9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

The role of prosody for the interpretation of rhetorical questions in German

Jana Neitsch, Bettina Braun, Nicole Dehé

University of Konstanz, Germany
{jana.neitsch, bettina.braun, nicole.dehe}@uni-konstanz.de

Abstract

Questions can be marked as rhetorical by their prosodic realisation. In two eye-tracking experiments, we tested whether *wh*-questions can be interpreted as rhetorical (RQ) or information-seeking (ISQ) based on prosody. We manipulated nuclear pitch accent type (rise-fall with a late-peak L*+H vs. falling with an early-peak H+!H*) and voice quality (breathy vs. modal) and investigated the contribution of the modal particle *denn*. Participants had to decide whether they heard an RQ or ISQ by clicking on one of two labels. Experiment 1 presented listeners with *wh*-questions containing the modal particle *denn.* Experiment 2 replicated Experiment 1 without the particle. Results showed that late-peak accent and breathy voice quality led to a rhetorical interpretation, while early-peak accent with modal voice quality was interpreted as information-seeking. The presence of the particle slightly strengthened these interpretations. Listeners decided faster when presented with late-peak/breathy and early-peak/modal compared to the other conditions. Fixation data showed different sensitivity to the prosodic cues depending on the presence of *denn*. In sum, listeners can use the prosodic realisation of *wh*-questions to interpret them as rhetorical or not, i.e. contextual linguistic information and other means (e.g., syntactic or lexical) are not strictly necessary.

Index Terms: rhetorical questions, information-seeking questions, nuclear pitch accent, voice quality, modal particle, perception, *wh*-questions, German

1. Introduction

I*nformation-seeking questions* (ISQs) have been described as eliciting information from the addressee [1, 2], while *rhetorical questions* (RQs) are usually assumed to imply an answer that is already known (or at least inferable) to all interlocutors [3, 4, 5]. RQs, in contrast to ISQs, thus characteristically exhibit a mismatch between their interrogative form and their function [6, 7, 8].

A rhetorical reading can be signalled lexically, e.g., by strong *Negative Polarity Items* (e.g., Who *lifted a finger* to help her?) [8]. In German, modal particles such as *schon* and *auch* are explicitly associated with an RQ interpretation [2], while *denn* can occur in both illocution types (RQs and ISQs) [9], indicating that it might not bias either one of the two possible readings. However, the potential pragmatic influence of *denn* on question interpretation has not yet been empirically investigated.

Although RQs have been the subject of semantic and pragmatic investigations for decades, knowledge about their prosody is still scarce. In a first systematic analysis of their prosodic characteristics, we compared the realisations of string-identical pairs of polar and string-identical pairs of German *wh*-questions, where one member of the pair was produced in a rhetorical context, the other in an information-seeking context (see [10] for relevance of these cues). Our results for *wh*-questions showed that L*+H (late-peak) nuclear pitch accents occurred most often in rhetorical contexts, but rarely in information-seeking contexts. On the other hand, ISQs were mostly produced with L+H* and H*. The H+L* early-peak accent occurred predominantly in ISQs and hardly in RQs. Phonetically, RQs were realised with longer durations of the *wh*-word and the sentence-final object noun. Furthermore, a breathier voice quality in RQs than in ISQs seemed to play a crucial role in the realisation of RQs [10, 11].

In this paper, we investigate a) whether prosody is sufficient for listeners to identify *wh*-questions as rhetorical or not, when they are presented out of linguistic context and without lexical markers, b) the effects of nuclear pitch accent type and voice quality on the identification of a *wh*-question as rhetorical or not, c) whether the particle *denn* affects the use of these prosodic cues, and d) the time course of interpretation. For comparisons, we used a late-peak (L*+H) and an early-peak (H+!H*) pitch accent. According to [10], L*+H was the most frequent accent type in productions of RQs; H+!H* was chosen despite not being the most frequent accent type in ISQs because it rarely occurred in RQs and is clearly distinct from L*+H [12, 13])[1]. Voice quality was manipulated on the object noun in sentence-final position to make the cue available at the same time as accent type. In Experiment 1, all stimuli contained the particle *denn* [15]. For Experiment 2, the particle was removed. This allowed us to test the contribution of the particle to question interpretation while keeping the prosodic form of the questions the same.

