A case for systematic sound symbolism in pragmatics:

The role of the first phoneme in question prediction in context

Anita Slonimska
s4415000
Research Master's in Language and Communication
Radboud University
Nijmegen, The Netherlands

Supervisor: Dr. Sean Roberts
Second reader: Dr. Sara Bögels

2016

TABLE OF CONTENTS

Abstract

Conversation is a socially and cognitively demanding endeavor in which interlocutors have to continuously monitor what is being said in order to react fast and appropriately. This is even more demanding with questions as they put an obligation on an addressee to respond. In the present study we investigated whether the first phoneme of question words helps in predicting an incoming turn as a question. Importantly, given that conversation always occurs in context, we investigated how the type of previous sequential turn influences question recognition.

We addressed this topic by first investigating the hypotheses in naturally occurring conversations in a corpus. Then, we tested these findings in a controlled setting. In the corpus study we used the method of *the decision trees* to assess the influence of the first phoneme and the context on probability of an incoming turn being a question. In the experimental study, we designed a behavioral task in which participants had to predict an incoming turn once they heard the recording (from the same corpus) of the previous turn and the first segment of an incoming turn.

Both studies confirmed that the first phoneme of an incoming turn and the context play a role in question prediction. Namely, we found that if an incoming turn starts with a phoneme from question words (i.e., /w/ in English), participants are more likely to think that an incoming turn is a question in comparison to other phoneme or no phonemic cue at all. Also, questions are expected more, if a turn is preceded by a non-initiating turn in comparison to an initiating turn. Interestingly, the corpus study suggests that the phoneme is the strongest factor in question recognition and also that this effect should be stronger in non-initiating context. Nevertheless, in the experiment we find that context is a stronger factor than phoneme and there is no significant interaction between phoneme and context, even though the trend is in the predicted direction.

The present study provides the first support for the hypothesis that early phonemic cue plays a role in question recognition, also with context available. Moreover, this is the first study to approach this phenomenon in ecologically valid and controlled ways. Both similarities and differences in the results from both studies highlight the importance of such approach in research.

1. Introduction

The time that people spend on speaking is estimated to be 2-3 hours per day on average and during this time period speakers can produce up to 1200 turns (Levinson, 2016). Interestingly, even though conversation can be considered the predominant form of language use (Levinson, 2006), only relatively recently it has been taken notice that the mechanism of conversation itself is quite remarkable in its own right (Sacks, Schegloff & Jefferson, 1974; Levinson, 2016).

When people talk to each other they take turns to deliver speech acts. This turn – taking is a puzzling phenomenon as it happens surprisingly fast. Within an average of 200 ms speakers are capable of delivering an appropriate speech act in response to the previous turn (Levinson & Torreira, 2015). This is even more surprising with questions as they put an obligation on an addressee to provide an answer tailored to the question (Sacks, Schegloff & Jefferson, 1974). Accordingly, there is a social pressure that, in turn, puts cognitive pressure on the addressee to comprehend and at the same time prepare the response in time. It has been proposed that there are cues early in a turn that can help in recognizing the speech act as a question and accordingly help in planning the response so that it can be delivered right after the question (Levinson, 2013).

Slonimska & Roberts (in prep.) put forward quite a controversial hypothesis in regard to a phonetic cue to questions. Namely, they argue that the fact that content question words tend to match in regard to their first phoneme indicates that it is a likely cue to question recognition. In the present paper we investigate whether the first phoneme of the turn is actually used as a cue in question prediction. Moreover, given that turn-taking never happens in isolation but is built on sequences, we are also interested in how previous context influences question recognition. Accordingly, the research questions of the present paper are:

- Is the first phoneme of content-question words a cue for question prediction?
- Does the sequential type of context influence question prediction?

Importantly, this is one of the first papers that aims to address this topic from both ecologically valid and experimentally controlled settings. Accordingly, we address these research questions by means of two studies. First, we explore a large corpus of natural conversations and subsequently use the insights from the corpus study to design an experiment in which we test the hypotheses in a controlled setting by using stimuli from the same corpus.

As such, the present project not only informs the theoretical field in regard to question recognition, but it also makes a case for a new approach to research – namely, by creating a synergy between ecologically valid qualitative analysis and experimentally controlled quantitative insights of the phenomena.

The paper is structured as follows: first, we provide background information on turn-taking, response planning and cues to question recognition. Then, we proceed to the first study – we analyze a large corpus of spontaneous conversations in American English by means of the method of *the decision trees* (Strobl, Malley, & Tutz, 2009) in order to explore whether we can find patterns of question recognition based on the first phoneme and context in natural data. Next, in order to test the hypotheses in a controlled setting, we carry out an experiment that is based on the findings of the corpus study. Finally, we compare findings from both studies, interpret the results and provide conclusions.

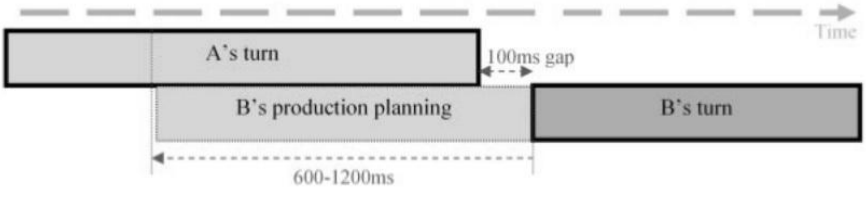
2. Background

2.1. Turn-taking

Conversation progresses though exchanging bursts of information – mostly through use of language – that are orchestrated in consecutive turns produced by the speakers (Sacks, Schegloff & Jefferson, 1974; Levinson, 2016). Thus, in a nutshell, conversation is an exchange of speech acts packed into turns of the interlocutors. The surprising aspect of turn - taking is that it is orchestrated in a remarkably tight manner. It has been estimated that, on average, gaps between turns are only 200ms long (Levinson & Torreira, 2015). Indeed, previous research shows that speakers tend to minimize gaps and overlaps between the turns (Stivers et al., 2009; Kendrick & Torreira, 2015). In other words, long overlap between turns appears to be rare and once it occurs one of the speakers retracts so that only one turn is maintained (Levinson & Torreira, 2015). Also longer gaps between turns are not common in conversation. In this regard, research suggests that delayed turns (i.e., turns that are longer than 350ms) can be interpreted as a hesitation, especially when initial turn of the sequence is produced in order to receive information (e.g., answer to a question, uptake of an offer, compliance with a request) (Roberts & Francis, 2013; Kendrick & Torreira, 2015; Levinson, 1983). Research shows that negative answers (e.g., *No* in contrast to *Yes*) are expected when there is a greater delay after the question and not when there is a shorter gap (Bögels, Kendrick, & Levinson, 2015). Importantly, these findings appear to be universal as many languages from different language families and areas exhibit similar patterns of short gaps between turns, minimal amount of overlapping turns and minimal amount of longer gaps between turns (Stivers et al., 2009). Thus, while languages themselves differ, the way they are used in conversation is quite similar.

The surprising fact that turns are produced in such a tight window of time becomes even more puzzling if we take into account that it takes a minimum of 600ms to plan (i.e., message, syntactic, phonological encoding) a single word (Schriefers, Meyer, Levelt, 1990; Levelt, 1993) (see Fig.1). In this context, one has to ask a question – how is it possible that the gap between turns is shorter than the planning of the response? The answer to this question is suggested to be prediction (Sacks, Schegloff & Jefferson, 1974; Levinson, 2013). In other words, research suggests that people are capable of projecting what the current speaker is roughly going to say and when his turn will end (Holler and Kendrick, 2015; Bögels & Torreira, 2015). Thus, the next speaker can start preparing their turn in advance so that it can be delivered on time.

Figure 1. Overlap of comprehension and production in conversation (Levinson, 2013, p.104)



The next logical question then is as follows– how is it possible that people are capable of predicting the incoming turn? The answer might lie in the fact that people make use of early cues (e.g., context, intonation, eye gaze) to predict what kind of turn is about to be produced (see Holler, Kendrick, Casillas, & Levinson, 2015 for a review). This aspect, namely predicting the specific type of a speech act, is extremely important as different speech acts have different social and cognitive pressures on speakers. For example, if we are greeted, the greeter expects a greeting in response. Just as when we are asked a question, we are socially obliged to give an answer. In terms of Sack, Schegloff & Jefferson (1974) the current speaker has selected the person to whom the question was referred as the next speaker. On the other hand, statements do not pose such a social pressure, as they do not require a specific responding action. In this light, the current speaker can maintain the floor or another person can self-select to speak.

In regard to questions, there is a social pressure for the person to respond. Thus, social pressure puts also a pressure on cognition in order to respond in a rapid way to be able to minimize the gap between the turns. For example, greetings are quite automatic while responses require thinking and retrieving - all this in the shortest period of time. Previous research suggests that planning of the response starts as soon as an answer can be retrieved. We review this in the next section.

2.2. Planning of a response to a question

Research suggests that the onset of articulation of a response is based on the turn-end cues of the speakers (Torreira, Bögels & Levinson, 2015). However, planning of the response to a question occurs immediately after (i.e., within a half of a second) the answer can be retrieved (Bögels, Magyari & Levinson, 2015; Bögels, Casillas, & Levinson, 2016).

In their experiment, Bögels, Magyari & Levinson (2015) presented participants with two kinds of questions – questions that had a crucial word for the answer retrieval in the middle of the sentence (e.g., *Which character, also called 007, appears in the famous movies*?) and questions that had the crucial word for the answer retrieval at the end of the sentence (*Which character from the famous movies, is also called 007?).* By means of ERP measures, they show that participants start planning the response right after they hear the crucial information. They also show that at this point in time they switch from comprehension to production planning. These findings were also replicated in their recent study, in which they also show that focus on production planning can interfere with comprehension (Bögels, Casillas, & Levinson, 2016).

This research clearly indicates when planning of the response to a question occurs. However, even before production planning, speakers first have to recognize that they are being asked a question. In previously described experiment the design of the study was framed as a quiz game. In other words, participants knew that they are being asked only questions. In real conversation there is a necessity to continuously monitor the incoming speech acts in order to first recognize what they are and only then react to them adequately (e.g., plan a response to a question). In other words, the preparation of the response is only possible when an addressee knows that what they are hearing is a question. Thus, even before starting to plan an answer to a question, recipient first has to recognize that the speech act that is being produced is a question and has to be answered to. Accordingly, there must be cues that give an "early start" for an addressee in regard to question recognition and "prepare" for answer planning.

