The processing of prosodic cues to rhetorical question interpretation: Psycholinguistic and neurolinguistics evidence

Mariya Kharaman, Manluolan Xu, Carsten Eulitz, Bettina Braun

Department of Linguistics, University of Konstanz, Germany

{mariya.kharaman,Bettina.braun,Carsten.eulitz}@uni-konstanz.de,
xumanluolan@163.com

Abstract

In many languages, rhetorical questions (RQs) are produced with different prosodic realizations than string-identical information-seeking questions (ISQs). RQs typically have longer constituent durations and breathier voice quality than ISQs and differ in nuclear accent type. This paper reports on an identification experiment (Experiment 1) and an EEG experiment (Experiment 2) on German *wh*-questions. In the identification experiment, we manipulated nuclear pitch accent type, voice quality and constituent duration and participants indicated whether they judged the realization as ISQ or RQ. The results showed additive effects of the three factors, with pitch accent as strongest predictor. In the EEG experiment, participants heard the stimuli in two contexts, triggering an ISQ or RQ (blocked). We manipulated pitch accent type and voice quality, resulting in RQ-coherent and ISQ-coherent stimuli, based on the outcome of Experiment 1. Results showed a prosodic expectancy positivity (PEP) for prosodic realizations that were incoherent with ISQ-contexts with an onset of ~120ms after the onset of the word with nuclear accent. This effect might reflect the emotional prosodic aspect of RQs. Taken together, participants use prosody to resolve the ambiguity and event-related potentials (ERPs) react to prosodic realizations that do not match contextually triggered expectations.

Index Terms: Prosody, voice quality, duration, pitch accent type, rhetorical question, event-related potentials (ERP), prosodic expectancy positivity (PEP)

1. Introduction

In linguistics, questions are formally distinguished into constituent questions (often referred as *wh*-questions), polar questions (often referred to as yes/no questions) and alternative questions. In terms of meaning, not all questions serve a purely information-seeking purpose (filling a knowledge gap). In this paper we are concerned with a special type of non-information seeking questions, rhetorical questions (RQs).

RQs have the surface structure of an interrogative, but are often used to "extract a commitment to the rhetorical point" from an interlocutor [1], p. 304. While for information-seeking questions (henceforth ISQs) there is a high degree of uncertainty as to the answer on the part of the speaker, for RQs there is no uncertainty. Instead, the answer is in the common ground and obvious to all interlocutors [1-3] or is intended to be added to the common ground [1]. Biezma and Rawlins [1] further argue that for a question to be interpreted as an RQ, it "must conventionally indicate the speaker's attitude […] that the question they are asking is non-inquisitive in context" [1] , p. 306-307. Corpus analyses have shown that rhetorical questions differ from information-seeking questions in their prosodic realization [4, 5], regarding boundary tones and pitch accents. Previous experimental work in German has shown that questions in a rhetorical context differ from questions in an information-seeking context in terms of their prosodic realization [6]. RQs are more often realized with nuclear L*+H accents than ISQs, have longer constituent durations and more often a breathy voice quality, irrespective of question type (*wh*-question or polar question). Neitsch, Braun and Dehé [7] used the visual world paradigm to test listeners' sensitivity to pitch accent type and voice quality (while keeping duration constant) in *wh*-questions (e.g., *Who likes lemons*?). Their results showed that listeners interpreted nuclear L*+H accents with breathy voice quality as RQ in 90% of the cases, while nuclear H+!H* (peak accent with the peak before the accented syllable) with modal voice quality was interpreted as ISQ in 90% of the cases.

In this paper we go two steps further. We first add the additional cue duration in a 2x2x2 design, manipulating constituent durations (lengthened or shortened by 10%), voice quality (breathy vs. modal voice quality on the object) and nuclear pitch accent type (L*+H vs. H+!H*). Second, we study event-related brain potentials (ERP) to different degrees of expectancy violations which should be reflected in subcomponents of the P300 group including the prosodic expectancy positivity (PEP). These components have been successfully used to investigate intonation processing [8-10] and the processing of speech rhythm [11, 12]. The idea is to investigate if prosodic patterns associated with different question types are also distinguishable in their processing as reflected in different ERP responses. As a future goal we plan to establish a hierarchy of the cues enabling the categorization of IS and RQ prosodic patterns.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Sixteen native German speakers (aged 19-30 years, mean age 23.1 years, SD=3.2, 11 female) who were unaware of the purpose of the experiment participated for a small fee.