Based on [10], we predict that both a late-peak nuclear accent (L*+H) and a breathy voice quality increase RQ interpretations of *wh*-questions in German. Following [9], we predict that the presence of the particle has little impact on the interpretation of *wh*-questions as RQ or ISQ.

2. Experiments

Two eye-tracking studies with a forced-choice identification task were carried out. Pitch accent type and voice quality were manipulated within-subjects, presence/absence of the particle between-subjects. Participants' mouse clicks and click latencies were monitored, participants' fixations were tracked.

2.1. Methods

2.1.1. Materials

For Experiment 1, we created 32 *wh*-questions containing the German modal particle *denn* (e.g., *Wer mag denn Vanille*, 'Who likes PRT vanilla'). Each question started with the *wh*-

[1] Note that following [14], we do not assume two distinct phonological categories for the two types of early-peaks H+!H* and H+L*.

10.21437/SpeechProsody.2018-39

word *wer* ('who') followed by a finite verb, the modal particle *denn* and a sentence-final object noun (e.g., *Vanille* 'vanilla'). All object nouns were mostly sonorant, consisted of three syllables and carried lexical stress on the penultimate syllable. For each object noun, we selected a corresponding colour picture (500x500 pixels).

The questions were audio-recorded by a trained female native speaker in a sound attenuated booth. She first produced each *wh*-question with a late-peak accent (L*+H) and an early-peak accent (H+!H*) in modal voice quality on the object noun (note that the early-peak accent questions had an additional prenuclear H* accent on the *wh*-word to make the contour more natural). Based on [10], the final boundary tone was low (L-%). After each modal version, the speaker recorded the question with the same contour but breathy voice quality on the object noun. This procedure helped to achieve acoustic similarity of global intonation contours between the two realisations with the same accent type in different voice qualities. In some recordings, f0-maxima and minima within the object were slightly lower in breathy voice than in modal voice quality. Therefore, f0-maxima and minima were resynthesized (PSOLA, [16]) to achieve an average scaling across voice quality versions. As shown in the production study by [10], *wh*-RQs are characterised by significantly longer durations of the overall utterance, the first constituent and the object noun as compared to their information-seeking counterparts. This might be a cue to a rhetorical interpretation as well. In our recordings, breathy voice versions were also longer than modal voice versions. To avoid potential influences of durational differences, durations for each word were manipulated such that they had the average duration of the modal and breathy recording of that question pair.

In total, we used 32 *wh*-questions in four prosodic realisations, i.e. 128 experimental items (32 interrogatives x 4 conditions, see Figure 1 for example contours).

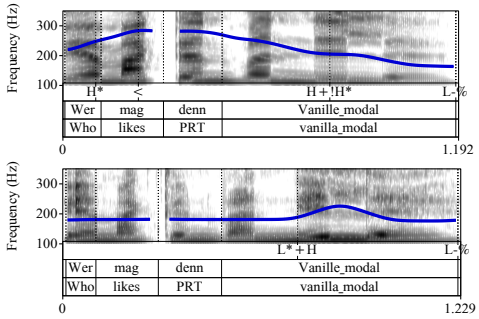

Figure 1: *Example contours for the two pitch-accent conditions in Experiment 1 (top: early-peak accent (H+!H*) on the noun, bottom: late-peak accent (L*+H) on the noun).*

To corroborate the voice quality manipulation acoustically, we extracted spectral tilt, amplitude differences between H1*-A3*, following [17] (see [18] for difficulties with H1-H2). H1*-A3* was measured in the middle of the vowel of the *wh*-word, and in the stressed vowel of the object noun. Higher value indicates a breathier voice quality. Results showed no differences in the *wh*-word (p>0.6), but significantly higher values in the object noun of the breathy versions than of the modal ones (32.48dB vs. 28.58dB, p<0.0002).

In Experiment 2, we used the same stimuli as in Experiment 1, but the particle *denn* was cut out without affecting the naturalness of the experimental items.