2.3. Recognizing a question

Levinson (2013) suggests that question recognition is possible due to front-loading of the cues at the beginning of a turn. For example, front-loading can be observed in use of intonation (Levinson, 2013), pitch (Sicoli et al. 2014) and eye-gaze (Rossano, Brown & Levinson, 2009; Rossano, 2013). Sicoli et al. (2014) argue that speakers use pitch at the beginning of the utterance to differentiate between questions that are to be perceived directly – requesting information, and question that are to be perceived indirectly. As such, pitch can play an important role in not only helping people recognize a question, but also in differentiating whether this question is actually used with it's primary scope (i.e., requesting information). Rossano, Brown & Levinson (2009) suggest that speakers are more likely to maintain eye gaze when asking a question rather than shifting eye gaze away from the addressee.

Shifting question words to the initial position of the utterance (e.g., English) appears to be one of the most evident examples of front-loading (Levinson, 2013). This, however, is not a universal feature of all languages. There are languages that do not relocate the question words at the beginning and use them *in situ.* In other words, the question word takes place of the missing information it is inquiring about (e.g., statement: I go to the *store*. Question – You go *where*? *Store* – focus of inquiry). In English, for example, it is acceptable to have both, front-loaded and *in-situ* questions. Interestingly, however, Levinson (2013) also highlights that in colloquial interactions speakers tend to rephrase the sentences in such way that question words are fronted also in some languages (e.g., Japanese) that according to formal grammatical rules should leave the question words *in situ*. These qualitative insights suggest that front-loading of question words might be helpful in predicting incoming questions. Surprisingly, though, there is no quantitative research investigating whether this feature actually helps in question recognition.

Slonimska & Roberts (in prep.) were the first to quantitatively assess whether question words, also called *wh-words*, are plausible candidates as a cue to content question recognition. They argue that for wh-words to be able to help in predicting a question, they should be systematically similar. In other words, if question words tend to sound similar, it makes easier for the addressee to predict a question, given that in such way a specific phoneme would be associated with a specific pragmatic function – signaling about an incoming question.

Even though there is some qualitative research arguing that there is no systematicity of wh-words within a language (Cysouw, 2004), Slonimska & Roberts (in prep.) show that there is a statistical tendency for wh-words to sound similarities within languages. They analyzed 172 languages from 65 language families and from 18 different geographic areas. They show that matching first phoneme of the wh-words (within languages) is an occurrence above chance. They also show that similarity of the first phoneme of the question words is higher than for random and conceptually related words. Moreover, an analysis shows that question words are more detectable than other words (i.e., the first phonemes of wh-words are less likely to be found in other words). Thus, this indicates that there could be viable phonetic cues to questions. Finally, they show that there is a tendency for the first phonemes of question words to match more in languages that use front-loading of the question words in comparison to languages that do not.

Importantly, Slonimska & Roberts (in prep.) control their findings for historical contact. Namely, they control whether there is influence on the results based on the language family and/or area. While they do find the effect, the similarity of question words within languages still stays significant independently from these factors. Accordingly, Slonimska & Roberts (in prep.) conclude that the fact that question words tend to have matching first phonemes is not due to chance or historical factors. Instead, it is possible to argue that this phenomenon constitutes a property of cultural evolution that is selected for due to its benefit in interaction – i.e., rapid question recognition.

This study, however, was purely observational (i.e., based on word lists). The current project seeks to find experimental evidence for these observations.

Conversation, however, is a continuous stream of information. It is built on sequences (Sacks, Schegloff & Jefferson 1974) and thus the cues that are available in the question itself might actually be preceded by cues that come from the context in which conversation occurs. For example, Gisladottir, Chwilla, & Levinson (2015) show that people can recognize the type of a speech act at an early stage if it occurs in highly constraining context. In other words, they find neurological evidence for participants recognizing the speech acts early in the turn if they form an *adjacency pair* (Sacks, Schegloff & Jefferson, 1974) like an answer to a question or offer to a request. On the other hand, participants use the entire utterance to recognize less "sequence dependent" speech acts like pre-offers. Based on these findings Gisladottir, Chwilla, & Levinson (2015) argue that context of a previous turn helps speakers to project an incoming turn. Accordingly, given that *question-answer* can be considered a prototypical adjacency pair (Enfield et al., 2010), it seems logical to assume that context, or in other words the previous turn, plays an important role in question prediction. To be more specific, the sequential type of the previous turn should have an effect on question prediction. An initiating turn (e.g., a question) requires a responding action, while a non-initiating turn does not. As such, non-initiating turns should be better predictors of a question than initiating turns.

To summarize, there is extensive research on how various paralinguistic and supra-segmental cues contribute to question recognition. In contrast, there is almost no research investigating how and whether this can be achieved with phonemic cues as well. Moreover, it is not clear how and whether context modulates the effectiveness of such cues.

3. The present study

Based on the reviewed literature we argue that people recognize incoming turns as questions based on the first phoneme of the incoming turn and the sequential type of the previous turn. However, there are no previous studies on which we could base these predictions. Thus, we are interested in exploring whether there is evidence for a phonetic/sequential cues in natural conversation (corpus study) and consecutively test whether people actually use these cues to predict upcoming turns (experimental study).

In the present study we aim to fill the gap in regard to whether the systematicity of the first phoneme of the question words contributes to (content) question prediction. Based on the findings of Slonimska & Roberts (in prep.), our first hypothesis is as follows:

- People are more likely to think that an incoming turn is a question if it starts with the first phoneme of a *wh-word*.

Given that conversation always occurs in context, we also aim to provide first insights on how this impacts the prediction of the turn being a question. Considering that *question-answer* is a highly restricting adjacency pair, we expect that if a turn is preceded with a question, people will be less likely to think that an incoming turn is a question, considering that an answer to a question should be expected. Accordingly, the second hypothesis is as follows:

- People are less likely to think that an incoming turn is a question if it is preceded by another question.

Accordingly, if an incoming turn starts with the first phoneme of the wh-words and the previous turn is not a question, people would be more inclined to think that an incoming turn is a question than in any other combination, considering that both factors suggest that it could be the case. Thus, the third hypothesis is:

- There is an interaction between phoneme and context in question prediction: people are more likely to think that a turn is a question if it starts with the first phoneme of *wh-words* and is not preceded by a question.

To assess whether we can gain support for our hypotheses, we first carry out an exploratory corpus analysis of naturalistic and therefore ecologically valid data –

i.e., spoken conversations. We address this by means of the method of *binary decision trees*, also known as *recursive partitioning* (Strobl, Malley, and Tutz, 2009).

Roberts et al. (2015) suggest that it is possible to use insights from a binary decision tree to generate predictions that can be consecutively tested in an experimental setting. What is more, it is also possible to use real conversational data to create stimuli for controlled testing of these predictions (e.g., De Ruiter et al., 2006, Bögels & Torreira, 2015). Thus, we first assess our predictions by comparing them with the predictions produced by a decision tree. We then use the findings to inform the design of the experiment and use the same corpus to construct the stimuli for this experiment

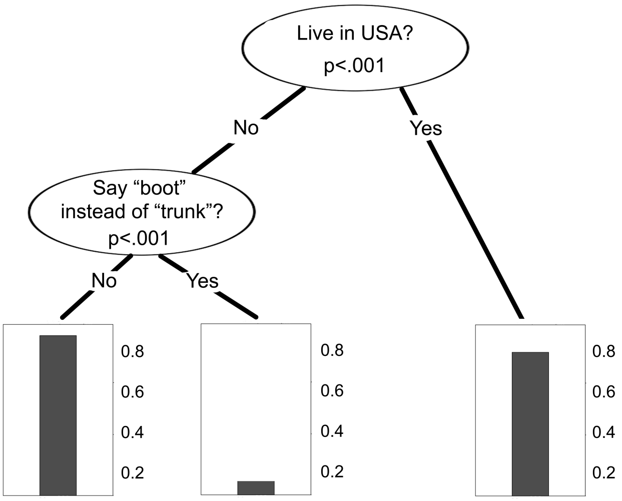
Such approach gives more saturated understanding of the phenomena under investigation as it is based on generating hypotheses from the data in "the wild" (i.e., in the corpus), testing them experimentally, and then referring back to "the wild" in order to draw conclusions about similarities/differences of the results from both approaches. Thus, we start with the corpus study. In next section, we describe the method of the decision trees, the data we used for the analysis, and interpret the results.

3.1. Corpus study

3.1.1. Method

A binary decision tree can be roughly compared to a simple cognitive model of a rational agent trying to decide the order in which to ask a series of yes/no questions in order to make the best decision (see Roberts et al., 2015). For a hypothetical example, let's imagine that an agent tries to predict whether someone is American versus British and it has information on whether they use "boot" instead of "trunk" when they speak and whether they live in Great Britain or USA (see Fig.2). According to the decision tree, for an agent the best choice would be to first ask: Do they live in the USA? We see that there is 80% chance for a person to be American if they live in USA (and not Great Britain). If they do not live in the USA (thus, they live in Great Britain), the agent should further ask the following question: Did they say "boot" instead of "trunk? If so, there is only 10% chance that they are American (accordingly, there is 90% chance that they are British), if they did not say "boot" – 90%.

Figure 2. A mock example of the decision tree for guessing whether someone is American versus British. The bars indicate the proportion of Americans.



In the current study we are interested in whether the first phoneme of the turn (first predictor) and context of the previous turn (second predictor) would help in recognizing an incoming turn as a content question (outcome variable). Namely, we are interested in whether data would be clustered in such way that specific first phoneme (/w/, /h/ versus other phonemes) of the current turn and specific type of previous turn (non-initiating turn versus initiating turn) would lead us to increasing the probability of the turn being a question (proportion of outcome variable in a cluster). Thus, if the first phoneme and context make a difference in a decision making, we expect that the best guess of the turn being a question will be made based on rational agent choosing the first phoneme being /w/ or /h/ phoneme and previous turn being non-initiating turn.

