The prosodic realization of rhetorical and information-seeking questions in German spontaneous speech

Bettina Braun, Marieke Einfeldt, Gloria Esposito, Nicole Dehé

Department of Linguistics, University of Konstanz, Germany
Bettina.Braun, Marieke.Einfeldt, Gloria.Esposito, Nicole.Dehe@uni-konstanz.de

Abstract

Previous results from laboratory experiments show that German speakers use prosody to distinguish between information-seeking questions (ISQs) and rhetorical questions (RQs). In the current paper we investigate whether pitch accents and edge tones (i.e., those that were typical for RQs and ISQs in the experimental data) are also used in spontaneous speech.

As compared to laboratory data, spontaneous speech data are syntactically and lexically more varied. However, notwithstanding more variation in their prosodic realization, RQs and ISQs in spontaneous speech essentially exhibit the same prosodic characteristics as RQs and ISQs in lab speech. Specifically, RQs were most often realized with an L*+H nuclear accent in both polar and *wh*-questions. Edge tones differed across question types. For polar ISQs, the most frequent edge tone was a high-rise H-^H%, while polar RQs were mostly realized with a high plateau (H-%) or a low-rise (L-H%). *Wh*-ISQs equally often ended in a low edge tone (L-%) and H-^H%, while *wh*-RQs most frequently terminated in L-%. RQs were furthermore produced with a slower speaking rate in both settings.

Given the similarities between the results for spontaneous vs. lab speech, the use of experimental data to investigate the prosodic realization of different illocution types is validated.

Index Terms: question, prosody, rhetorical question, information-seeking question, spontaneous speech, German

1. Introduction

Interrogative constructions are forms by which questions are linguistically expressed. The question in (1) can be interpreted as a genuine question, also known as information-seeking question (ISQ). Its function is the request of information and consequently the exigency of an answer from the addressee.

(1)   Wo wohnst du?

        'Where do you live?'

Kohler [1] investigated the intonation of polar and *wh*-ISQs in a corpus of German spontaneous speech from a pragmatic perspective. The underlying research question was concerned with the pragmatic and attitudinal meanings associated with falling or rising intonation patterns in polar and *wh*-questions. In total, 121 polar ISQs and 172 *wh*-ISQs were analyzed. The results showed that polar questions were mostly high-rising (39%), which corresponds to H-^H% in the German Tone and Intonation Break Indices system GToBI [2] or low-rising (L-H%; 30%) while *wh*-questions were mostly falling (L-%; 57%). However, rising contours were also found for *wh*-questions, which shows that rising and falling contours are present in both question types. The rising *wh*-questions were high-rising (10%) or low-rising (24%), cf. [1], p. 3.

Rhetorical questions (RQs) are formally interrogatives as well, but they differ from ISQs in that they have the feel of an assertion and are not necessarily expected to be answered [3]. The question in (2), for example, does not have to be answered since the answer is clear to everyone. It has the illocutionary force of an assertion stating that nobody wants to get stuck in traffic.

(2)   Wer steht schon gerne im Stau?

        'Who wants to get stuck in traffic?'

The fact that RQs are formally interrogatives but do not function as ISQs, makes them particularly interesting from a semantic and pragmatic perspective. It has been discussed how ISQs and RQs can be distinguished [4, 5] and how the interrogative form of RQs matches their function [3, 6]. Apart from the context and other lexical cues that help to interpret a question as information-seeking or rhetorical, it is commonly assumed that intonational cues are involved, too [3]. The influence of prosodic cues has been partly investigated, mostly for English, in work based on introspection or on corpus studies [7, 8]. For German, a pilot study by Wochner et al. [9] explored the differences in the prosodic realization of string-identical ISQs and RQs that were disambiguated by context in an interaction study.

Braun et al. [10] investigated prosodic differences between ISQs and RQs in a German production experiment. They examined final edge tones, nuclear pitch accents, frequent nuclear tunes and prenuclear accents. Further, the phonetic features initial pitch, constituent durations, and voice quality were investigated. To this end, 22 *wh*- and 22 polar interrogatives compatible with both a rhetorical and an information-seeking reading were constructed. Each question was placed in one of two different contexts triggering either an ISQ or an RQ. The verb and the nominal object of the target interrogatives were always lexically given in the context that was presented to the speaker, removing effects of information structure and information status. The contexts triggering RQs contained a sentence stating that it is generally known that nobody agrees with the proposition stated. The contexts triggering ISQs stated that the speaker was requesting information. The interrogatives in (3) and (4) serve as examples of *wh*- and polar questions used in the experiment.