2.1.2. Participants

Twenty-four native speakers of German participated in each experiment (Experiment 1: ∅=23.7 years, SD=3.2 years, 19 female, Experiment 2: ∅=22.8, SD=2.9, 17 female). They received a small payment and were tested individually.

2.1.3. Procedure

The 128 experimental items were divided into four lists of 32 items each (8 items x 4 conditions) following a Latin Square design (i.e. each participant listened to each experimental condition, but never for the same item). The experimental lists were pseudo-randomised to ensure that no more than two items from the same experimental condition immediately followed one another. Four practice trials were put at the start of each list. Each list was split into two blocks that contained 16 items (four per condition). A second version was created for each list by switching the two blocks in order to counterbalance potential training effects. The experimental lists were the same in both experiments. Participants were randomly assigned to one of the eight experimental lists.

Both experiments followed the same procedure. During the experimental session, participants were seated comfortably in front of an LCD screen. We used the desktop mounted EyeLink 1000 Plus system with head support. Participants' dominant eye were calibrated (pupil and corneal reflection) and validated prior to the experiment. Participants' fixations of the dominant eye were tracked and recorded during the experimental session with a sampling rate of 250Hz. An automatic drift correction was conducted after every fifth trial. Each trial started with a fixation cross that appeared for 300ms in the centre of the screen. Then, the picture (corresponding to the respective object noun) was presented for 2500ms on white background (this helped to situate the question, cf. [19, 20, 21]). Following the picture, the two labels *wirkliche Frage* ('real question', corresponding to ISQ) and *rhetorische Frage* ('rhetorical question', corresponding to RQ) were displayed side by side centred on the screen and each was framed by a rectangular box. The position of the labels (left vs. right) was counterbalanced such that a label never occurred in the same position for more than three trials in a row. The presentation of the auditory question started 1000ms after the appearance of the labels. Target sentences were presented over headphones. Participants were asked to indicate whether they had heard an RQ or an ISQ by clicking as quickly as possible on the corresponding label. No feedback was provided. Each experimental session took about 20 minutes.

Participants' mouse clicks and fixations were coded as pertaining to a particular label if they were directed within the frame of one of the two labels. Click latencies were measured relative to the offset of the acoustic stimuli.

2.2. Results

Per experiment, a total of 768 mouse clicks (24 participants x 32 items) were analysed. Results showed most clicks to the RQ label when *wh*-questions were produced with a late-peak accent (L*+H) in breathy voice quality (Experiment 1: 93%; Experiment 2: 73%, cf. Figure 2). In both experiments, the percentage of clicks to the RQ label dropped for questions with the same accent type in modal voice quality. Stimuli with an early-peak accent in modal voice were mostly interpreted as ISQs, i.e. RQ interpretations were lowest in this condition (Experiment 1: 7%; Experiment 2: 13%), whereas breathiness

in the same accent type category resulted in increased RQ interpretations. Henceforth, we will use the term *prototypical contours* to refer to the conditions that resulted in the most distinct interpretations (late-peak in breathy voice for RQ, and early-peak in modal voice for ISQ).

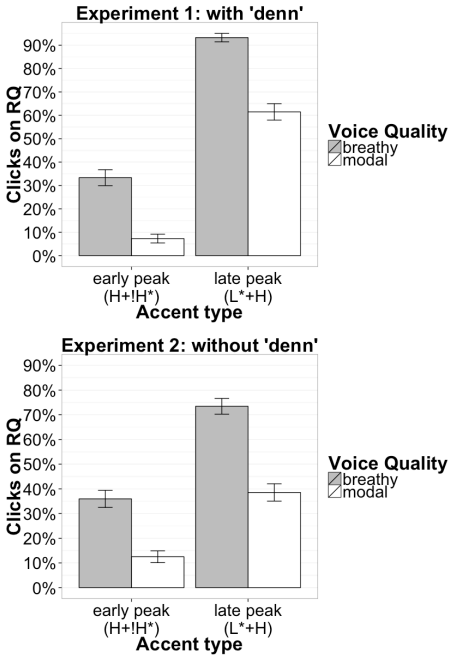

Figure 2: *Clicks on the RQ label by accent type (early-peak vs. late-peak) and voice quality. Whiskers indicate SE.*