Importantly, decision trees also allow assessing the effect of each predictor. Namely, data is first clustered based on the strongest predictor (e.g., the country in previous example), then, in each branch, the predictors are re-evaluated anew and split again based on the strongest predictor in the branch until the splits no longer produce significant differences in the two clusters. In other words, at the top of the graph (i.e., the first split) we see the most important predictor and if some predictors are not present in a decision tree it implies that they do not have an effect on the outcome variable. Thus, by using the method of binary decision trees we can assess whether both of the variables of interest in our study have an effect on outcome variable and also we can assess which predictor is stronger.

As such, the method of decision trees does not test hypotheses but serves an exploratory purpose in order to generate them. For the current study we do, however, have hypotheses. Given that the method of the decision trees is blind to those, we explore the existing corpus of natural conversation and see whether the decision tree generates comparable hypotheses to those of our study. In turn, if we do not find support for our hypotheses by assessing the decision tree we can still investigate how the data is clustered and make informed decisions in order to adjust initial hypotheses accordingly.

3.1.2. Materials and design

We used the Switchboard corpus (Godfrey et al., 1992; Calhoun et al., 2010) that consists of telephone conversations in American English. In these telephone conversations speakers (strangers to each other) talk about random topics like work, vacations, politics etc. Godfrey et al (1992) and Calhoun et al. (2010) transcribed and annotated these conversations in detail, also providing information on properties of the turns of the speakers. They also annotated the turns in regard to their dialog acts. These dialog acts, consist of speech acts, but also they include information on backchannels, laughter, etc. Thus, the annotation of the corpus is well suited for the current analysis. In addition to this annotation, we also use annotation specifying the sequence organization type and sequential turns of the dialog acts used in Roberts et al. (2015).

The data was prepared for the analysis in R and later analyzed by means of the package “party” (Hothorn, Hornik & Zeileis, 2006). First of all, we disregarded data from the first 5 seconds of all conversations. This was done with consideration that the beginning of the conversation always consisted of the introduction of the speakers – including greetings and general questions (e.g., *What is your name?).* We chose to disregard this part of the data considering that these sequences can be considered ritualized (Schegloff, 1979) and thus could potentially confound the findings in regard to the predictors under investigation. Also, we excluded all overlapping turns in order to ensure that both turns are clearly perceivable.

We used the annotation of Switchboard in the following way to extract the target speech acts and their preceding speech acts from the other speaker’s turn: each observation consisted of a transition between two turns between speaker A and speaker B. We used the first speech act of B’s turn (turn types are based on the dialogue act categories from Switchboard) for the target turn. We specified the outcome variable – *question* – according to whether B's turn was a question (content/open question) or not. We used the last speech act of A's turn for the previous turn. For this turn we created a predictor variable *context* specifying whether this turn was initiating or non-initiating (see Roberts et al., 2015 for dialog act categories according to their sequence organization type).

We assumed that fillers (e.g., *hmm, uh*) at the beginning of the turns do not contribute to the content of the incoming turn and recognition of the speech act. Thus, we excluded following fillers from the B's turn (from the current turn): *ahm*, *er*, *ah*, *hmm*, *oh*, *uh*, *aa, um, ow.* Then, the first phoneme from B's turns was extracted to create the predictor variable *phoneme*. This variable consisted of 34 unique phonemes (coded according to the symbols used by Switchboard): /aa/, /ae/, /ah/, /ao/, /aw/, /ax/, /ay/, /b/, /ch/, /d/, /dh/, /eh/, /er/, /ey/, /f/, /g/, /hh/, /ih/, /iy/, /jh/, /k/, /l/, /m/, /n/, /ow/, /p/, /r/, /s/, /sh/, /t/, /th/, /v/, /w/[1], /y/. Finally, we excluded all turns for which B's turn was a backchannel, considering that backchannel serves monitoring rather than informing function - they often appear in overlap and do not always need to be identified in the same was as other speech acts.

In the final data used in the decision tree we had 9185 turns in total out of which 221 turns were content or open questions (see Table 1). Out of all turns, 5052 were initiating and 1456 were non-initiating turns. Thus, it is clear that initiating turns are more common than not-initiating turns in our data set and logically questions are much less frequent in comparison to all the other speech acts combined together.

Table 1. Distribution of the data according to the previous turn being initiating or not initiating and whether the current turn is a question or not.

		A's turn: Previous turn		
		Initiating	*Non-initiating*	***Total***
B's turn: Current turn	*Not a content/open question*	4836	1451	**6287**
	A content/open question	216	5	**221**
	Total	**5052**	**1456**	**9185**

[1] The accents in this corpus do not have aspirated and un-aspirated allophones of /w/.

In total, there were 7830 current turns that started with a phoneme other than /w/ or /h/ (see Table 2). There were 1358 turns that started with /w/ or /h/.

Table 2. Distribution of the data according to the first phoneme of the current turn.

		First phoneme of B's turn		
		Other	*w/h*	***Total***
B's turn: Current turn	*Not a content/open question*	7703	1216	**8919**
	A content/open question	127	139	**266**
	Total	**7830**	**1355**	**9185**

For the analysis we had 2 predictor variables: *context* from the previous turn (initiating or non-initiating) and first *phoneme* of the current turn (34 unique phonemes). The outcome variable was whether the current turn was a content/open *question*. Accordingly, if none of the cues has an influence on the outcome variable, decision tree should not split the data at all and keep it as a single partition. On the other hand, if the cues are extremely strong, then the data should be divided perfectly into questions versus non-questions.

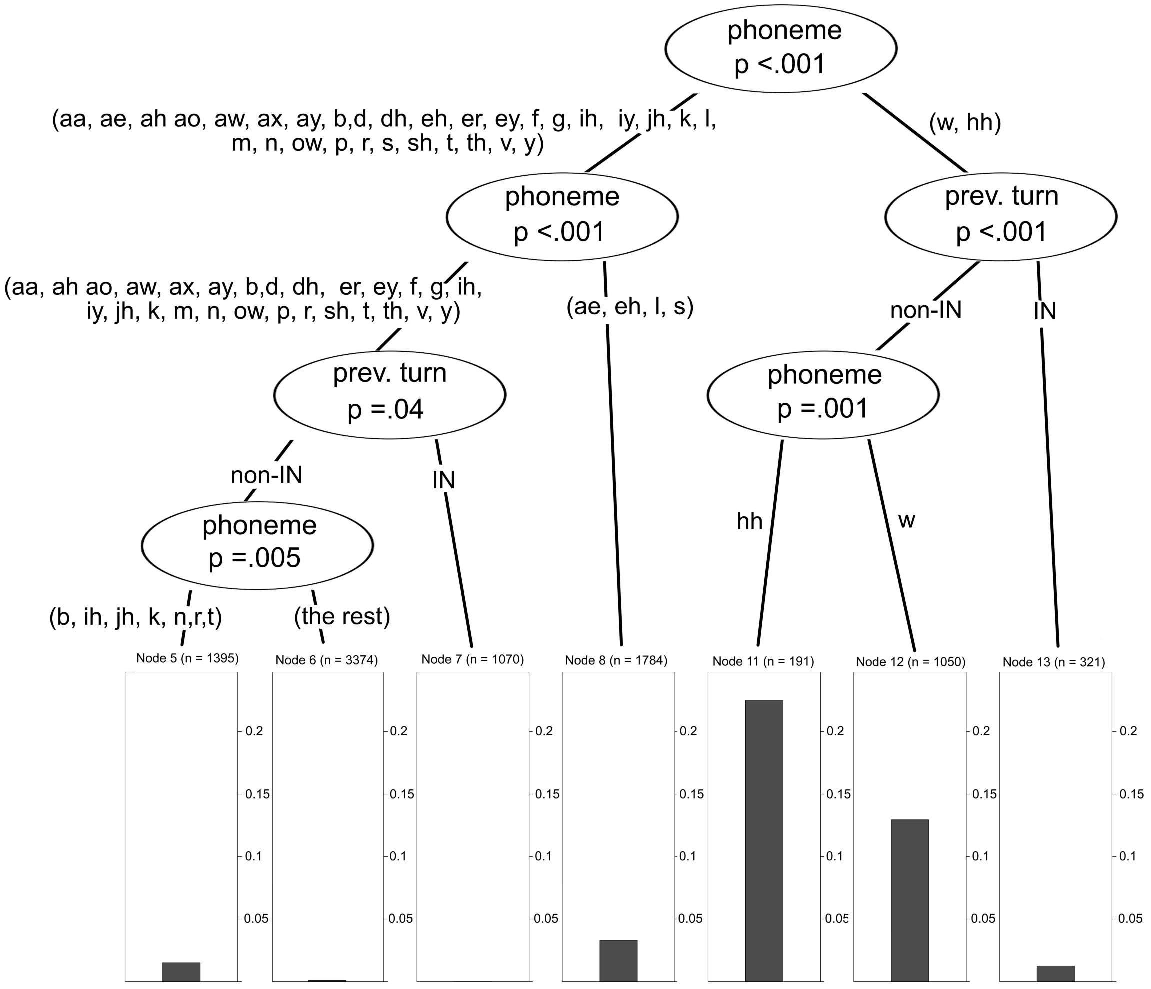
3.1.3. Results

The decision tree divides data at each node of the tree starting from the top of the figure. Leaves of the tree at the bottom of the figure show a proportion of turn being a question (i.e., *question* turns)(see Fig.3).

As noted above, there are more turns that are not questions in the data (6287 turns versus 221 turns). Accordingly, it is more likely overall that an incoming turn is not a question. Thus, the proportions of questions in each leaf of the decision tree provide an insight of how predictors that are included in the decision tree augment the probability of a turn being a question in a specific subset.

The decision tree splits data first based on the first phoneme of the turn. The exact division of the phonemes is as follows: /w/ and /hh/ versus all the other phonemes. Thus, the decision tree, which is absolutely blind to our predictions, splits the data exactly in line with these predictions. Note that larger proportions of question turns in the leaves of the tree are found on the right (i.e., node 11, node 12, node 13 in comparison to node 5, node 6, node 7, node 8) - under the data that is clustered according to the phoneme being /w/ or /hh/.

Figure 3. The decision tree of question turns split according to the sequential type of the previous turn and the first phoneme of the current turn.