2.1.2. Materials

The sentence materials were identical to Neitsch, Braun and Dehé [7]. They consisted of 32 short *wh*-questions, starting the *wh*-word *wer* ('who') followed by a finite verb, the modal particle *denn* and a sentence-final object noun (e.g., *Vanille* 'vanilla'). The object nouns were trisyllabic nouns with word stress on the penultimate syllable. The 32 questions were recorded in four conditions, crossing pitch accent realization

(nuclear H+!H* and L*+H) and voice quality on the object (breathy vs. modal). In the eye-tracking study in Neitsch, Braun and Dehé [7], the durations of the questions were normalized to have equal analysis windows, see Figure 1.

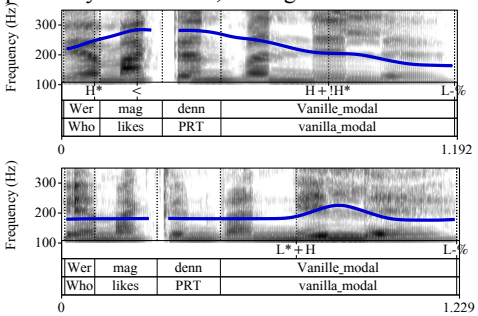

Figure 1. *Example contours for the two pitch-accent conditions in Experiment 1 (top: early-peak accent (H+!H*) on the noun, bottom: late-peak accent (L*+H) on the noun), from [5].*

For the current study, the normalized durations of each constituent were uniformly lengthened by 10% for the long version and shortened by 10% to arrive at a short version. For each object noun, a corresponding colour picture was selected (500x500 pixels), the same as in Neitsch, Braun and Dehé [7].

2.1.3. Procedure

Pitch accent type, voice quality and duration were manipulated within subjects. There were eight basic lists that contained all the 32 items, with pitch accent type and voice quality manipulated in a Latin Square Design and duration manipulated between items. Each participant was assigned to two of the eight lists. Hence, each participant heard each item (question) twice, in two of the eight conditions and responded to 64 experimental trials overall. The order of stimuli in the lists was pseudo-randomized with the constraint that the same conditions were separated by at least two other trials and that the same lexical item was separated by at least eight other trials.

Participants were informed that there are differences between ISQs and RQs by means of unambiguous examples (ISQ: *What time is it?*, RQ: *Who likes paying taxes?*). They were then informed that this difference is not always that clear cut and that it is sometimes marked by how the question is asked. They were then seated in front of a computer with a three-button box, whose buttons were labelled (real question, other, rhetorical question). Each trial started with a 2500ms display of the visual object, followed by an auditory stimulus (played over headphones). They had to respond as quickly and accurately as possible. There was no timeout and no feedback.

2.2. Results

Participants clicked the middle button in only 32 items (3% of the overall data), suggesting that due to their prosody, the stimuli were clearly identifiable as either ISQ or RQ. The results for RQ-responses are depicted in Figure 2.

The responses were analyzed with logistic mixed effects regression models using the lme4-package [13-15]. The initial model included all fixed effects and participants and items as crossed random effects. Random slopes were added if this improved the fit of the model, as estimated by comparison of the models' LogLikelihood. RQ clicks were most frequent in interrogatives with an L*+H accent, breathy voice and long duration, and became less frequent when any of the factors changed. Pitch accent type had the strongest influence ($\beta = 3.48$, SE = 0.45, z = 7.74, $p < 0.0001$). The effects of voice quality and duration were similar to one another, but smaller than the effect of accent type (voice quality: $\beta = 1.83$, SE = 0.21, z = 8.43, $p < 0.0001$; duration: $\beta = 1.56$, SE = 0.19, z = 8.14, $p < 0.0001$). None of the interactions was significant (all $p > 0.13$).

