.50 \) [95%CI: 0.34; 0.64], \( t = 5.87, df = 101, p_{\text{adj}} < 0.0001 \)) and a weak to moderate positive correlation for wh-questions (\( r = 0.41 \) [95%CI: 0.23;
3.1.3). Importantly, on the lexical tone (cf. random forest analysis, in or additional cue, which depends on the speaker and, most significantly, on the lexical tone (cf. random forest analysis, in Section 3.1.3).

Our findings showed that Standard Chinese RQs, as compared to ISQs, are realized with lower \( f_0 \), longer duration and more instances of non-modal (glottalized) voice quality, both in polar and in wh-questions, supporting H1. Hence, the prosodic differences between ISQs and RQs are not confined to a particular question type. We have thus shown that essentially the same prosodic cues are used in the tone language Standard Chinese that have also been observed for intonation languages such as English and German, among others (see Section 1.1 above). Given the primary lexical function of \( f_0 \) in tone languages, which it does not have in intonation languages, the similarities between the two types of languages are particularly noteworthy. With regard to the interplay between cues to RQs, we find that in most illocutionary pairs, \( f_0 \) is modified in tandem with duration and voice quality (cf. Q2). In the remainder of this section, we will first discuss the implications arising from the findings on the prosodic differences in Standard Chinese ISQs vs RQs (Section 4.1). From a cross-linguistic perspective, we will also elaborate on the common use of \( f_0 \), duration and voice quality as prosodic cues to RQs in typologically different languages, and potential explanations (assertive force, focus, speaker attitude) that unite the present findings on RQs in Standard Chinese with the findings on RQs in other, non-tonal languages. In Section 4.2, we address the implications of the interplay between prosodic cues to RQs for the modelling of the relation between prosody and meaning.

4. General discussion

4.1. Cues to RQs in Standard Chinese and other (typologically different) languages

This section is organized according to cues, starting with \( f_0 \). RQs were realized with lower mean \( f_0 \) than ISQs in Standard Chinese polar and wh-questions, in line with previous findings that Cantonese and Standard Chinese wh-RQs have lower \( f_0 \).
associated with the sentence-final particle than corresponding wh-ISQs (Lo & Kiss, 2020; Lo et al., 2019a). Lower mean f0 is not restricted to tone languages but also occurs in intonation languages (Beyssade & Delais-Roussarie, 2022 for French; Sahkai et al., 2022 for Estonian; other studies on intonation languages have focused on the differences in pitch accents and boundary tones rather than on global features of f0 to arrive at a more detailed and language-specific analysis). The f0 distinction between ISQs and RQs in Standard Chinese also fits in with previous research on pitch modifications for the purpose of marking affect in Standard Chinese (Li et al., 2011), as well as on prosodic characterizations of interrogatives as opposed to declaratives (Lee, 2005; Liu & Xu, 2005; Liu et al., 2016; Yang, 2018; Yang, Gryllia, Pablos, & Cheng, 2019; Yuan, 2006). Given that RQs have functionally been considered assertion-like (cf. Han, 2002, p. 202) and that assertions are typically realized as declaratives in terms of syntactic structure, it is not surprising that RQs prosodically resemble declaratives in that they display lower f0 compared to information-seeking questions (for Standard Chinese: Lee, 2005; Liu & Xu, 2005; Liu et al., 2016; Yang, 2018; Yang et al., 2019; Yuan, 2006). Conversely, higher pitch, both globally and at specific positions in the utterance (here final) have been associated with inquisitive utterances (information-seeking questions) as compared to statements in a variety of languages (cf. Hirst & Di Cristo, 1998, pp. 24-26 for overview).