We, first follow the node that clusters the data on the right (/w/, /hh/). The data is further clustered according to the type of the previous turn. If previous turn was an initiating turn (i.e., initiating) proportion of question turns is considerably lower than if previous turn was not an initiating turn (i.e., node 13 versus node 11 and node 12). Also, if previous turn is not initiating, the data is further split into whether the phoneme of the current turn is /hh/ or /w/. Note, that proportion of questions is higher in /hh/ (22%) leaf than in /w/ (13%). This can be explained by the fact, that a word *well*, which often is used as a filler, often occurs at the beginning of a turn and thus decreases the overall proportion of questions in /w/ leaf. Moreover, there are more turns overall that start with /w/ than with /h/. Thus, the proportion in /w/ leaf is also lower because the total number of turns is much higher than in /hh/ leaf.

In regard to the data clusters on the right (turns starting with phonemes other than /w/ and /hh/), it is evident that proportion of question turns is extremely low in all leaves of the tree. However, there is a larger probability of turn being a question if it starts with /ae/, /eh/, /l/, or /s/. This cluster most probably is due to words like *and* (e.g., *and how old the youngest?*), *anyway* (e.g., *anyway so where your favorite place to go?), like* (e.g., *like what*?) and *so* (e.g., *so which one are we gonna throw out?*). Note that often the next word tends to be exactly one of the content question words and, thus, these words most probably are used as fillers before the question.

3.1.4. Summary

Overall, the analysis confirmed our initial hypotheses. Namely, the analysis showed that there are phonetic cues to questions in the data - if the incoming turn starts with /hh/ or /w/ it is more likely that this turn is a question than if it started with a different phoneme. Thus, we find first support for phonemic cue in question recognition as argued by Slonimska & Roberts (in prep). Not only there is systematicity above chance for question words (Slonimska & Roberts, in prep.), but also this systematicity is a likely predictor of question in an incoming turn in English.

We also found confirmation that the turn is more likely to be a question if it was preceded by a non-initiating turn as opposed to initiating turn. What is more, based on the analysis we can also expect that recognition of a turn being a question will be boosted if both cues converge on a possibility of an incoming question (nodes 11 and 12) – namely, if an incoming turn starts with /w/ or /hh/ and previous turn is non-initiating. Thus, we could expect an interaction of context and phoneme – namely, that effect of phoneme will be stronger when the previous turn is an initiating action in comparison to the effect of phoneme in the context of non-initiating turn.

We first proposed to view a decision tree as a simple cognitive model of a rational agent. Accordingly, for an agent to predict whether the next turn is a question they should consider following facts: if the first phoneme of the incoming turn is /w/ or /h/ and if this incoming turn is preceded by a non-initiating action there is a larger probability that the incoming turn is a question (13% for /w/ and 22% for /hh/) than if the turn is preceded by an initiating action (1%) or if it starts with a phoneme other than /w/ or /hh/ (below 3%). Accordingly, the analysis suggests that phonemic cues are used in context. Thus, both predictors should be taken into account when assessing their efficacy on question prediction. Conversation is always a context-dependent phenomenon. Thus, exploring the effect of phonemic cue in isolation might be under-representing its actual strength in question recognition. Put differently, assessing the systematicity or lack thereof of wh-words might be actually only representing the surface of the potential effect of the phoneme. Its entire value appears to be evident exactly in conversational context. Accordingly, in the experimental design both factors should be included in order to explore how context and first phoneme influence question recognition and how these factors modulate the effect.

Based on these findings we make following predictions for an experimental testing in regard to question recognition:

- Participants will be more likely to think that a turn is a question if it starts with the first phoneme of the *wh-words* in comparison to other phonemes.
- Participants will be more likely to think that a turn is a question if it is preceded with a non-initiating turn in comparison to initiating turn.
- There will be an interaction between phoneme and context: participants will be more likely to think that a turn is a question if it starts with the first phoneme of *wh-words* in a non-initiating context.

3.2. Experimental study

3.2.1 Method

Participants

For the experiment 25 participants (14 male. 11 female) were recruited. Participants' age ranged from 21 – 70 years (M = 32, SD = 11). All participants were native speakers of English but had various (double) nationalities (e.g.. American, British, Canadian, Australian, Indian, Latvian). Thus, the participants spoke different dialects of English, which we divided into 3 main groups – American English, British English and Other. All participants had no hearing impairments. Nine participants were raised bilingual with English being their dominant language. Participants were paid 6 Euros for participation.

Materials and design

In this experiment participants listened to series of audio samples. Each sample consisted of a *context* (initiating versus non-initiating) produced by the first speaker and a *response*[2] produced by the second speaker. The response could be either the first phoneme of *wh-words* (i.e.. /w/ or /h/), a single phoneme other than /w/ or /h/, or no response (no audio from the second speaker).

We used the recordings from Switchboard corpus (Godfrey et al., 1992; Calhoun et al., 2010) analyzed in the corpus study to construct the samples. Each sample consisted of two turns that were taken from the same dialog. Thus, the first turn always came from one speaker in a conversation, but the second turn came from the other speaker in the same conversation (except for 2 items where we could not extract necessary second turns. In this case for the second turn we used an audio from a different conversation). This secured that background noise was kept constant across all samples in the same set. Turns were extracted by means of the software Praat (Boersma & Weenink. 2014).

The first turn of the sample constituted the first factor – *context* – with two levels: initiating and non-initiating. For the context with an initiating first turn we used yes/no questions and wh-questions; for the context with non-initiating first turn we used statements (see Table 3). The number of words in the first turn ranged from 3 to 25 in non- initiating turns and 4 to 33 in initiating turns. Independent t-test showed that number of words was comparable in both conditions (t(24)=0.87, *p*=.392).

The second turn of the sample (i.e., the response produced by the second speaker) constituted the second factor – *phoneme* – with 3 levels: *wh* (phonemes /w/ or /h/), *other* than in level *wh*, and *none*. For the second turn in the level *wh* audio was clipped to contain the first phoneme together with the beginning of the subsequent phoneme of turns that started with phoneme /w/ or /h/[3] – the critical level. Importantly, from each conversation 2 types of phonemes were extracted – from speech acts that were content questions and from speech acts that were not questions (e.g., statements starting with *well, we*). Thus, we could be able to assess whether the effect of other question cues (e.g., raised pitch at the beginning of the question word) contribute in question prediction.

[2] In order to reduce confusion in the text by “response” we refer to the response of the second speaker in the audio sample. We use the term “answer” to refer to the answers given by participants ("question"/"Not a question") in experiment.

[3] From now on we refer to phonemes /w/ and /h/ as *wh* phonemes.

Table 3. Example of two sets of samples - in each set there are 10 samples consisting of 2 types of first turn (initiating/non - initiating) and 5 types of second turn.

	First turn		Second turn				
			/W/		**OTHER**		**NONE**
SET			/w/ from quest.	/w/ not from quest.	not /w/ from quest.	not /w/ from non-quest.	no second turn
a	Non-initiating	***I do enjoy playing***	***Wh****[at your handicap]*	***W****[ell I wish that's all we had]*	***D****[o you have long waits uh to get on the course]*	***Q****[uite a while ago it's probbaly up to 20 now if I]*	-
	Initiating	***That's some cold golf too isn't it***	***Wh****[at your handicap]*	***W****[ell I wish that's all we had]*	***D****[o you have long waits uh to get on the course]*	***Q****[uite a while ago it's probbaly up to 20 now if I]*	-
b	Non-initiating	***I don't think uh hardly anybody lives there***	***Wh****[at is it]*	***W****[e went to california this last year]*	***Pr****[obably a city in itself kind of like huh]*	***M****[ost most of land is pretty borwn]*	-
	Initiating	***Oh where is that***	***Wh****[at is it]*	***W****[e went to california this last year]*	***Pr****[obably a city in itself kind of like huh]*	***M****[ost most of land is pretty borwn]*	-

For the second turn in level *other* we extracted the segments from turns that did not start with the *wh* phonemes. Also, in this level we extracted phonemes from two different types of speech acts – phoneme other than in *wh* from questions and non-questions. This made it possible to account for other possible cues available in the sample in predicting the turn as a question. Accordingly, dividing these two levels, *wh* and *other*, in sub-levels according to whether the phoneme came from actual question or not, we could have a clear-cut understanding of how the phoneme, and not the other cues, contributes to question prediction.

We used the software Praat to concatenate each first turn with each second turn (e.g., (first turn: statement) + (second turn: /w/ from wh-question)). Subsequently, each turn pair was processed in the software Audacity (Mazzoni & Dannenberg, 2000) by adjusting a gap between the turns, so that the gap between first and second turn was 250ms. This was done with consideration that differences in length of the gap might influence answers of the participants (see Roberts & Francis, 2013; Kendrick & Torreira, 2015; Roberts, Torreira & Levinson, 2015; Stivers et al., 2009), thus it was kept constant across all trials. Stivers et al. (2009) show that average gap between turns, also polar questions, is 200ms. Thus, we chose to have a slightly longer gap considering that we were interested in content, thus more cognitively demanding, questions, and to ensure that participants can differentiate between the end of the first turn and beginning of the second turn.

This resulted in a set of 8 audio samples - 4 samples started with a statement and every phoneme as a beginning of a second turn and 4 samples started with an initiating turn and the same 4 phonemes as for the statements as a second turn.

Finally, for the second factor – phoneme – a general control level *none* was added in which the second turn was absent. Thus, one sample in a set contained only the first turn with initiating context and one item contained first turn with non-initiating context. This control level provided a baseline in regard to the added efficacy in question prediction of hearing the first phoneme of the second turn. In other words, for these samples the decision regarding the type of the next turn could be made purely on the basis of the first turn. Thus, the final set consisted of 10 samples.

We created 25 sets in total, resulting in 250 unique audio samples – this was a fully crossed design. There were 50 unique first turns out of which 25 were initiating and 25 were not initiating. Each of these first turns was paired with a unique phoneme across all sets but that repeated twice within the same set - once with an initiating first turn and once with non-initiating turn of the same set. In total there were 25 unique phonemes for each sub-level of factor - *phoneme* (Level *wh*: 24 different variants of phoneme /w/ and 1 phoneme /h/ extracted from real questions, 25 different variants of phoneme /w/ extracted from speech acts that were not questions; *Level* other: 25 different phonemes than in level *wh* extracted from real questions, 25 different phonemes than in level *wh* extracted from non-questions).

These 250 items were divided in 5 blocks so that in each block first and second turns occurred only once (i.e., participants never heard the same first turn or second turn more than once). Each block was randomly administered to one-fifth of the participants. Each block contained 50 samples with equal number of trials across sets and conditions (25items from each context level– initiating and non-initiating first turn, 10 items from each *phoneme* (sub)level – 10 *wh* phonemes from question and 10 from non-question. 10 non *wh* phonemes from question and 10 from non-question. and 10 samples without a second turn). Each block contained 2 items from the same set – initiating and non-initiating first turn for which second turns varied.