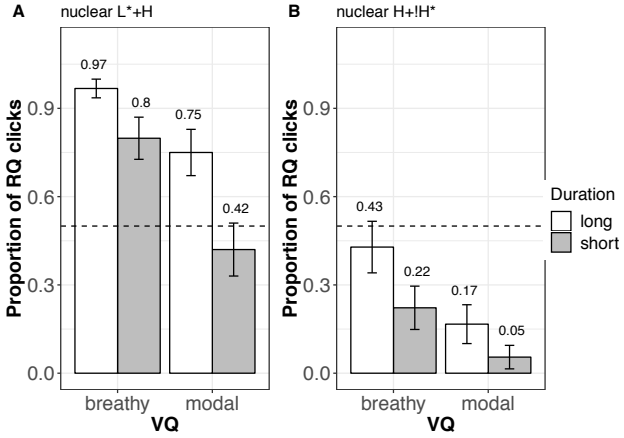

Figure 2. *Clicks to RQs in questions with a nuclear L*+H accent (panel A) and nuclear H+!H* accent (panel B), split by voice quality and duration. The horizontal line indicates chance level of 0.5. Whiskers represent standard errors.*

Reaction times (RTs) were recorded relative to the end of the question. RTs smaller 0 and larger than 2000ms were removed. This left 924 data points (out of 992). The remaining RTs were converted to square-root to make their distribution more normal. The average sqrt-RTs are shown in Figure 3 for both ISQ clicks (top panels) and RQ clicks (bottom panels).

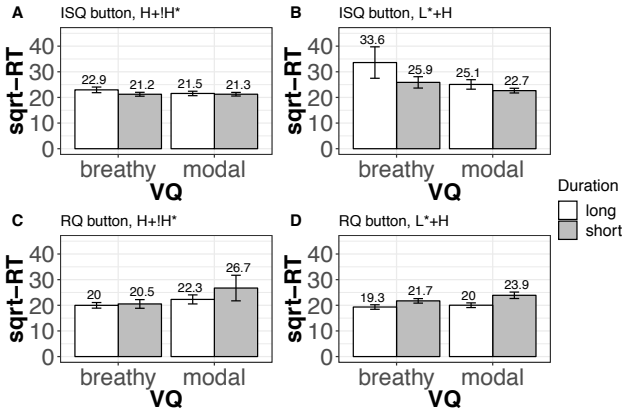

Figure 3. *Mean RTs when participants clicked the ISQ button (panels A and B) and when they clicked the RQ-button (panels C and D).*

RTs were analyzed with a linear-mixed effects regression model with button, accent type, duration and voice quality as fixed factors and participants and items as crossed random factors. The most complex model was tested first and the model was simplified using the step-function in R. The most parsimonious model showed three two-way interactions, all involving the participants' choice: there were significant interactions between button and voice quality ($F(1,893) = 24.7$, $p < 0.0001$), button and duration ($F(1,875) = 35.8$, $p < 0.0001$), and between button and accent type ($F(1,902) = 19.3$, $p <$

0.0001). An anonymous reviewer suggested to use general linear mixed models with raw RTs and a Gamma linking function [16]. This analysis yielded very similar results. The panels with most of the data are panel A for ISQ interpretation and panel D for RQ interpretation because these had the congruent pitch accent. The statistical analysis of reaction times for these conditions showed that for RQ choices (panel D in Figure 3) a change to shorter durations led to longer reaction times ($F(1,14.8) = 15.8$, $p = 0.001$), while the change to modal voice quality had no effect ($p = 0.1$). For ISQs (panel A in Figure 3) there were no effects of duration or voice quality on RTs and no interaction (all $p > 0.2$).

2.3. Discussion

Pitch accent type had the strongest effect on participants' interpretations, while duration and voice quality do not differ in relative importance. Our current generalization is that pitch accent type is a strong but not a sufficient cue to illocution type (L*+H for RQs, H+!H* for ISQs). To unambiguously mark illocution type, it must be complemented by one of the phonetic parameters, i.e., duration or voice quality. This ties in with results from classification and regression trees [17]. The analysis of reaction times (change to shorter utterance durations led to longer reaction times, while the change to modal voice quality had no effect on reaction times) suggests that longer duration is more constitutive to RQs than breathy voice quality.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Twelve native German speakers (19-25 years, mean age 22.6, SD= 2.15, 9 female) with no history of hearing problems or neurological impairment signed an informed consent before participating in the study. All were right-handed as tested by the Edinburgh Handedness Inventory [18] with the lateralization quotient > 90. An audiogram (AudioConsoleVers. 2.4.8 for Windows, Inmedico; frequency selection 0.125 KHz - 8 KHz) was measured before the experiment (participants could not participate if they did not reach 20 dB hearing acuity threshold). They received a small fee for their time and effort. As only trials with correct identification of the question types were analyzed, two participants with low correctness rates had to be excluded.