The f0 modulation we observe cannot solely be explained by the marking of assertive force, otherwise we would have observed a difference in register only (e.g., RQs uniformly lower than ISQs). Yet, the f0 trajectories of ISQs vs RQs paint a more detailed picture that goes beyond a register difference. There is an expansion of the f0 excursion for the pronominal subject 你 (anyone) in polar questions and the wh-word 什么 (who) in wh-questions in RQs as compared to ISQs. From the second word onwards (i.e., the verb), the f0 trajectories diverged, with RQs becoming increasingly lower than ISQs. Hence, it seems that speakers increased the f0 range to mark an interrogative as rhetorical in the beginning of the utterance, and after that reduced the f0 range in RQs as compared to ISQs (cf. Yuan, 2006, who shows the distinction in f0 between statements and polar questions to become larger towards the end of the interrogative). This observation of an increased f0 range would be compatible with prosodic focus marking on the first word in the RQs, followed by post-focal compression (Gårding et al., 1983; Jin, 1996; Xu, 1999; Xu & Xu, 2005; see Chen, 2005 for overview). Such an interpretation is in line with the fact that the first word was also lengthened most strongly (in both polar and wh-questions), further increasing their prominence. A post-hoc prominence rating task by a native speaker of Standard Chinese, who indicated whether the first word 你 (anyone) in polar questions and 什么 (who) in wh-questions) or another part of the question sounded most prominent to her, showed that RQs were often perceived as having the main sentence prominence on the first word (37% in polar and 56% in wh-questions); the main prominence for ISQs was perceived on the sentence-final object noun in almost all cases (100% in polar and 96% in wh-questions). Interestingly, Dehé and Braun (2020b) observe a similar shift in nuclear accent position for English polar questions: About 25% of the RQs in their data set were realized with the nuclear accent on the subject pronoun ‘anyone’ and no accent on the sentence-final noun (‘Does ANYONE eat lemons?’). Capsitals indicate the word that carries the nuclear pitch accent). This pattern never occurred for ISQs. For polar questions, a focus on anyone can turn the indefinite subject pronoun into a negative polarity item, which is only compatible with RQs and not with ISQs (e.g., Han, 2002). Likewise, a focused wh-word may strengthen the salience of the empty set interpretation intended by the RQs (“Nobody likes lemons”).

Our data-driven idea of focus being realized on the first word in RQs, but not in ISQs, however, challenges the semantic assumption that the wh-word is considered to have focus in neutral wh-questions (e.g., Deguchi & Kitagawa, 2002; Lambrecht & Michaelis, 1998; Yang et al., 2020). The theoretical question that arises is whether – based on our empirical data – we would still have to assume focus in wh-ISQs on the wh-word, and how to theoretically differentiate the two types of wh-questions (ISQ vs RQ) if both have focus on the wh-word, but perceptually, the prominence location differs. One possibility might be that in ISQs, we observe a misalignment between semantic focus and its prosodic manifestation, i.e., semantic focus on the wh-word, prosodic manifestation in sentence-final position (see Chen, 2006, for an example of durational marking of corrective focus not being self-contained on the focused element, but spilling over to the following syllables; cf. Rooth, 2008, on scope of focus). It might also be the case that the prosodic realization of focus in RQs is more salient than that in ISQs, which might have influenced the native speaker in the prominence rating task towards more ‘first word’ ratings in RQs as compared to ISQs. In that case, focus marking might be a matter of different degrees of emphasis between ISQs and RQs, with RQs being more strongly marked than ISQs – along the lines of Chen and Gussenhoven (2008) for different degrees of emphasis in focus marking. There are hence several possibilities that could explain our findings of the expansion of the f0 excursion for the first word in RQs as compared to ISQs, which would indeed be worthy of further research in the future.

From a cross-linguistic perspective, Standard Chinese resembles other, typologically different languages in that it also uses f0 to mark a question as rhetorical. Even though the implementation shows language-specific aspects (i.e., certain pitch accent types or edge tones are favoured in RQs in intonation languages, cf. Dehé et al., 2022), there seem to be cross-linguistic similarities with respect to the position of sentence accent (Dehé & Braun, 2020b).