Procedure

Participants were tested in Nijmegen, the Netherlands and Riga, Latvia. Even though, location differed in regard to where participants were tested, all participants were seated in a quiet room in front of a computer and used headphones to listen to the audio samples. The experiment was presented via the online software *Qualtrics* (Snow & Mann. 2010). First, participants read general description of the experiment (see Appendix I) and pressed a button for consent of usage of their data. Subsequently, they filled out a questionnaire about their age, nationality, native language and knowledge of other languages. Then, participants were informed that they would listen to short fragments of dialogues in which they heard what the first person says and also the beginning of what the second person says. They were also instructed that sometimes they would not hear anything from the second speaker. Their task, as written in the instructions, was to determine whether the second person would ask a question or not by means of completing a sentence *"The Second turn is ____"* on the screen by pressing one of the buttons on the screen below the sentence: *not a question* or *a question*.

Then, 2 test trials followed ensuring that participants understood the task. One test trial consisted of an item that had both turns and one of the items consisted of the first turn only. The difference in one item having a second turn and other not having a second turn was explicitly mentioned. Thus, participants were familiarized with two different types of dialogues that they might hear – one where they hear the beginning of the speech of the second person and one where they hear only the first person. Also, participants were encouraged to ask experimenter for elaboration if they were not sure about the task.

Given that the main objective of the study was to concentrate on the response of the participants in regard to what they heard and not on the timing of their response we chose to allow participants to listen to the fragments twice, ensuring their understood the short fragment. They were instructed, however, to do so only if they have not understood the speech. Thus, any data on reaction times would not be informative for this task and they were not recorded. Moreover, given that participants never heard what followed after the first syllable of the second speaker, reaction time could not indicate the exact moment when decision was made, naturally as participants were instructed to listen to the whole fragment from start to end and only then make a decision. Once the participants have completed the test trials and pressed a button confirming that they have understood the task, the experiment started.

There were 50 experimental trials presented auditorily through headphones. The order of the trials was randomized for each participant. Participants would click on the play icon to listen to the trial. Afterwards they would indicate whether second turn they heard was a question or was not a question. Once they have made a decision, they would press an arrow that would lead them to the next trial that appeared on a new screen.

Analysis

We analyze the data in R by using package lme4 (Bates, Maechler, Ben Bolker & Walker, 2015). We use the method of linear mixed models to test the effect of context and first phoneme on prediction whether the second turn of the dialog is or is not a question. We chose to use linear mixed models in order to be able to account for individual differences of both participants and experimental items. By using linear mixed models we could examine not only the fixed effects of context and the first phoneme, but also include random effects of the stimuli samples by accounting for variability in context samples and phoneme samples. More so, linear-mixed models allow modeling not only random intercepts but also random slopes and thus accounting for even more fine-grained individual variation that might have influence on the outcome of the analyses.

We assumed that following random effects should be included in the model: context sample and response (phoneme) sample. Given that the audio samples used in the experiment were not exhaustive, or in other words they were meant to represent (and not cover completely) all possible samples of the conditions, we had to account for their individual differences. It would be impossible to include all samples of initiating and not initiating context. As well it would be impossible to include all possible variants of the first phoneme of the response. Thus, we considered both context and response samples as random effects in order to account for their individual differences and be able to generalize to other samples.

Another way to view the use of random effects is that if we include a random intercept (i.e., random effect) for the context sample we account for variability that some samples from the context (1st turn) are generally more powerful in eliciting "question" responses from the participants than others. For example, this might be due to the semantic content of the turn or some other aspect besides the type of sequence organization that we are interested in. The same can be said about the random intercept for the phoneme sample – we control for the specific sample in the second turn having a generally larger effect on the participant's response or, in other words, not due to the phoneme itself but due to sample's individual properties.

It is also possible that the effect of context and/or phoneme is stronger for some participants and not for the other participants. Thus, in order to account for this aspect we chose to include random slopes of context and phoneme by participant. Accordingly, the individual differences of participants in regard to how sensitive they were to one of or both predictors were also considered. Furthermore, we run series of models to account for possible confounding factors, e.g., trials, strategies of participants in answering to samples, age, gender and type of English spoken by participants. The significance is derived from model comparisons. The general procedure of assessing whether there is an effect of a factor on the outcome variable is by comparing a baseline model to a model to which factor is added. If there is no difference between baseline model and the model with factor included, this indicates that it does not have an effect. This can be repeated continuously by accounting for various confounding factors and subsequently comparing the factors of interest to the baseline model that includes random effects and confounding factors.

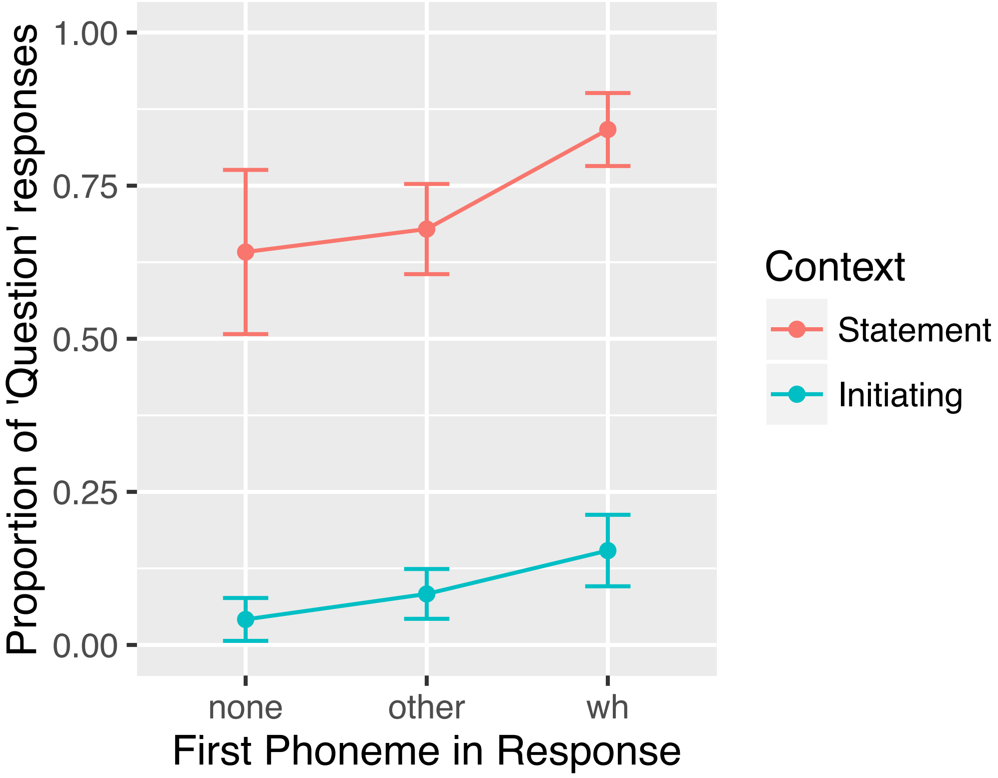
3.2.2. Results

In the present experiment we tested whether participants predict that an incoming turn is a question based on two factors - the first *phoneme* of the incoming turn (*wh* phonemes versus other phonemes or none) and the *context* of the previous turn (initiating versus non-initiating).

We excluded 1 participant from the analysis due to the fact that they took 3 times longer to complete the experiment than other participants (38 minutes compared to average of 12 minutes). Thus, we assumed that either this participant did not understand the task or this participant was listening to the audio samples more than twice. The results are not influenced if the data points from this participant are kept in the analyses. However, in order to be conservative, we report the results with this participant excluded. Accordingly, the final analysis is based on 24 participants.

The results section is divided as follows: first, the random effects are reviewed and the baseline model defined. Next, the design of the study is reviewed by controlling for possible confounding factors. Finally, we assess the impact of the key factors *context* and *phoneme* on prediction of a question in an incoming turn (for the full summary of the results, see Appendix II). It appears that there is a large effect of context, possible effect of phoneme and a trend for an interaction (see Fig.4).

Figure 4. Raw proportions of participants answering that an incoming turn is a question based on the previous context and the first phoneme of the incoming turn. Error bars indicate 95% CI of observations grouped within participants.



Assessment of the random effects

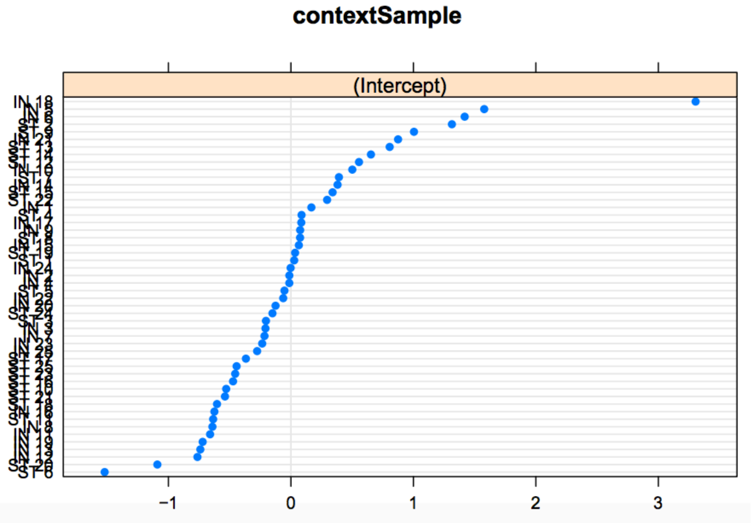
We first run series of models to examine the impact of random effects. The baseline model included the random effect by subject only. Analysis revealed that the best fit of model was when random effects of context sample, response sample and participant, and a random slope for context and phoneme by subject were included ($\chi^2(7) = 19.39$).

Accordingly, the baseline model for the main analysis included random effects of context and response sample, random effect of participant, and random slopes for context and phoneme by participant. In next section we control for possible confounding factors by comparing this model to models with these factors included. Finally, in subsequent section we compare this model to the models with fixed effects of interest (i.e.. context and phoneme) included. Before the main analysis, we controlled whether the design of the study was reliable.