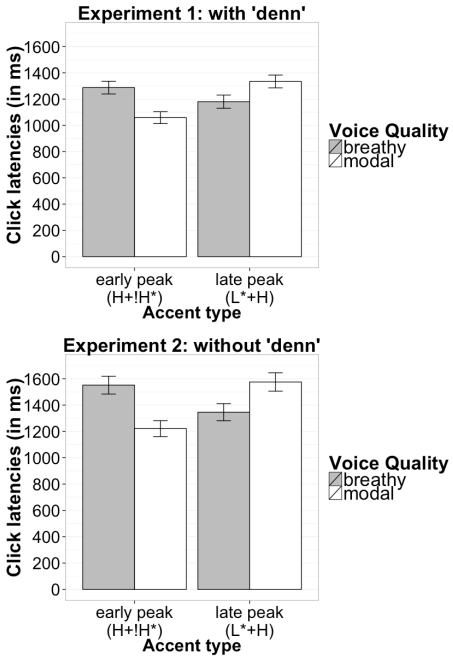

Figure 3: *Click latencies (in ms) from object noun offsets. Whiskers indicate SE.*

Clicks (coded as click on RQ) were statistically analysed by calculating a mixed-effects logistic regression model in RStudio (version 0.99.902, R version 3.2.2 [22]) with *accent type* (early-peak vs. late-peak) and *voice quality* (modal vs. breathy) as fixed factors and *subjects* and *items* as crossed random factors, allowing for random adjustments of intercepts [23]. *P*-values were calculated using the Satterthwaite approximation in the R-package lmerTest [24]. In what follows, values in square brackets indicate the 95% confidence interval of the estimate. Results showed a significant effect of *pitch accent type* (Experiment 1: β=4.90 [4.16; 5.77], *SE*=0.41, $p<0.0001$; Experiment 2: β=1.81 [1.44; 2.19], *SE*=0.19, $p<0.0001$) and a significant effect of *voice quality* (Experiment 1: β=3.35 [2.68; 4.12], *SE*=0.37, $p<0.0001$; Experiment 2: β=1.68 [1.32; 2.07], *SE*=0.19, $p<0.0001$). There was no interaction between *accent type* and *voice quality* (*p*-values in both experiments >0.6). A three-way interaction between *particle*, *voice quality* and *accent type* revealed that decisions were clearer in Experiment 1 than in Experiment 2 ($p<0.0003$).

For the analysis of *click latencies,* measurements greater than 2.5SD above the grand mean were excluded (Experiment 1: N = 46, Experiment 2: N=44). Click latencies were lowest for the prototypical contours. Linear mixed-effects regression models of *click latencies* revealed a significant interaction between *accent type* and *voice quality* in each experiments (both *p*-values <0.0001; cf. Figure 3). Click latencies were generally longer in Experiment 2 than Experiment 1 ($p<0.0003$) but there was no three-way interaction ($p=0.9$).

Figure 4 shows the evolution of fixations to the RQ label. Note that it takes about 150ms to plan a saccade (e.g., [25]). The fixation proportions did not differ during the processing of the *wh*-word or the verb in either experiment, nor for the particle in Experiment 1.