(3)   Wer   mag   denn   Sellerie?
        Who   likes   PRT   celery
        'Who likes celery?'

(4)   Mag   denn   jemand   Sellerie?
        Likes   PRT   anyone   celery
        'Does anyone like celery?'

[10] tested twelve native speakers of German and analyzed 249 polar questions (125 ISQs, 124 RQs) and 253 *wh*-questions

(126 ISQs, 127 RQs). The participants were instructed to read the context descriptions presented on a screen silently, and to produce the subsequent target interrogative aloud. They then produced the questions in such a way that they were suitable in the given context.

The results of the prosodic analysis showed no clear distinction between final rise and final fall in polar ISQs vs. polar RQs, but instead revealed the use of a mid-level plateau H-% in polar RQs (67% of the cases vs. <1% in polar ISQs). Furthermore, the high-rising H-ˆH% edge tone appeared in 23% of polar RQs vs. 88% in ISQs [10]. Thus, the type of high edge tone helps distinguish between ISQs and RQs in polar questions. In *wh*-questions, RQs ended typically in a low edge tone L-% that also occurred in *wh*-ISQs, but there was more variation and a higher proportion of rising edge tones in *wh*-ISQs than in *wh*-RQs. According to these findings, the edge tone is not a reliable marker to distinguish between ISQs and RQs within *wh*-questions. Table 1 compares the edge tones for ISQs and RQs in German reported in [1]'s corpus data and [10]'s experimental data.

Table 1: *Comparison of most frequent edge tones in German ISQs and RQs. We use GToBI [2] for both authors.*

		Kohler (2004)	Braun et al. (2019)
Polar	**ISQ**	Rise: H-ˆH%	Rise: H-ˆH%
	RQ		Rise: H-%
Wh-	**ISQ**	Fall: L-%	L-%, L-H%, H-ˆH%
	RQ		Fall: L-%

Regarding phonetic parameters in German questions, Niebuhr et al. [11] investigated speech rate and voice quality in German declarative questions and statements. They showed that declarative questions have a faster speech rate and a breathier voice quality than statements. Van Heuven and van Zanten [12] reported similar results: they examined the speech rate in polar questions and statements for three different languages: English, German and Dutch. They found a faster speech rate (by 15%) for questions in all three languages relative to statements. With respect to speech rate, [10] reported longer constituent durations in RQs compared to ISQs, on average a lengthening of 15.8% for polar questions and 17.6% for *wh*-questions.

2. Corpus data

We extracted ISQs and RQs from the German TV cooking show *Kitchen Impossible*. The broadcasting company (RTL) was informed about the usage of the data for prosodic analysis. In the show, two prominent star or television chefs compete against each other. Each chef sends their opponent to two different countries where they have to recreate a local culinary specialty without knowing the complete list of ingredients or the preparation procedure. Finally, a jury knowing the original dish awards points and selects the winner. Since the show is interactive, emotional, and contains authentic dialogues, it is well suited to collect interrogatives for the analysis of spontaneous speech. The conversations can be divided into three parts: (i) Conversations between the chef whose turn it is to guess the ingredients and to prepare the meal and a person in the background, (ii) the conversations during the preparation of the dish (mostly with a person behind the camera), and (iii) the conversations between the two competitors when watching the preparation of the meal together in retrospect.

2.1 Data selection

The data was collected from the second season of the TV show consisting of seven episodes, as well as five episodes from the third season, with a length of 180 minutes each. In total the data was collected from 36 hours of film material. The interrogatives were recorded from the episodes using the Windows 10 screen recorder. The audio files were then converted to WAV-files, annotated and analyzed using *Praat* [13].

The data set consists of interrogatives produced by nine German speaking chefs (all native L1 speakers; eight male and one female; average age = 46.7 years, SD = 3.7 years). Another female speaker was excluded because she is Swiss and spoke Standard German with a heavy Swiss German accent. The regional background of the remaining speakers is heterogeneous: Two speakers are from Schleswig-Holstein, all other speakers are from other German-speaking regions, including Baden-Wuerttemberg, Bavaria, Berlin, North Rhine-Westphalia and Thuringia in Germany, Tyrol in Austria and South Tyrol in Northern Italy. Initially, 402 interrogatives were collected. To match the syntactic structure of the interrogatives used in [9] and [10] as closely as possible, only verb-first (V1) polar questions and *wh*-questions with the *wh*-pronoun in first position were used in the analysis. The data were preselected by the third author. Of the 402 interrogatives, 150 were excluded because they did not match the target structure syntactically or pragmatically. Interrogatives which were excluded were tag questions (n=47), declarative questions (n=39), syntactically incomplete questions (n=12), self-addressed questions (n=9) and alternative questions (n=6), as well as 37 interrogatives that had to be excluded for acoustic reasons, e.g., because they were incomprehensible, or because they had a bias induced by negation, e.g., *Ist das nicht der Hammer?*, 'Isn't that awesome?'; cf. [14] for bias in polar questions.