Individual differences by items

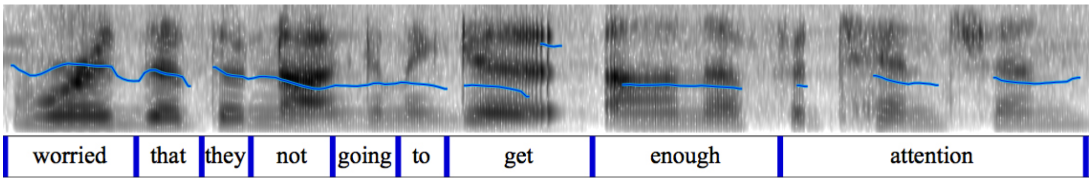
We first examined the individual differences of the context samples (see Fig.5). It is evident that samples are treated quite differently. We also looked for the outliers. It appears that initiating context from set 18 was treated differently than other samples. Namely, participants were more likely to answer that a turn was a question if it was preceded by this context sample (i.e.. yes/no question: *worried that they're not going to get enough attention)*.

Figure 5. Individual differences of the context samples in regard to eliciting an answer "question". The x axis represents the model estimate in the logit probability scale.



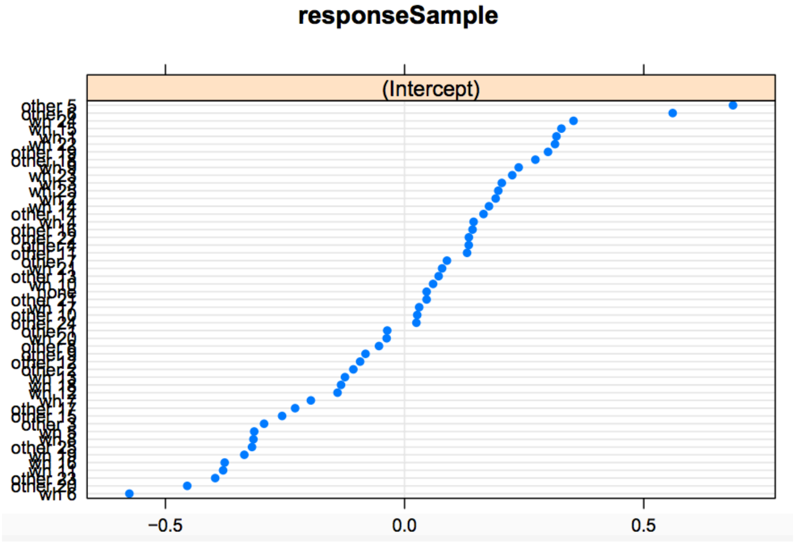
When we examined the item, we found that the intonation of the speaker was not rising considerably at any point of the turn (see Fig.6). Thus, it was likely to be perceived as a statement. Even though this context sample was an outlier, we chose to keep it in the analysis in order to keep fully balanced design of the study. Given the properties of the method of the liner mixed models, the analysis is adjusted in regard to the individual differences if random effect of *context sample* is included. Thus the fact that there are individual differences of context samples (including the outlier) can be accounted for.

Figure 6. The spectrogram and intonation contour of the initiating context sample from the set 18.



We also examined the individual differences of the samples of the first phoneme of the response (see Fig.7). There were no considerable deviations, but two samples were treated slightly differently than others. Namely, these samples elicited more "question" responses. However, note that range of overall variation is quite narrow. Considering that the baseline model included random effect of *phoneme sample* we could be certain that these minor differences are accounted for and thus do not confound the results.

Figure 7. Individual differences of the phoneme samples in regard to eliciting an answer "question". The x axis represents the model estimate in the logit probability scale.

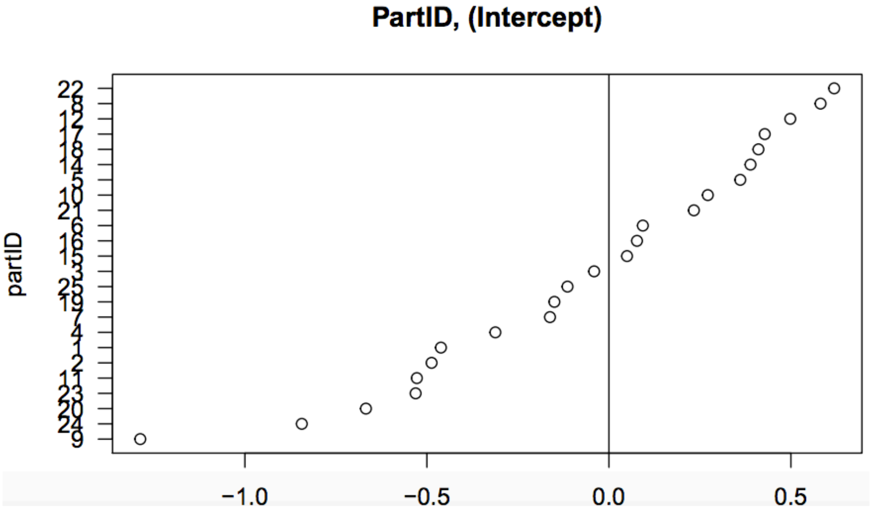


Individual differences by subjects

It is plausible that some effects were stronger for some participants than others. Thus, it is important that we also adjusted the intercept according to these differences. This was done by means of random slopes for context and phoneme by participant.

We found that there were some individual differences in regard to how participants tended to answer to the experimental samples (see Fig.8). Namely. there were some participants that tended to answer "question" more on a general level and there were some participant that tended to answer "not a question" more on a general level. Importantly, it appears that one participant (i.e., partID - 9) was more likely to answer “not a question” than all the other participants. Thus, the decision to include random effect of participant is valid in order to account for the individual differences (i.e.. sensitivity to these factors) of the participants (including the outlier). To be more conservative we also add random slopes of context and phoneme by participant to account for the variability in sensitivity to these factors.

Figure 8. The individual differences of participants in regard to answering that a turn is a question. The x axis represents the model estimate in the logit probability scale.



Possible confounding factors

We compared the baseline model (containing random effects of context and response samples, random effect of participant and random slopes for context and phoneme by participant) to possible confounding factors: trial number, question block, previous answer of the participant, sex of the speakers in the audio samples, type of English spoken by participants, age and sex of the participants.

We found an effect of trial[4] ($\chi^2(1) = 4.80$, $p = .03$) and an interaction of trial and context ($\chi^2(1) = 12.81$, $p < .001$). Participants were more likely to answer that a turn is a question in later trials and this effect was larger for non-initiating context. We address this finding in a discussion section. Considering that the effect of trial was significant. the factor *trial* and the interaction of trial and context were included as fixed effect in the baseline model. In other words, the effect due to trial number was accounted for when assessing other effects.

[4] We recentred the intercept of the trial so that it would reflect the differences in the middle of the experiment.

There was no effect of the question block administered to the participants ($\chi^2(1)$ = 1.13, *p* = 0.29). Thus, none of the blocks contained samples that were "easier" or "more difficult" in predicting an incoming question. There was no effect of previous answer of the participant ($\chi^2(1)$ = 1.73, *p* = .19). This indicates that participants did not develop any specific strategies to respond to the experimental items and we can assume that their answers were genuine. There was no effect of the sex of the speakers - nor in the context ($\chi^2(1)$ = 1.53, *p* = .22) nor in the response (χ^2 (1)= 0.02, *p* = .89) samples. Thus, the answers of the participants were not biased in this regard. In regard to participants, there was no effect of type of English spoken ($\chi^2(2)$ = 2.09, p = 0.35), age ($\chi^2(1)$ = 0.81 . p = 0.37) nor sex ($\chi^2(1)$ = 0.02. *p* = .89) of the participants.

Assessment of the predictors - context and phoneme

A linear mixed model was fit to assess the effect of *context* and *phoneme* on participants' answers in regard to whether an incoming turn was a question, which was a binary decision (i.e.. second turn IS or IS NOT a question). The predictor variables were *context* (*initiating/non-initiating)* and *phoneme* (*wh, other, none*). These predictors were coded as fixed effects and compared to a baseline model (described above), which included fixed effect of *trial*, random effect of *context sample* and *phoneme sample*, random effect of *participant* and random slopes for context and phoneme by participant.

Table 4. Summary of the best-fit model in a logit scale in regard to prediction of an incoming turn as a question.

		95% CI				
	Estimate	**Lower b.**	**Upper b.**	**SE**	***z* value**	***p* value (Wald-*z*)**
(Intercept)	2.14	1.43	2.85	0.36	5.91	>.001
TrialNumber	0.75	0.36	1.14	0.20	3.74	>.001
Context - IN	-4.41	-5.41	-3.41	0.51	-8.63	>.001
Phoneme - NONE	-1.30	-2.54	-0.06	0.63	-2.06	.04
Phoneme - OTHER	-1.23	-1.89	-0.57	0.34	-3.63	>.001
Context -IN:Phoneme - NONE	-0.47	-1.80	0.85	0.68	-0.70	.49
Context - IN:Phoneme - OTHER	0.23	-0.68	1.13	0.46	0.49	.62
TrialNumber: Context - IN	-1.23	-1.92	-0.55	0.35	-3.52	>.001

There was a significant main effect of *context* ($\chi^2(1) = 45.74$, $p < .001$). Indeed, regardless of the type of the first phoneme of an incoming turn, participants were more likely to rate the turn as a question in non-initiating than initiating context. Table 4 shows the results of the main model.

There was a significant main effect of *phoneme* ($\chi^2(2) = 13.83$, $p < .001$). In both contexts turns that started with *wh* phonemes were more likely to be rated as questions in comparison to turns starting with other phonemes or without the response from the second speaker. The model estimated that the probability of considering a turn a question was 90% for *wh* phonemes compared to 71% for *other* and 70% for *none* in non-initiating context. In initiating context this was 9% compared to 4% for *other* and 2% for *none* (see Table 5). There were no significant differences in question prediction between other phoneme and no response. Considering that in the experimental samples only one instance of /h/ phoneme was present, we ran the analysis with the samples containing this phoneme excluded. The results did not differ (see supporting information in Appendix II).

Table 5. Model estimate of the probability of participants rating a turn as a question based on the previous context and the first phoneme of the incoming turn.