3.1.2. Materials

The sentence materials were identical to those used in Experiment 1, including manipulations on pitch accent realization (H+!H* and L*+H) and voice quality on the object (breathy vs. modal). No durational manipulations were applied to the original stimuli of [7], i.e., the constituent durations were matched for the two voice quality conditions. This resulted in four different prosodic conditions.
For cueing the question types, two pictures were used. For RQ it was always a pile of money to symbolize an unambiguous RQ sentence like *Who likes paying taxes?* (see left panel in Figure 4). For ISQ it was always a watch to symbolize an unambiguous ISQ sentence like *What time is it?* (see right panel in Figure 4).

3.1.3. Procedure

The task for the participants was to judge for each sentence they heard as to whether or not it matched with the cued illocution type (visual RQ or ISQ cue).

Who likes paying taxes? *What time is it?*

Figure 4. *Visual and written contexts to cue an RQ (left panel) or ISQ (right panel) interpretation in Experiment 2.*

A single trial started with the presentation of the visual RQ or ISQ cue (which was the same for all trials) and a written sentence which was presented below the picture for 2000 ms. This was followed by a central fixation cross. 500 ms after the appearance of the fixation cross, the spoken sentences were presented via headphones. The fixation cross remained on the screen until the end the auditory presentation of the sentence. Then, a question mark appeared on the screen together with a repetition of the visual question-type cue and prompted the participants to give their judgements. The right button was used for questions that were judged as coherent with the visual cue, the left button for questions that were judged as incoherent. The next trial started after the button press or 4000 ms after the onset of the question mark (time-out).
The 32 sentences were presented in all four prosodic conditions (within-subjects manipulation) and presented both with the visual ISQ and RQ cue, resulting in 256 trials. There were 32 blocks. Within each block, the visual cue remained constant. Within a block with the same visual cue, the four prosodic conditions were randomized. The overall duration of the experimental session was dependent on the decision time of the participants and lasted between 20 and 30 minutes.
The EEG was recorded using BrainAmp (version 1.20.0502) and the Brain Vision Recorder (Brain Products GmbH) from 64 sintered Ag/AgCl electrodes using an electrode cap (EASY CAP equidistant 64 channels montage, EASYCAP GmbH) in the frequency range 0.016-250 Hz with Cz as the reference. Horizontal and vertical eye movements were registered by the electrodes positioned under the eyes and on the forehead, the ground electrode was attached to the right cheek. Impedances were kept below 5 kΩ. The EEG signal was digitized with a sampling rate of 500 Hz. The data were off-line processed with the FieldTrip [Matlab-based tool, 19]. Detrending was applied to the continuous raw data in the 4 s intervals. Eye movements were removed with the ICA procedure. Bad channels were interpolated, and the data were re-referenced to a common average reference. Epochs of 1000 ms including a 200 ms pre-stimulus baseline were extracted from the cleaned data, averaged and baseline corrected.

3.2. Results

To receive a sufficient number of data for analysis, two of the four prosodic conditions were pooled: The prosodic realization of the questions is termed coherent if the pitch accent was congruent with the context indicated by the picture (i.e., nuclear L*+H for an RQ context, nuclear H+!H* for an ISQ context), irrespective of voice quality. We kept the pitch accent contrast, because this was the strongest cue in Experiment 1. ERPs were only considered for "correct" judgments, i.e. trials for which participants judged the match between visual cue and auditory stimulus correctly (false for incoherent combinations and true for coherent combinations). The average correctness was

between 83% and 87% in all conditions in the remaining artefact-free trials.