The remaining 252 interrogatives were classified based on their semantic and pragmatic properties as either ISQs or RQs. To this end, the response behavior of the addressee was taken into account. If the addressee gave an answer that provided information filling the knowledge gap in the question, the question was classified as an ISQ. Questions that did not elicit a genuine answer were classified as RQs. This applied to interrogatives that were not answered as well as those which were answered by the addressee but the interrogative itself had the feel of an assertion (see [3] and [15]). Given the nature of the data (oral speech extracted from a TV show), it cannot be excluded that prosody also played a role in the process of classifying target interrogatives as either RQs or ISQs. Crucially, however, for all interrogatives, their situational and linguistic contexts were consulted before taking a decision as to their illocution types, and semantic and pragmatic properties were the key decision criteria.

Of the 252 interrogatives, 30 were polar RQs; the three other categories had higher numbers. To arrive at equal numbers for each of the four categories (polar ISQs, polar RQ, *wh*-ISQ, *wh*-RQ) more interrogatives were excluded based on the criterion of syntactic complexity (eliminating the most complex ones). The final data set for prosodic analysis consists of 120 questions, 30 for each category.

2.2 Analysis

For the transcription and annotation of the data, we used the GToBI set of conventions [2]. The analysis treats the final edge tones, nuclear accents and frequent nuclear tunes as phonological variables. As a phonetic variable, we included speech rate as a proxy for constituent durations. An analysis of voice quality was not possible due to too much background noise.

2.1.1. Prosodic analysis

The data were transcribed and annotated separately by the first and the last author. 13 data points were removed because the background noise was too strong for reliable analysis. Agreement for the complete nuclear tune was 64%, with a kappa of 0.84 for nuclear pitch accents and 0.68 for edge tones. The most frequent disagreements were between accent types L*+H and L+H*, which are often confused by labelers [16-18], and between edge tones H-% and H-^H%. The 36 cases with disagreements were also labelled by the second author and one more researcher trained in GToBI, and the majority vote of all 4 annotators was taken. 18 cases of disagreements remained unresolved (two for whole nuclear contour, 7 for accent type only, and 9 for edge tone only). This leaves 98 data points for the analysis of nuclear pitch accents and 96 data points for the analysis of edge tones, see Table 2 and 3.

Table 2: *Number of illocution and question types in the data set for the analysis of nuclear pitch accents.*

Illocution type	Polar questions	Wh- questions
ISQs	23	28
RQs	23	24

Table 3: *Number of illocution and question types in the data set for the analysis of edge tones.*

Illocution type	Polar questions	Wh- questions
ISQs	25	27
RQs	18	26

2.1.2. Speech rate

We measured the speech rate dividing the intended number of syllables by the utterance duration in seconds [19]. To this end, the start and the end of the utterances were marked by the first two authors, relying on the broad band spectrogram, the auditory input and standard segmentation criteria [20]. In case of doubt, because of background noise, reliable boundaries were used (e.g., vowel onset in phrase initial [bi] sequence when a burst was not detectable). Utterances with hesitations or pauses were excluded from the analysis of speech rate, leaving 100 data points. The average deviance between the two labelers for the utterance start and end was 5.8ms (SD 1.8ms).

3. Results

The distribution of nuclear pitch accents and edge tones across the two question types and illocution types is shown in Figures 1 and 2.

3.1 Pitch accents

For statistical analysis, we recoded the dependent categorical variables pitch accent type and edge tone as binomial [21]. To this end, we coded whether a particular accent type (e.g., L*+H) was present or not. This binomial dependent variable was then modelled using logistic hierarchical regression models [22] in R working with the package lme4 [23], first with an interaction term between the two independent variables *illocution type* and *question type* and with *speaker* as random effects factor. The interaction term was removed if it was not significant at $p < 0.1$. Then, random intercepts were included if this improved the fit of the model. To account for the fact that multiple variables were tested, we adjusted the p-values using the Benjamini-Hochberg correction [24].