		Phoneme		
Context		None	Other	*wh*
	Non-initiating	0.698	0.713	0.895
	Initiating	0.017	0.037	0.094

Importantly, we also assessed whether participants could differentiate between the type of the response sample (a question or not) from which the phoneme was extracted. We found no effect of the response type ($\chi^2(1) = 0.11$, $p = .75$). Thus, participants answered comparably to the phoneme samples that actually were questions and samples that were not questions. Most importantly, there was no interaction between response phoneme and the type of the response ($\chi^2 = 0.008$, $p = 0.93$). Thus, participants treated *wh* phonemes from real questions comparably to *wh* phonemes from other speech acts.

There was no significant interaction between *context* and *phoneme* ($\chi^2(2) = 1.34$, $p = 0.51$). However, the trend appears to be in the predicted direction (see Fig.4). Namely, if the incoming turn starts with *wh* phoneme and is preceded by non-initiating turn participants are more likely to think that the turn is a question that in initiating context. We address the lack of significant interaction between context and phoneme in the discussion.

3.2.2. Summary

We found a significant main effect of context and phoneme, but no interaction between these factors. Participants were more likely to rate an incoming turn as a question if it was preceded by non-initiating context in comparison to initiating context. Importantly, they were also more likely to rate an incoming turn if it started with *wh* phoneme in comparison to other phonemes or absence of any phoneme. There were no significant differences between answers to other phoneme and absence of any response. We also found that participants did not draw on other cues of questions in order to make their answer, considering that they answered to samples coming from both, real questions and not questions, comparably. There were slight individual differences in experimental samples and participants. Also, we found an effect of trial number – participants were more likely to answer "question" in later trials. The main effects were robust to all controls.

4. Discussion

People need to predict upcoming turns due to social and cognitive constraints in conversation, so they may use early cues to turn types to help them. In the present paper we were interested in carrying out a study that would provide ecologically valid but at the same time experimentally controlled insights about the phenomena under investigation – the cues to question recognition. We aimed to explore our hypotheses in an ecologically valid way with a corpus study. Subsequently, we wanted to ensure that the findings were supported in a controlled setting using an experimental study.

Based on previous findings, we hypothesized that people take into account context - the sequential type of the previous turn (initiating/ non-initiating) - in order to predict the type of an incoming turn. Based on some previous preliminary hypotheses (Slonimska & Roberts, in prep.) we also assumed that the first phoneme of the incoming turn might be used in order to infer/predict what kind of turn is being produced –a question or not a question. We predicted that people should be more likely to think that an incoming turn is a question if it was preceded by a non-initiating turn. We found support for this assumption in both corpus and experimental studies. We also hypothesized that the first phoneme of *wh-words* could function as an early cue in predicting that the turn will be a question. Indeed, this hypothesis was also confirmed in both studies.

We also assumed that the effect of the phoneme should be larger than the effect of context, considering that the corpus data was first split by phoneme in the decision tree. Surprisingly, we find the reversed pattern in the experimental study. Namely, we find a very strong effect of context and a weaker effect of phoneme.

Finally, based on the corpus study we also hypothesized that when people have both *pro-question* cues in the signal –*wh* phoneme in the incoming turn preceded by a non-initiating turn – they will be more likely to think that an incoming turn is a question. We did not gain statistical support for this hypothesis in the experimental study. Nevertheless, the trend was in the predicted direction.

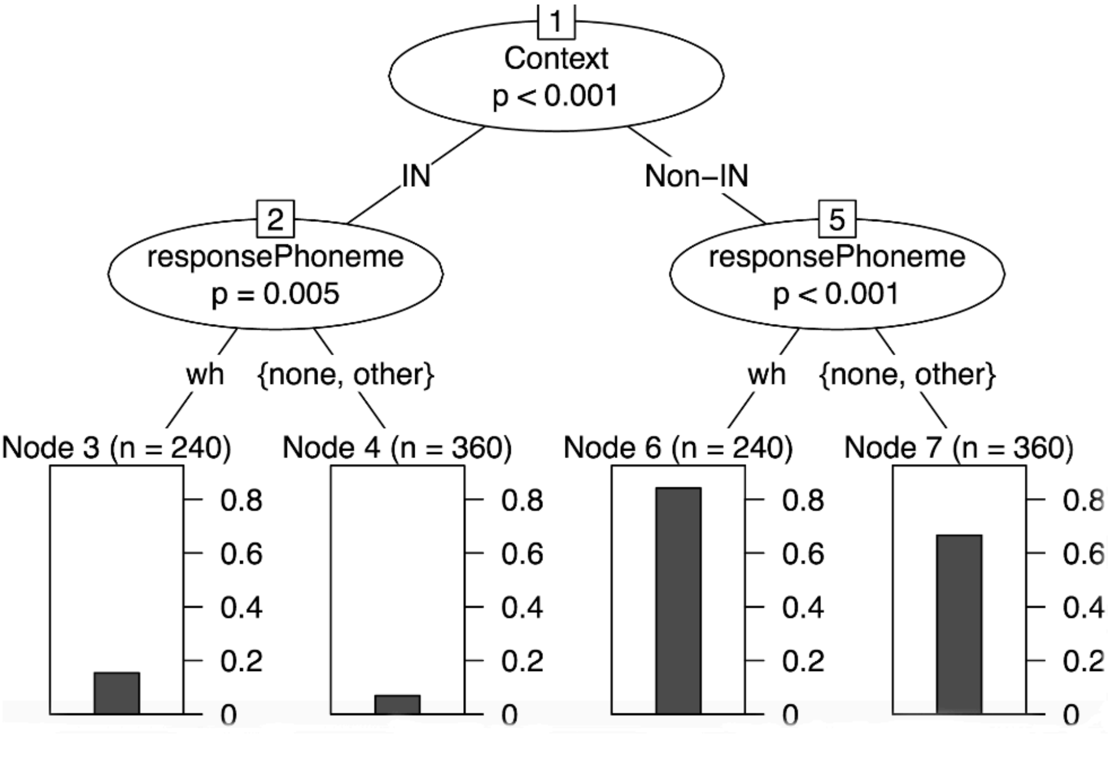
4.1. Question recognition: the role of the context and the first phoneme

In the corpus study we provided the decision tree with three variables. Dependent variable – turns that are content/open questions, and predictor variables – sequential type of the previous turn (initiating/non-initiating) and the first phoneme of the incoming turn. The data in the corpus was first split based on the phoneme and then based on the context. Note that the proportions of questions in the leaves of the decision tree showed that there was bigger probability overall to have a non-question in an incoming turn. Thus, if we would follow the decision tree in predicting an incoming turn as a question, we would benefit the most by first asking whether the first phoneme of the turn is a *wh* phoneme or not. Consecutively, if the turn started with a *wh* phoneme we would ask whether the previous turn was initiating or non-initiating. If it was non-initiating we could be more likely to expect that the incoming turn is a question than in any other case scenario. We expected to find the same pattern in the experimental study.

We created a decision tree for the data collected in the experimental study in order to have a clearer comparison with the results from the corpus study (see Fig.9). In this tree the proportion of a “question” answers in an incoming turn is much higher than in the decision tree of the corpus study. This, however, is due to large differences in the number of observations in both studies and a much higher proportion of possible question turns in the experimental study (see *Method* sections of both studies).

The decision tree of the experimental study first splits the data based on the context and only then based on the phoneme in both branches. Note that three levels of the factor *phoneme* are split in two leaves – *wh* leaf and *other & none* leaf. Thus, *wh* phonemes are treated differently from other two factors, while the decision tree does not differentiate between having another phoneme in the incoming turn or not having the turn at all. The pattern of the decision-making is exactly reversed in the decision tree of the experimental study in comparison to the decision tree in the corpus study. There are various possible interpretations to the discrepancies between the results from both studies.

Figure 9. The decision tree splitting the data of "question" answers of the participants based on predictor variables - *context* and *phoneme.*



The data in the corpus study comes from natural conversations. In contrast, in the experimental study the participants had only one sentence available to understand the context of the conversation. Thus, unlike in the experiment, the speakers in natural conversations not only have information on the preceding turn of the incoming speech act, but they also have the information about the unfolding of the conversation as a whole. This fact can be interpreted in two ways that either benefits or hinders question recognition.

4.1.1. Extensive context – benefiting question recognition

On one hand, the speakers in the corpus have more common ground between each other. In this light, the context aligning with the expectation of a question might have not come from a single speech act preceding the target turn, but from an entire turn or even from a sequence of turns. In contrast, in the experimental study participants were provided with a single sentence to form an expectation. Accordingly, there was much more attention paid to this sentence than it would be "in the wild" and thus it was taken advantage of.

Thus, in natural conversation the first phoneme of the turn was a stronger predictor of a question than a single previous turn while in experimental setting previous turn was the only context available. Thus, it overruled the information from the incoming first phoneme.

4.1.2. Extensive context – hindering question recognition

On the other hand, conversation is a stream of information, which is being updated continuously. According to Christiansen & Chater (2016), processing of the conversation can in part be interpreted as *Now or Never Bottleneck* – information has to be processed rapidly as it is pushed out of the memory very quickly. Moreover, linguistic regularities in input allow an addressee to process incoming information in such way (Christiansen & Chater, 2016). The matching phoneme of *wh-words* in English constitutes such regularity. Thus, in real conversation people might be more biased to first process incoming information - the first phoneme in the current case - and only then update the prediction of an incoming turn based on the previously available information. It is possible to argue that in natural conversation prediction of the next turn could be influenced to a greater extent by the early cues of the incoming turn rather than analyzing the speech acts just produced.

Also, keep in mind that, in a dialog, the context (i.e., previous-turn) is provided by the speaker who has to anticipate the incoming turn. In the corpus the speakers are all involved in an on-line task. In the experiment, the participants were passive listeners - they themselves were not actively involved in the conversation. In other words, in real conversation, it might be less demanding to monitor the partner in the dialog than monitor oneself and the entire unfolding of the conversation in order to draw on the cues to questions. The speakers in the corpus study had a cognitively demanding task – they had to plan their own turn as well as comprehend the last one. In the experiment, both - the phoneme of the incoming turn and the context - came from other speakers. Thus, the participants had to monitor others and not themselves. Moreover, there was no competing context information- there was only one sentence constituting the entire context. Accordingly, in comparison to the speakers in the corpus, they had high cognitive resources to process the information from context and use it.