**Duration.** RQs in Standard Chinese were produced with longer overall duration, in line with findings in a variety of typologically different languages (Beyssade & Delais-Roussarie, 2022 for French; Braun et al., 2019 for German; Dehé & Braun, 2020a for Icelandic; Dehé & Braun, 2020b for English; Dehé et al., 2018 for Icelandic; Lo et al., 2019b on Cantonese; Miura & Hara, 1995 for Japanese; Sahkai et al., 2022 for Estonian; Soraniello, 2018 for Italian). Longer duration hence seems to be a stable characteristic of RQs across languages, both occurring in lab-like settings and in RQs in spontaneous productions (Braun et al., 2020). This again ties in with durational differences in other speech acts, such as the distinction between statements and interrogatives in Standard Chinese, both for polar and wh-questions (X. Liu et al., 2016; Yang,
Also, faster speaking rate has been observed in declarative questions as compared to declaratives (Niebuhr et al., 2010, for German; van Heuven & van Zanten, 2005, for Manado Malay, Orkney English, and Dutch; and for exclamatives vs ISQs for German, Wochner, 2022). We can hence generalize that utterances with an assertive force (statements and rhetorical questions) are longer than genuine information-seeking questions, which lack assertive force. Given that similar prosodic differences — both for f0-related and durational cues — have been reported for the distinction between statements and questions in Standard Chinese (Lee, 2005; Liu & Xu, 2005; Liu et al., 2016; Yang, 2018; Yang et al., 2019; Yuan, 2006), a logical next step in future studies will be to compare Standard Chinese RQs to string-identical assertions (Wochner, 2022 on German ISQs, RQs, and assertions).

**Voice quality.** We find that Standard Chinese RQs are more often realized with glottalized voice than string-identical ISQs (see also Lo & Kiss, 2020, on creaky voice in sentence-final particles in wh-RQs). This was the case for both the first word, the final syllable of the noun and the sentence-final particle, although glottalization was generally more frequent towards the end of the utterance. Similarly, voice quality differentiates between ISQs and RQs in a number of languages (Braun et al., 2019; Dehé & Braun, 2020b; Dehé & Wochner, 2022; Sahkai et al., 2022). While the voice quality contrast is spelled out differently across languages, either as a contrast between modal and breathy (English, German, Icelandic) or between modal and glottal (Standard Chinese, Estonian), the generalization is that RQs are more often produced with non-modal voice quality than ISQs. In Standard Chinese, the presence of glottalization was influenced more by lexical tone identity (and hence f0) and speaker identity than by illocution type (cf. random forest), which suggests that glottalization is not a strictly independent cue. Breathy voice, a cue to RQs in other languages (Braun et al., 2019; Dehé & Braun, 2020b; Dehé & Wochner, 2022), might be more independent of f0. Future research needs to further study the relation between voice quality and f0 in the marking of illocution type in other languages.

Given that the possible answers to the RQs in our study were negative, e.g., nobody eats lemons, it might be the case that speakers additionally convey a negative attitude when asking an RQ. Conceivably, stronger attitudes may have led to a stronger marking of illocution type (cf. a post-hoc analysis of Neitsch, 2019, for the influence of attitude on the prosodic marking of RQs in German). For instance, lower f0, longer duration and glottalized voice have also been reported to mark disgust, a negative attitude, in Standard Chinese (Li et al., 2011; Liu & Pell, 2012; Yuan et al., 2002). As indicated earlier, a pre-test confirmed that our materials were on average ambiguous as to whether participants liked them or not. Clearly though, for individual items there might be differences in the strength of attitude, which may have interacted with the marking of illocution type.

4.2. **Interplay between cues to RQs in Standard Chinese**

In response to our second research question, we analysed the interplay between the different cues to illocution type. From other languages and linguistic phenomena it is known that prosodic cues may either have trading relations (with one cue compensating for the other, e.g., Kim, 2020; Niebuhr, D’Imperio, Gili Fivela, & Cangemi, 2011; cf. Schertz & Clare, 2019) or, conversely, prosodic cues may be used in tandem (Braun, 2006; Kim, 2020). For Standard Chinese, as a tone language, we considered the possibility that speakers use fewer f0 modifications and instead make more use of duration and voice quality. Our findings revealed that this is not the case and that f0 and duration are largely modified in tandem, as evidenced by a positive correlation between the differences in mean f0 and sentence duration across illocutionary pairs. Hence, the stronger the marking in terms of f0, the stronger the durational difference; if voice quality distinctions occurred, they did so in addition to other cues.