For the accent type **L*+H,** there was no interaction between illocution type (RQ, ISQ) and question type (*wh*, polar) ($p > 0.9$), no main effect of question type ($p > 0.6$), but there was a significant effect of illocution type (ß = -1.7, SE = 0.6, z = -2.9, p = 0.008). For the low accent, **L*,** the main effect of illocution type was reversed: There was no interaction between illocution type and question type ($p > 0.9$), no main effect of question type ($p > 0.9$), but a significant main effect of illocution type (ß = 1.77, SE = 0.6, z = 2.9, p = 0.02). For **L+H*,** there were neither main effects nor an interaction (all $p > 0.3$). For the other accent types, there were too few data points to conduct statistical analyses.
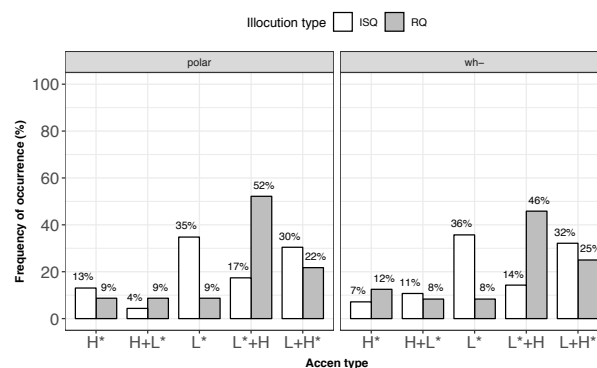

Figure 1: *Proportion of nuclear accents across illocution types and question types.*

3.2 Edge tones

For the edge tone **L-%,** the interaction between illocution type and question type approached significance (p = 0.07). For polar questions, there was no effect of illocution type on the frequency of L-% ($p > 0.5$), but the difference in frequency of L-% approached significance for *wh*-questions (ß = -1.13, SE = 0.6, z = 1.87, p = 0.08). For the high-rise, **H-^H%,** there was no interaction between illocution type and question type ($p > 0.6$). The effect of question type approached significance (p = 0.07), and there was a significant main effect of illocution type (ß = 1.98, SE = 0.5, z = 3.5, p = 0.009). The effects for the high plateau, **H-%**, were similar, but the effect of illocution type went in the opposite direction: There was no interaction between illocution type and question type ($p > 0.7$), no main effect of question type ($p > 0.18$), but a significant effect of illocution type (ß= -1.76, SE = 0.7, z = -2.4, p = 0.05). For the other edge tones, there were too few instances to carry out statistical comparisons.
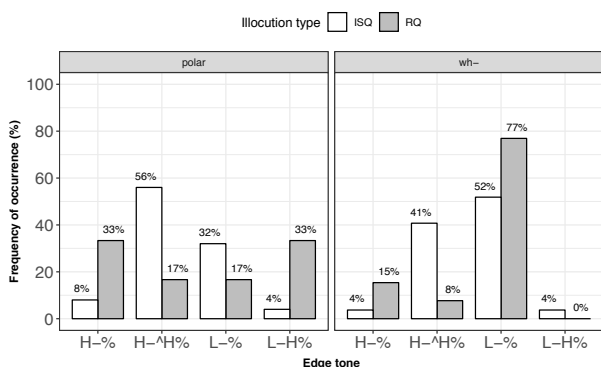
Figure 2: *Proportion of edge tones across illocution types and question types.*

3.3 Speech rate

Speech rate was analyzed with linear mixed effects regression models with *speaker* as random effects factor. The modeling was similar to the categorical data. There was no interaction between illocution type and question type (p > 0.19), no effect of question type (p > 0.18), but a significant effect of illocution type (ß = 1.1, SE = 0.4, t = 86.6, p = 0.04), i.e., ISQs were faster: they were produced with almost 1.3 syllables more per second than RQs. Relative to the speech rate of ISQs, this is a decrease in speech rate of 17% for RQs.
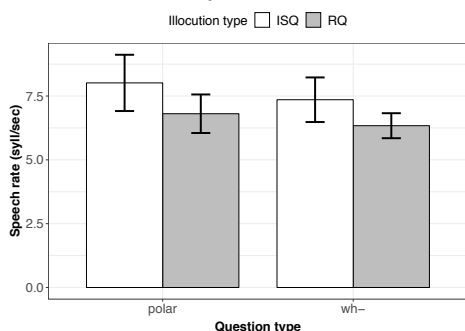