In the following, we first compare the ERP responses to the two types of auditory questions (RQ prosody and ISQ prosody, determined by the pitch accent) in a visual context that was incoherent or coherent with the prosodic form of the question. This results in difference waveforms (incoherent minus coherent prosodic condition), see Panel A of Figure 5. The red line shows the prosodic condition with an RQ prosody (incoherent minus coherent), the blue line with an ISQ prosody incoherent minus coherent). The zero point on the x-axis is time-locked to the onset of the final noun in the sentences. As shown in the top left of Figure 5, the ERP responses over frontal electrode sites to RQ realizations (red line) elicited more positive going amplitudes starting at about 50 ms after the onset of the sentence-final grammatical object noun in incoherent ISQ context cues compared to coherent RQ cues. In parietal and occipital electrode sites (top right of Figure 5), the ERP responses to RQ realizations resulted in more negative going amplitudes in incoherent compared to coherent contexts. For prosodic ISQ realizations (blue lines), the type of visual context cue lead to an opposite pattern: a negativity in frontal regions (top left) for an incoherent vs. coherent visual cue and a positivity in parietal and occipital regions (top right)

Next, we calculated topographies of the ISQ and RQ difference waveforms by subtracting the waveform triggered by a coherent visual context from the coherent context condition. The mean topographies from 0 to 900ms after noun onset (bottom panel B of Figure 5) support the differences between the prosodic realizations. For prosodic RQ realizations (bottom right), the topographies show the positivity (orange color shades) over frontal electrode sites and the negativity (blue color shades) over the parietal and occipital regions. For prosodic ISQ conditions (bottom left), we see negativity over frontal regions and positivity over parietal and occipital regions. The described amplitude difference was present in 8 of the 10 participants.

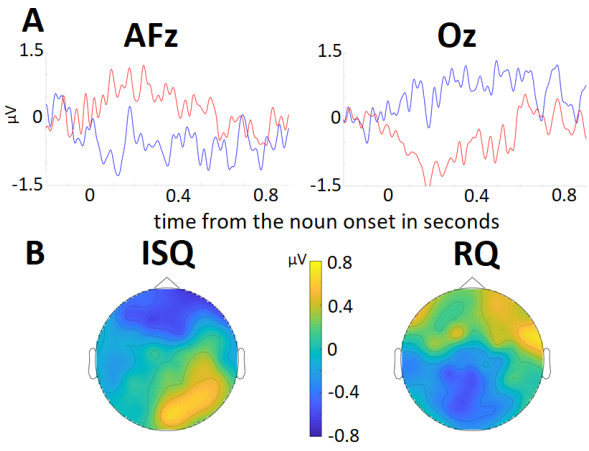

Figure 5. *A. Difference waveforms (incoherently minus coherently cued context) for ISQ (blue) and RQ (red) are shown for two selected electrode sites over the frontal and occipital brain regions. B. Mean topographies of the prosodic ISQ and RQ conditions (differences between incoherent and coherent visual cues) for 0-900 ms after the onset of the noun.*

3.3. Discussion

With respect to ERP effects to the object noun, we found a difference for the expectancy violations between the prosodic realizations of the ISQ and RQ conditions. Moreover, this difference occurred relatively early (~50 ms after the onset of the object noun). A possible explanation is that prosodic cues before the onset of the object noun were sufficient to affect the processing of the stimulus and the evaluation of the match to the visual context cue. Furthermore, the timing and the efficiency of these prosodic cues were different for the ISQ and the RQ condition. To support this interpretation, alternative data analyses using the whole sentence will be necessary. These data analyses might also pave the way to see error detection responses, which could be expected based on the literature (error-detection component N2b, prosodic expectancy positivity of P3 group), e.g., [8-10].

RQ realizations may come across more emotionally than ISQ realizations. Considering the emotional aspect, the early positivity in frontal regions might reflect the combined linguistic and emotional expectancy violation processing as observed by Paulmann et al [10], though it is not as broadly distributed in our case than in [10]. It is possible that a clearer pattern emerges when the prosodic conditions are fully separated instead of pooled over voice quality conditions. On the whole, the expectancy violation based on the pitch accent type categorization (RQ/ISQ) was observed in the ERPs.

On the whole, cuing the expectation with a single picture for each illocution type was successful.