What does this joint modification of cues in the encoding of illocution type imply for the relationship between prosody and meaning? Recently, so-called ‘prosodic constructions’ have been suggested in the literature (Ward, 2019; Ward & Gallardo, 2017) in order to explain how prosodic cues combine to form meaningful configurations or constructions. Prosodic constructions are defined as “recurring temporal patterns of prosodic activity that express specific meanings and functions and which typically involve not only pitch contours but also energy, rate, timing and articulation properties” (Ward & Gallardo, 2017, 3f., but see Burdin, 2020; Huttenlauch, Feldhausen, & Braun, 2018, for arguments against prosodic constructions). The finding of joint modification of f0, duration, and voice quality is compatible with a prosodic construction approach. However, given that these cues are also jointly employed to encode focus (f0, duration), interrogativity (f0, duration), and emotions/attitudes (f0, duration, and voice quality), considering the observed pattern ‘prosodic constructions’ does not add explanatory value. Based on our findings, we hence argue that a bundle of cues is used to convey rhetorical illocution, just like for emphatic focus (Chen & Gussenhoven, 2008), but we refrain from confining the present combinations of cues to RQs alone. The weighting of cues to RQ interpretation needs to be further investigated in perception studies (see Kharaman, Xu, Eulitz, & Braun, 2019 on German; Miura & Hara, 1995 on Japanese). Such experiments assess how speakers weigh cues to identify and distinguish RQs from ISQs. Kharaman et al. (2019), for instance, orthogonally crossed nuclear pitch accent (late- vs early-peak accent), duration (short vs long), and voice quality (breathy vs modal) and asked German listeners to judge target interrogatives as either ISQs or RQs. While the interplay of all cues (late-peak nuclear accent, long duration, breathy voice) led to 97% of RQ responses, the presence of two cues still resulted in more than 75% of RQ identifications (80% for late-peak and breathy voice, and 75% for late-peak and long duration). We will pursue this issue for Standard Chinese in future research, with a particular focus on the interaction between lexical tone and intonation.

5. Conclusion

To conclude, our study reveals that speakers of Standard Chinese jointly employ lower f0, longer duration, and – less consistently more glottalized voice quality – to prosodically mark the illocutionary force of a question as rhetorical. Impor-
tantly, these prosodic differences are not confined to a particular question type, but equally apply to both polar and wh-questions. Our findings further reveal that f0 and duration are largely modified in tandem, while voice quality seems to be an additional cue. From a cross-linguistic perspective, we conclude that f0 and duration modifications are robust cues to rhetorical illocution, with non-modal voice quality being a more optional choice (at least in the languages tested). As shown above, there are three possible explanations according to which cross-linguistic differences can be unified: (A) The marking of assertive force (duration and f0), (B) the marking of focus (accent position, expanded f0 range, longer duration), and (C) the marking of speaker attitude (lower f0, non-modal voice quality). Future work needs to find ways to check whether and how these possible mechanisms can be disentangled.

Funding

The research was funded by the DFG as part of research unit ‘Questions at the Interfaces’ (FOR 2111, P6), Grant No BR 3428/4-1,2 and DE 876/3-1,2.

Author contributions

Bettina Braun and Nicole Dehé are Co-PIs on the funded project on rhetorical questions mentioned in the Section Funding. They designed the experiment and supervised testing. Katharina Zahner-Ritter annotated the data, prepared the data for analysis, conducted the acoustic and statistical analyses and drafted the manuscript. Yiya Chen advised on language-specific aspects in the acoustic analysis and interpretation and conceptualization. All authors jointly discussed the experimental results their implications, and are responsible for the final version of the manuscript.

CRediT authorship contribution statement

Katharina Zahner-Ritter: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing – original draft. Yiya Chen: Conceptualization, Writing – review & editing. Nicole Dehé: Conceptualization, Investigation, Methodology, Writing – review & editing, Funding acquisition. Bettina Braun: Conceptualization, Investigation, Methodology, Writing – review & editing, Funding acquisition.

Acknowledgements

We are grateful to Manluolan Xu for translation of the experimental stimuli, testing the participants, and help with annotation, to Meiqiao Li and Tianyi Zhao for help with annotation and for native advice, and also to Jasmin Rimpler for help with data preparation. We also thank Cesko Voeten for helpful practical advice on GAMM modelling, and Haiping Long, Tian Li, Jun Liu, Mengzhu Yan and the audiences at P&P 2019 in Düsseldorf and virtual Speech Prosody 2020 in Tokyo for discussion and feedback. Finally, we thank the reviewers and editor for their helpful comments.