Figure 3: *Speech rate in syllables per seconds across illocution types and question types.*

4. Discussion

Overall, we find important commonalities between spontaneous speech and lab data, as well as some minor differences. Like in the experimental production study in [10], L*+H is the most frequent accent type in RQs, independent of question type (polar or *wh*-question). It was used in about half of the RQ productions, and was considerably less frequent in ISQs (<20% here, approx. 2% in lab speech). L*+H must therefore generally be considered an important cue to RQ interpretation. The most frequent accent type in spontaneous ISQs was a low accent, L* in both question type, while in the lab data, L* was the most frequent in polar ISQs, but not in *wh*-ISQs (outnumbered by L+H*). Regarding edge tones, the high-rising H-^H% is the most frequent accent type in polar ISQs in both data types. The high plateau (H-%) and the low-rising L-H% were equally frequent for spontaneous polar RQs, while in lab speech, H-% occurred with higher frequency. Across data type, *wh*-questions most frequently ended in a low L-%, but

more so for *wh*-RQs. *Wh*-ISQs were moreover frequently produced with a high-rising H-^H%, and more so in spontaneous than in lab speech. Spontaneous speech and lab speech also differ in the occurrence of L-H% such that L-H% is less frequent in spontaneous *wh*-ISQs than in *wh*-ISQs produced in the lab. At the same time, L-H% is more frequent in the spontaneous production of polar RQs than in lab speech polar RQs.

One reason for the fact that there is more variation in the results for accent types and edge tones in the present study compared to the experimental data – other than production settings – may be the more heterogeneous regional origins of the speakers. More than 50% of the spontaneous data was produced by speakers from northern Germany, while the lab data were recorded in the south. However, it is important to note that despite the differences in the results for spontaneous speech vs. lab speech, with regard to intonational categories, the tendencies are the same across varieties and data type.

Regarding speech rate, our data show that the overall duration for RQs is longer than for ISQs both for *wh*- and polar question. This result clearly goes in the same direction and has the same magnitude as [10]. Previous perception data have shown that listeners make use of prosodic cues to RQ vs. ISQ meaning [25, 26] in an experimental setting. Here, we show that speakers use these cues also in spontaneous settings to help their listeners deduce the intended meaning. More variation in prosodic realization is perhaps possible here, because the situational context serves as a cue to intended meaning along with linguistic context. Also, some of the examples from the TV cooking show were lexically unambiguously rhetorical or at least had a very strong bias for rhetorical interpretation (e.g., *Wollt ihr mich verarschen?* 'Are you kidding me?', *Ist das euer Ernst?* 'Are you serious?', *Ist der wahnsinnig?* 'Is he crazy?', *Wer will denn schon verlieren?* 'Who wants to lose?'). Questions like these may need less accurate prosodic marking to be understood compared to the experimental data, which were carefully controlled for lexicon and syntax.

Based on two methodologies (controlled experiments, spontaneous speech) we identified some very reliable cues to rhetorical meaning (e.g., L*+H nuclear accent and H-% for RQs). Future research will further determine the role of the situational and linguistic context, as well as the relative strength of context-independent cues to rhetorical vs. information-seeking interpretation.

5. Summary and Conclusions

In sum, while there is a lot of variation in these spontaneous data, we find important commonalities between lab speech and spontaneous speech to mark rhetorical vs. information-seeking meaning in interrogatives, a finding which lends validity to the experimental data collected under laboratory conditions.

6. Acknowledgements

7. References

[1] K.J. Kohler, "*Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions*", in *From traditional phonology to modern speech processing - Festschrift für Professor Wu Zongji's 95th Birthday*, G. Fant, et al., [Eds], Foreign Language Teaching and Research Press: Bejing. p. 205-215, 2004.

[2] M. Grice, S. Baumann, and R. Benzmüller, "*German Intonation in Autosegmental-Metrical Phonology*", in *Prosodic Typology. The Phonology of Intonation and Phrasing*, J. Sun-Ah, [Ed], Oxford University Press: Oxford. p. 55-83, 2005.

[3] M. Biezma and K. Rawlins, "*Rhetorical questions: Severing asking from questioning*", in *Proceedings of SALT 27*, D. Burgdorf, et al., [Eds]. p. 302-322, 2017.