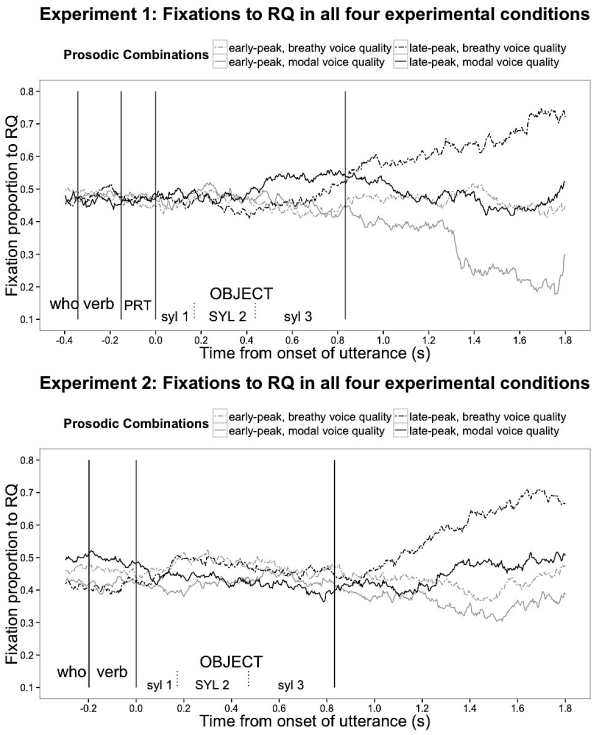

Figure 4: *Evolution of fixation proportions to RQ. Straight vertical lines indicate acoustic landmarks.*

In Experiment 1, fixation proportions to the RQ label began to differ in the object noun region. Starting from around 0.8s after the onset of the noun, fixations to the RQ label were higher in the late-peak conditions than in the early-peak

conditions, irrespective of voice quality (dashed and solid black lines in Figure 4, upper plot). To statistically corroborate this observation, fixations were analysed in 0.1s time windows from object noun onset (cf. [26]) until 1.8s after object noun onset. Following [27], *empirical logits* (elogs) were calculated by dividing fixations to the RQ label by fixations directed elsewhere. They were analysed in the same way as click latencies. Results showed a significant effect of *accent type,* starting at 0.8s after noun onset (β=0.66 [0.15;1.17], *SE*=0.26, $p<0.02$). An additional effect of *voice quality* started at 1.1s after noun onset (β=0.54 [0.07; 1.02], *SE*=0.24, $p<0.03$), i.e. after the offset of the noun. None of the analysed time windows revealed an interaction between the two variables (all *p*-values >0.5). Fixations for the prototypical contours started to differ around 0.8s after noun onset (β=0.71 [0.01; 1.41], *SE*=0.36, $p<0.05$). This fixation pattern was driven by the processing of the stressed syllable of the object noun.

For Experiment 2 without *denn,* all effects occurred after the offset of the object noun. Ranging from 1.2-1.7s after noun onset, there were significantly more fixations to the RQ label in items with breathy voice (β=0.27 [0.04; 0.50], *SE*=0.12, $p<0.05$; cf. Figure 4, lower plot). An additional effect of *accent type* started at 1.3s after noun onset (β=0.40 [0.13; 0.68], *SE*=0.13, $p<0.007$). There was no interaction between *accent type* and *voice quality* in the analysed time windows (all *p*-values >0.2). Fixation proportions for the prototypical contours started to differ significantly around 1.1s after noun onset (β=-0.36. [0.45; 1.87], *SE*=0.16, $p<0.03$), i.e. after noun offset.

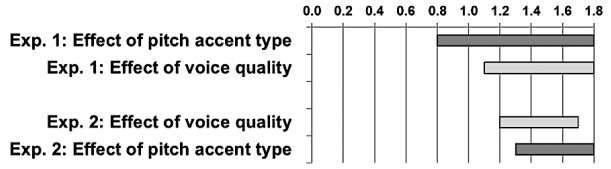

Figure 5: *Summary of effects found for fixations relative to the onset of the object noun (in s).*

To corroborate the differences across experiments, we calculated whether the above effects interacted with experiment. Only at 0.8-0.9s after the object noun onset, there was an interaction between *particle* and *accent type* ($p<0.02$, cf. Figure 5).

3. Discussion

The click data indicate that German *wh*-questions with a *nuclear late-peak accent* (L*+H) on the object noun that are produced with a *breathy voice quality* are reliably interpreted as conveying a rhetorical illocution, while a nuclear *early-peak accent* (H+!H*) in *modal voice quality* evokes predominantly information-seeking interpretations. In addition, participants decide faster when they are presented with these prototypical contours compared to one of the other combinations. When *breathy voice* occurs in *early-peak accents* and *modal voice* in *late-peak accents*, participants not only take longer to click, but they are also less confident in their decision. Removing the particle *denn* from the recordings leads to a similar overall pattern, but with less distinct choices and longer click latencies.