4. General Discussion

In Experiment 1, participants judged questions without any context. Pitch accent type was a strongest cue for the decision of whether an auditory question was considered as RQ or ISQ, while voice quality and durational cues were less important to listeners. Hence, participants are able to consistently use prosodic cues (here pitch accent type, duration, and voice quality) to classify the illocution type in interrogatives even when there is no contextual information given.

The reliance of prosodic cues observed in the behavioral data also surfaces in participants' neurophysiological responses. In the EEG study presented in Experiment 2, we triggered two different expectations by means of a visual cue and then presented the auditory stimuli. Given that accent type was the strongest predictor in Experiment 1, we classified the prosodic realizations according to accent type (e.g., breathy and modal L*+H were both classified as RQ). The results showed that prosodic patterns, which did not match the expectations triggered by the visual context were reflected in the ERPs. ERPs showed early effects after the onset of the noun that carried the pitch accent. Currently, we plan further analyses on the additive effects of the less powerful cues shown in Experiment 1, such as voice quality and duration. Furthermore, we will test the cue weighting process during the RQ/ISQ categorization.

5. Acknowledgements

The work was supported by grants from the German research foundation to Bettina Braun (BR 3428/4-1,2) and Carsten Eulitz (EU 39/8-2) as part of the research unit FOR 2111 (Questions at the interfaces). We thank Jana Neitsch for providing the auditory stimuli and Katharina Zahner for valuable comments on an earlier version of the manuscript.

6. References

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

[2] Caponigro, I. 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.

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

[4] Hedberg, N., et al. *Prosody and pragmatics of wh-interrogatives*. in *2010 Meeting of the Canadian Linguistics Association*. 2010.

[5] Sicoli, M.A., et al., *Marked initial pitch in questions signals marked communicative function.* Language and Speech, 58(2): p. 204-223. 2015.

[6] Braun, B., et al., *The prosody of rhetorical and information-seeking questions in German.* Language and Speech, advance online publication. 2018.

[7] Neitsch, J., B. Braun, and N. Dehé. *The role of prosody for the interpretation of rhetorical questions in German*. in *9th International Conference on Speech Prosody*. Poznan, Poland, 2018.

[8] Kung, C., D.J. Chwilla, and H. Schriefers, *The interaction of lexical tone, intonation and semantic context in on-line spoken word recognition: an ERP study on Cantonese Chinese.* Neuropsychologia, 53: p. 293-309. 2014.

[9] Liu, M., Y. Chen, and N.O. Schiller, *Online processing of tone and intonation in Mandarin: Evidence from ERPs.* Neuropsychologia, 91: p. 307-317. 2016.

[10] Paulmann, S., S. Jessen, and S.A. Kotz, *It's special the way you say it: an ERP investigation on the temporal dynamics of two types of prosody.* Neuropsychologia, 50(7): p. 1609-20. 2012.

[11] Bohn, K., et al., *The influence of rhythmic (ir)regularities on speech processing: Evidence from an ERP study on German phrases.* Neuropsychologia, 51(4): p. 760-71. 2013.

[12] Henrich, K., R. Wiese, and U. Domahs, *How information structure influences the processing of rhythmic irregularities: ERP evidence from German phrases.* Neuropsychologia, 75: p. 431-40. 2015.

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

[14] Baayen, R.H., D.J. Davidson, and D.M. Bates, *Mixed-effects modeling with crossed random effects for subjects and items.* Journal of Memory and Language, 59(4): p. 390-412. 2008.

[15] Bates, D., et al., *Fitting Linear Mixed-Effects Models Using lme4.* Jounral of Statistical Software, 67(1): p. 1-48. 2015.

[16] Lo, S. and S. Andrews, *To transform or not to transform: using generalized linear mixed models to analyse reaction time data.* Frontiers in Psychology, 6: p. 1171. 2015.

[17] Braun, B., et al. *Classification of interrogatives as information-seeking or rhetorical questions*. in *17th Speech Science and Technology Conference*. Sydney, Australia, 2018.

[18] Oldfield, R.C., *The assessment and analysis of handedness: the Edinburgh inventory*. Neuropsychologia, 9(1): p. 97-113. 1971.

[19] Oostenveld, R., et al., *FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data.* Comput Intell Neurosci, 2011: p. 156869. 2011.