[4] C.-H. Han, "*Interpreting interrogatives as rhetorical questions*". Lingua, 112: p. 201-229. 2002.

[5] J.A.G. Groenendijk and M.J.B. Stokhof, "Studies on the semantics of questions and the pragmatics of answers", Institute for Logic, Language and Computation (ILLC), 1984

[6] I. Caponigro and J. Sprouse, "*Rhetorical questions as questions*", in *Proceedings of Sinn und Bedeutung 11*, E. Puig-Waldmüller, [Ed], Universitat Pompeu Fabra: Barcelona. p. 121-133, 2007.

[7] C. Bartels, "*The Intonation of English Statements and Questions. A Compositional Interpretation*". New York & London: Garland Publishing, 1999.

[8] A. Banuazizi and C. Cresswell. "*Is that a real question? Final rises, final falls, and discourse function in yes-no question intonation*". in *35th Annual Meeting of the Chicago Linguistics Society* Chicago, 1999.

[9] D. Wochner, et al. "*The prosody of rhetorical questions in German*". in *Proceedings of Interspeech*. Dresden, Germany, 2015.

[10] B. Braun, et al., "*The prosody of rhetorical and information-seeking questions in German*". Language and Speech, 62(4): p. 779-807. 2019.

[11] O. Niebuhr, et al., "*Intonationsfragen hinterfragt! - Die Vielschichtigkeit der prosodischen Unterschiede zwischen Aussage- und Fragestäzen mit deklarative Syntax [On the complexity of prosodic differences between declaratives and declarative questions]*". Zeitschrift für Dialektologie und Linguistik, 77: p. 304-346. 2010.

[12] V.J.v. Heuven and E.v. Zanten, "*Speech rate as a secondary prosodic characteristic of polarity questions in three languages*". Speech Communication, (47): p. 87-99. 2005.

[13] P. Boersma and D. Weenink, "*Praat: doing phonetics by computer*": http://www.praat.org/, retrieved 11 May 2018, 2018

[14] F. Domaneschi, M. Romero, and B. Braun, "*Bias in polar questions*". Glossa: a journal of general linguistics, 21(2): p. 23. 2017.

[15] H. Rohde, "*Rhetorical questions as redundant interrogatives*", in *San Diego Linguistics Papers 2*. p. 134-168, 2006

[16] M. Grice, et al. "*Consistency in transcription and labelling of German intonation with GToBI*". in *4th International Conference on Spoken Language Processing (ICSLP)*. Philadelphia, 1996.

[17] J.F. Pitrelli, M. Beckman, and J. Hirschberg. "*Evaluation of prosodic transcription labelling reliability in the ToBI framework*". in *3rd International Conference on Spoken Language Processing (ICSLP)*. Yokohama, Japan, 1994.

[18] R. Herman and J.T. McGory, "*The conceptual similiarity of intonational tones and its effect on intertranscriber reliability*". Language and Speech, 45(1): p. 1-36. 2002.

[19] J. Trouvain, et al. "*Articulation rate measures and their relations to phone classification in spontaneous and read German speech*". in *ISCA Workshop on Adaptation Methods for Speech Recognition*. Sophia Antipolis, France, 2001.

[20] A. Turk, N. Satsuki, and M. Sugahara, "*Acoustic segment durations in prosodic reserach: A practical guide*", in *Methods in Empirical Prosody Research*, S. Sudhoff, et al., [Eds], De Gruyter: Berlin, New York. p. 1-28., 2006.

[21] A. Agresti, "*Categorical data analysis*". Wiley series in probability and statistics. New York: John Wiley, 2002.

[22] H.R. Baayen, "*Analyzing linguistic data. A practical introduction to statistics using R*". Cambridge: Cambridge University Press, 2008.

[23] D. Bates, et al., "*lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://CRAN.R-project.org/package=lme4*", 2014

[24] Y. Benjamini and Y. Hochberg, "*Controlling the false discovery rate: a practical and powerful approach to multiple testing*". Journal of the Royal Statistical Society, Series B 57: p. 289-300. 1995.

[25] M. Kharaman, et al. "*The processing of prosodic cues to rhetorical question interpretation: Psycholinguistic and neurolinguistics evidence*". in *Interspeech*. Graz, Austria, 2019.

[26] M. Xu, "Die Rolle der Dauer bei der Perzeption von rhetorischen Fragen [The role of duration in the perception of rhetorical questions]", MA thesis, Department of Linguistics, University of Konstanz, 2019