Concerning the online processing of illocutionary force, fixation patterns of Experiment 1 show that listeners do not differ in their fixations to the RQ label early in the utterance, suggesting that the particle *denn* and the prenuclear H* do not have an impact. Fixations to either label immediately increase once the stressed syllable has been processed, i.e. when nuclear pitch accent and voice quality are available. In contrast, Experiment 2 reveals effects only after the object noun offset. Here, the effect of voice quality sets in before the effect of accent type. The order in which accent type and voice quality are used by listeners in Experiment 2 is thus the reverse of what we find in Experiment 1.

Fixation data hence indicate differences in the time course of interpretation of the nuclear pitch accent and voice quality, depending on whether *denn* is present or not. One explanation for this timing difference is that the intonation contours were not ideal for *wh*-questions without the particle, which may have confused listeners. After all, the contours were modelled on production data containing a particle. This might explain the slightly earlier effect of voice quality compared to the later effect of accent type in Experiment 2 (see [28] for adaptive perception theories). It should be noted, however, that the identification of illocution type was still rather high in *wh*-questions without the particle (RQs: >70%, ISQs: >90%). Listeners were thus still able to interpret the prototypical contours for *wh*-questions containing a particle when confronted with *wh*-questions with a different syntactical structure (i.e. no particle). Admittedly though, more effort was necessary for identification. In a future production experiment, we will therefore investigate how *wh*-questions (RQs and ISQs) without the particle *denn* are realised.

Note that click latencies were comparatively long in both experiments, suggesting that the task was rather difficult. In future experiments, we will compare decision times for RQ and ISQ interpretations for less ambiguous linguistic structures to investigate whether RQ and ISQ judgments are difficult per se or whether the difficulty arose from the interpretation of prosody.

4. Summary and Conclusion

Primarily, the current data show that pitch accent type and voice quality facilitate listeners' disambiguation of string-identical *wh*-questions in German. Linguistic context and specific lexical markers such as polarity items or specialised particles are not necessary. Specifically, *wh*-questions with a late-peak nuclear accent in breathy voice are reliably identified as RQs, while *wh*-questions with an early-peak nuclear accent in modal voice quality are reliably identified as ISQs. The data without particle show less distinct interpretations and longer click latencies. The online eye-tracking data furthermore revealed that participants primarily relied on pitch accent type when the particle was present and on voice quality information when the particle was removed. This suggests that the relevance of these cues differs over time but further investigation is needed in this regard. For instance, stimuli with breathy voice quality on the *wh*-element will be analysed in a further study in order to provide a detailed interpretation of the fixations concerning breathiness as an early prosodic cue.

5. Acknowledgements

This research was supported by a project grant from the German Research Council (DFG), awarded to the last two authors. We thank Marianne Kusterer and Katharina Zahner for support with data analysis and Maria Biezma for discussion on the pragmatics of rhetorical questions.

6. References

[1] J. Groenendijk and M. Stokhof, *Studies on the semantics of questions and the pragmatics of answers.* Amsterdam: Universiteit van Amsterdam PhD Thesis, 1984.

[2] J. Meibauer, *Rhetorische Fragen* (Linguistische Arbeiten 167). Berlin: De Gruyter, 1986.

[3] I. Caponigro and J. Sprouse, "Rhetorical Questions as Questions," in E. Puig-Waldmüller (ed.), *Proceedings of Sinn und Bedeutung*, vol. 11, pp. 121–133, 2007.

[4] M. Biezma and K. Rawlins, *Rhetorical Questions.* Manuscript. University of Konstanz, John Hopkins University, 2016.

[5] R. Stalnaker, "Common ground," *Linguistics and Philosophy*, vol. 25, no. 5-6, pp. 701-721, 2002.

[6] J. M. Sadock, Queclaratives. In Papers from the Seventh Regional Meeting of the Chicago Linguistic Society, 223–331. Chicago: Chicago Linguistics Society, 1971.