Appendix A

List of all target questions (polar and wh-version) used in the experiment. The last tone in the sentence-final noun is additionally indicated in the final column (items are grouped by the last tone in the sentence-final noun, 6 × Tone 1, 6 × Tone 2, 4 × Tone 3, 6 × Tone 4). In total, we made six adjustments compared to the German materials in previous work (cf. Braun et al., 2019) for cultural reasons. For instance, the item angora (安哥拉 [ångərə]) was replaced by sable fur (貂 [diáo] 貂 [pí]) as this was judged by native informants to be better known in Chinese culture.

List of all target questions (polar and wh-version) used in the experiment.

<table>
<thead>
<tr>
<th>wh-questions</th>
<th>polar questions</th>
<th>Last tone</th>
</tr>
</thead>
<tbody>
<tr>
<td>谁(shéi)吃(chi)小(xiǎo)蝦(xiā)？</td>
<td>有(yǒu)人(rén)吃(chi)小(xiǎo)蝦(xiā)？ (ma/me)</td>
<td>Tone1</td>
</tr>
<tr>
<td>谁(shéi)喜(xi)欢(huān)肝(gān)？</td>
<td>有(yǒu)人(rén)喜欢(huān)肝(gān)？ (ma/me)</td>
<td>Tone1</td>
</tr>
<tr>
<td>谁(wh-who)吃(chi)鸡(jī)肝(gān)？</td>
<td>有(yǒu)人(rén)吃(chi)鸡(jī)肝(gān)？ (ma/me)</td>
<td>Tone1</td>
</tr>
<tr>
<td>谁(wh-who)吃(chi)鸡(jī)肝(gān)？</td>
<td>有(yǒu)人(rén)吃(chi)鸡(jī)肝(gān)？ (ma/me)</td>
<td>Tone1</td>
</tr>
<tr>
<td>谁(wh-who)吃(chi)鸡(jī)肝(gān)？</td>
<td>有(yǒu)人(rén)吃(chi)鸡(jī)肝(gān)？ (ma/me)</td>
<td>Tone1</td>
</tr>
</tbody>
</table>
### Appendix A. Random forest: Predicting voice quality in final syllable of the noun

Using the R-package randomForest (Liaw & Wiener, 2002), we fitted a random forest model to predict the voice quality label (modal vs glottal) for the final syllable in the sentence-final noun as a function of a number of variables, including illocution type, lexical tone in the last syllable of the sentence-final particle, speaker, and position (the second to last syllable of the interrogative in polar questions, and the last syllable of the interrogative in wh-questions). To train the random forest, we randomly selected 80% for training and 20% for test. The number of trees was set to 1000. Random forests extract the importance of the individual variables using the Gini-index (Liaw & Wiener, 2002), see Fig. B.1. Note that higher values indicate greater importance of a variable. Hence, our model revealed the lexical tone in the last syllable of the noun to be most important in predicting the occurrence of voice quality labels at the last syllable of the noun, followed in importance by speaker, illocution type, and position. The accuracy for the random forest models on 20% of unseen data (training set) was 82.1%.