[7] J. M. Sadock, Toward a linguistic theory of speech acts. New York: Academic Press, 1974.

[8] C. Han, "Interpreting interrogatives as rhetorical questions," *Lingua*, vol. 112, no. 3, pp. 201-229, 2002.

[9] M. Thurmair, "Zum Gebrauch der Modalpartikel 'denn' in Fragesätzen: Eine korpusbasierte Untersuchung," in E. Klein, F. Pouradier Duteil & K. H. Wagner (ed), *Betriebslinguistik und Linguistikbetrieb;* Linguistische Arbeiten 260, Tübingen: Niemeyer, pp. 377–388, 1991.

[10] D. Wochner, J. Schlegel, N. Dehé and B. Braun, "The prosodic marking of rhetorical questions in German," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, pp. 987–991, 2015.

[11] J. Neitsch, D. Wochner, K. Zahner. N. Dehé. B. Braun, "Who likes liver? How German speakers use prosody to mark questions as rhetorical," Phonetics and Phonology in Europe, June 12 -14 Cologne, Germany (accepted), 2017.

[12] K. J. Kohler, "Categorical Pitch Perception," *Proceedings of the XIth International Congress of Phonetic Sciences*, pp. 331–333, 1987.

[13] K. J. Kohler, "Categorical Speech Perception Revisited," *Proceedings of From Sound to Sense: 50+ years of discoveries in speech communication, MIT, Cambridge*, pp. 157–162, 2004.

[14] T. Rathcke, and J. Harrington, "Is there a distinction between H+!H* and H+L* in standard German? Evidence from acoustic and auditory analysis," *Proceedings of the 3rd international conference of speech prosody, May 2-5, Dresden, Germany,* pp. 783-786, 2006.

[15] M. Kusterer, *Prosodic cues to question interpretation: The influence of pitch accent and voice quality on the interpretation of rhetorical questions.* Master thesis, University of Konstanz, 2016.

[16] P. Boersma and D. Weenink, *Praat: doing phonetics by computer,* http://www.praat.org, accessed in March 2016.

[17] C. Mooshammer, "Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German," *Journal of the Acoustic Society of America,* vol. 127, no. 2, pp. 1047-1058, 2010.

[18] A. Simpson, "The first and second harmonics should not be used to measure breathiness in male and female voices," *Journal of Phonetics* 40(3), pp. 477-490, 2012.

[19] C. Ilie, "Question-response argumentation in talk shows," *Journal of Pragmatics,* vol. 31, no. 8, pp. 975–999, 1999.

[20] I. Koshik, "Wh-questions used as challenges," *Discourse Studies,* vol. 5, no. 1, pp. 51–77, 2003.

[21] D. Schaffer. Can rhetorical questions function as retorts? *Journal of Pragmatics,* vol. 37, no. 4, pp. 433–460, 2005.

[22] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016.

[23] R. H. Baayen, *Analyzing linguistic data*: *A practical introduction to statistics using R*, 1st edn. Cambridge: Cambride University Press, 2008.

[24] A. Kuznetsova, P. B. Brockhoff and R. H. B. Christensen, *lmerTest: Tests in Linear Mixed Effects Models*: *R package version 2.0-32,* https://CRAN.R-project.org/package=lmerTest, 2016.

[25] M. K. Tanenhaus, "Spoken language comprehension: insights from eye movements," in M. G. Gaskell (ed.), *The Oxford Handbook of Psycholinguistics,* pp. 309-326, 2007.

[26] P. D. Allopenna, J. S. Magnuson, and M. K. Tanenhaus, "Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models," *Journal of Memory and Language*, vol. 38, no. 4, pp. 419-439, 1998.

[27] J. D. Barr, T. M. Gann, and R. S. Pierce, "Anticipatory baseline effects and information integration in visual world studies," *Acta psychological,* vol. 137, no. 2, pp. 201-207, 2011.

[28] L. C. Nygaard, M. S. Sommers, and D. B. Pisoni, "Speech perception as a talker-contingent process," *Psychological Science,* vol. 5, no. 1, pp. 42–46. 1994.