<table>
<thead>
<tr>
<th>wh-questions</th>
<th>polar questions</th>
<th>Last tone</th>
</tr>
</thead>
<tbody>
<tr>
<td>谁(shéi)喜(xi)欢(huán)鱼(yú)肝(gān)油(yōu)?</td>
<td>有(yōu)人(rén)喜(xi)欢(huán)鱼(yú)肝(gān)油(yōu)?</td>
<td>Tone2</td>
</tr>
<tr>
<td>'Who likes cod-liver oil?'</td>
<td>'Does anyone like cod-liver oil?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)吃(chī)鱼(yú)吗(mɑ)?</td>
<td>有(yōu)人(rén)吃(chī)鱼(yú)吗(mɑ)?</td>
<td>Tone2</td>
</tr>
<tr>
<td>'Who eats durian?'</td>
<td>'Does anyone eat durian?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)吃(chī)鱼(yú)吗(mɑ)?</td>
<td>有(yōu)人(rén)吃(chī)鱼(yú)吗(mɑ)?</td>
<td>Tone2</td>
</tr>
<tr>
<td>'Who eats dates?'</td>
<td>'Does anyone eat dates?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)项(xiàng)鱼(yú)吗(mɑ)?</td>
<td>有(yōu)人(rén)项(xiàng)鱼(yú)吗(mɑ)?</td>
<td>Tone2</td>
</tr>
<tr>
<td>'Who breeds worms?'</td>
<td>'Does anyone breed worms?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)跳(tiào)鱼(yú)吗(mɑ)?</td>
<td>有(yōu)人(rén)跳(tiào)鱼(yú)吗(mɑ)?</td>
<td>Tone3</td>
</tr>
<tr>
<td>'Who dances breakdance?'</td>
<td>'Does anyone dance breakdance?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)想(xiǎng)去(qù)鱼(yú)馆(guǎn)?</td>
<td>有(yōu)人(rén)想(xiǎng)去(qù)鱼(yú)馆(guǎn)?</td>
<td>Tone3</td>
</tr>
<tr>
<td>'Who wants to go to the museum?'</td>
<td>'Does anyone want to go to the museum?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)知(zhī)道(dào)鱼(yú)经(jīng)xìng吗(mɑ)?</td>
<td>有(yōu)人(rén)知(zhī)道(dào)鱼(yú)经(jīng)xìng吗(mɑ)?</td>
<td>Tone3</td>
</tr>
<tr>
<td>'Who knows Houttuynia?'</td>
<td>'Does anyone know Houttuynia?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)喜(xi)欢(huán)数(shǔ)?</td>
<td>有(yōu)人(rén)喜(xi)欢(huán)数(shǔ)?</td>
<td>Tone4</td>
</tr>
<tr>
<td>'Who studies algebra?'</td>
<td>'Does anyone study algebra?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)吃(chī)喜(xǐ)欢(huán)果(guǒ)?</td>
<td>有(yōu)人(rén)吃(chī)喜(xǐ)欢(huán)果(guǒ)?</td>
<td>Tone4</td>
</tr>
<tr>
<td>'Who eats pasta Bolognese?'</td>
<td>'Does anyone eat pasta Bolognese?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)吃(chī)喜(xǐ)欢(huán)菜(cài)?</td>
<td>有(yōu)人(rén)吃(chī)喜(xǐ)欢(huán)菜(cài)?</td>
<td>Tone4</td>
</tr>
<tr>
<td>'Who eats innards?'</td>
<td>'Does anyone eat innards?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)喜(xi)欢(huán)蛋(dǎn)黄(huáng)?</td>
<td>有(yōu)人(rén)喜(xi)欢(huán)蛋(dǎn)黄(huáng)?</td>
<td>Tone4</td>
</tr>
<tr>
<td>'Who likes mayonnaise?'</td>
<td>'Does anyone like mayonnaise?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)吃(chī)蛋(dǎn)黄(huáng)?</td>
<td>有(yōu)人(rén)吃(chī)蛋(dǎn)黄(huáng)?</td>
<td>Tone4</td>
</tr>
<tr>
<td>'Who reads biographies?'</td>
<td>'Does anyone read biographies?'</td>
<td></td>
</tr>
<tr>
<td>谁(shéi)喜(xi)欢(huán)菜(cài)?</td>
<td>有(yōu)人(rén)喜(xi)欢(huán)菜(cài)?</td>
<td>Tone4</td>
</tr>
<tr>
<td>'Who likes celery?'</td>
<td>'Does anyone like celery?'</td>
<td></td>
</tr>
</tbody>
</table>
Fig. B.1. Results of the random forest for the voice quality label at the sentence-final noun. 1. The Mean Decrease Gini plot (right panel) reveals the importance of the predictors (last tone in sentence-final noun > speaker > illocution type > position); the Mean Decrease Accuracy plot shows how much of the accuracy the model loses when excluding a variable (the higher the loss, the more important the variable).

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wocn.2022.101190.

References