Importantly, these two interpretations in regard to the previous speech act being less informative in a conversation can be tested experimentally. Namely, it is possible to design a study that would investigate whether prediction of a question is based on extensive context going beyond the last speech act and greater cognitive demands due to time constraints or whether it is due to a cognitive advantage of drawing on online cues provided by an interlocutor. This should be an endeavor for future research.

To sum up, in the experimental study the participants may be focusing more on the prior context, considering that they can afford to think about it more. In active conversation, however, it might be cognitively less demanding to focus on information available in the present moment (see Table 6 for the differences between studies). Thus, context plays a crucial role when amount of information available is low and cognitive resources high. In such settings it overrides or at least diminishes the informational benefit provided by phonemic cue of an incoming turn. On the other hand, when cognitive resources are low (i.e., in real conversations), online phonetic cues appear to be more important than context.

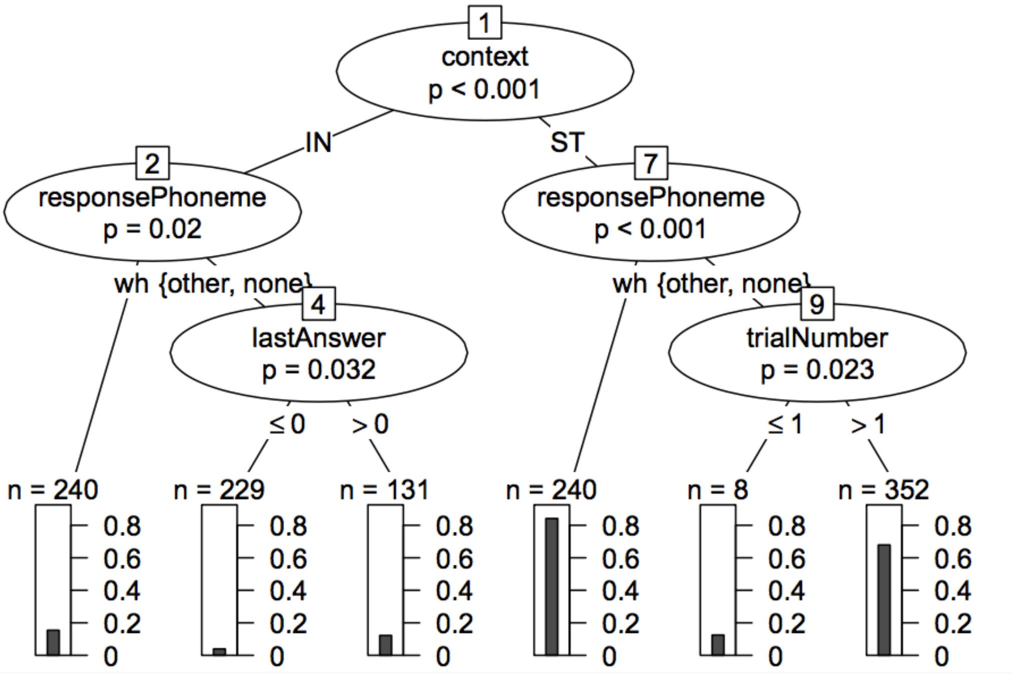
Table 6. The differences between corpus and experimental studies in regard to type of processing, amount of context and cognitive demands.

	Corpus study	**Experimental study**
Context	A lot	Little
Cognitive demands	High	Low
Processing	Active	Passive

4.2. First phoneme as a reliable cue to questions

We then run the same decision tree, but this time we included the possible confounding factors: trial number, last answer, age, sex and type of English of the participant (see Fig. 10). Just like in the previous tree, the data is first split based on the context and then on the phoneme. However, now it is evident that last response also plays a role in the initiating context and trial number plays a role in non-initiating context. Crucially, this split only regards the node of the data containing answers to *other & none* response samples. Answers to *wh* phoneme are not influenced. This indicates that participants were quite certain about how they respond to the samples with a second turn containing a *wh* phoneme, while they were more design-dependent in the other two cases. In other words – they were more certain about how to answer when they heard *wh* phoneme in comparison to hearing other phoneme or hearing nothing.

Figure 10. The decision tree splitting the answers of participants based on predictor variables - *context* and *phoneme* and possible confound variables – *previous answer* (lastAnswer, value <=0 indicates that previous answer was *Not a question* and value >0 – *A question*), *trial number*, *type of English spoken* by the participant, *sex* and *age* of the participant. The bar charts show the proportion of trials where participants thought the next turn was a question.



Also, we found that participants did not pick up on paralinguistic cues, if there were any, in the incoming turn. Words that started with *wh* phoneme but were not questions were treated as if they were questions. In other words, people could not tell the difference between questions and non-questions from the first phoneme. Also, the corpus study indicates that the first phoneme is a reliable cue for question recognition – thus, the finding that participants did not differentiate between real questions and not supports this observation.

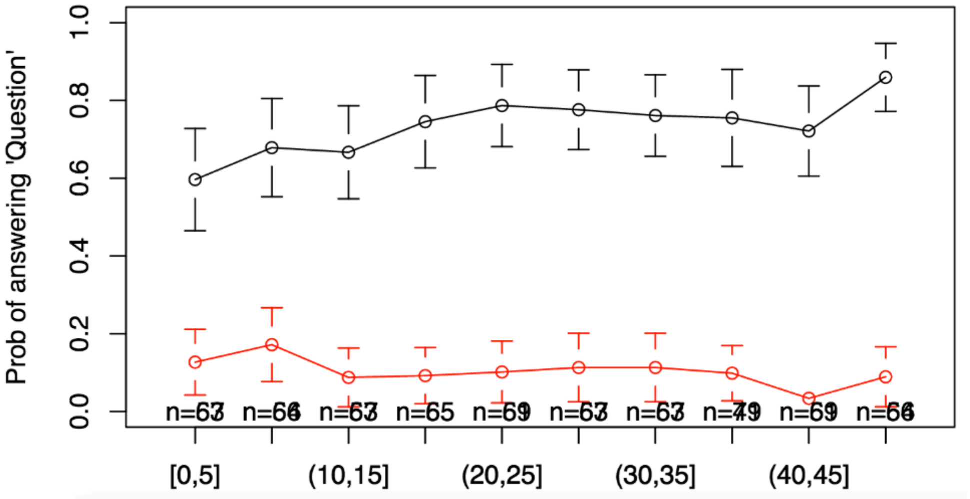
It is plausible that this early in the turn it is yet impossible to differentiate and take advantage of the changes in pitch and intonation to update the information about the incoming turn. Thus, it is plausible to assume that the phoneme itself is the very first trigger for the participants to considering an incoming turn as a question that is consecutively updated when other cues are starting to come in.

4.3. Added benefit of converging "pro-question" cues

In contradiction to our hypothesis we did not find an interaction between context and the first phoneme, even though the trend was in the predicted direction. It appears surprising that the convergence of two cues that lead to question recognition do no boost question recognition in comparison of having only one cue. For example, it is logical to assume that having an initiating context and *wh* phoneme in the incoming turn might create some hesitance due to the context not aligning with the expectation of a question. Also, the same can be said about non-initiating context and absence of the phonemic *wh* cue. In such case, there are many other possible speech acts that might be used.

We would like to argue that the lack of the interaction was not due to the fact that participants failed to draw on the benefit of having both cues in the signal. Instead, this was due to the fact that the participants would have needed an extra practice in order to understand the information they were presented with. It is plausible that the 2 test trials that participants had at the beginning of the experiment were not enough in order to understand what kind of information is at their disposal and thus they could not take advantage of it. Accordingly, participants might have used the beginning of the experiment to explore the samples and did not provide reliable responses. This assumption can be supported by the fact that there was a significant effect of trial and an interaction between trial and context (see Fig. 11). Namely, the trial factor was stronger in non-initiating contexts than in initiating contexts. In later trials participants were more likely to answer "question" in non-initiating context than initiating context. This might indicate that participants needed some time in order to understand that there are different types of previous turns available. Once this was taken up on, participants started using this cue. Note, that differences in the probability of a turn being a question were much higher in the experiment than in the corpus.

Figure 11. The answers "question" to experimental samples across trials. Black line indicates non-initiating context, red line- initiating context.



4.4. Shortcomings of the study

The goal of the study was to, first, explore natural conversational data in order to gain support and better understanding about the dynamics of question prediction and, second, use the findings to test the hypotheses in a controlled setting. The main findings are in line with the hypotheses in both studies. There are, however, some differences between the two in regard to the strength of the effect of the predictors.

We assumed that this is mainly due to the speakers in the corpus having more background information than participants in the experimental study. If so, it is possible to argue that the design of the experimental study is not appropriate to assess the findings of the corpus study. Future design should include more extensive background information for the participants to be able to make predictions about the incoming turn. Nevertheless, discrepancies that we found in both studies clearly highlighted that the amount of information available to the listener is of crucial importance.

Furthermore, given that the participants are only passive listeners of the audio samples it makes the comparison even more debatable. There should be some differences expected in regard to whether participants actually participate in the dialog or only observe/listen to it. The design of such study, however, is extremely challenging and would require an excessive consideration for a plethora of confounding factors.

Given that context turned out to be such a strong factor on question prediction, the choice of allowing participants to listen to the samples twice seems justified. However, the information on reaction times, if answer was required right after the first listening, could have been informative not only for assessing which factors ease question recognition, but also provide a more direct measure of cognitive processing. Possibly, the interaction that we could not confirm in our study would become evident in reaction times. One can argue that if the same experiment would include more extensive context as mentioned before, there would be no need to listen to the fragment twice and thus reaction times could be recorded as well.

Based on the findings we argued that the first phoneme of the *wh-words*, namely phonemes /w/ and /h/ would boost question recognition. We found support for this in the corpus study. However, there was only one instance of /h/ phoneme in the experimental samples. Thus, generalization to *wh* phonemes might seem too far fetched and not entirely valid. We run analyses with /h/ samples excluded and found no difference in the results. Thus, it is secure to argue that phoneme /w/ does indeed boost question recognition in English. The picture is less clear about phoneme /h/. Slonimska & Roberts (in prep.) argue that the similarity of the initial phoneme of the question word might serve as a cue to question recognition. Their argument is that question words tend to sound similar at the beginning of the word within a language to trigger question recognition. If this assumption is correct, we should expect that phoneme /w/ is a better predictor of a question than phoneme /h/, considering that there are more question words starting with /w/ and thus exactly this phoneme should be associated with a question. Nevertheless, in the corpus study we found that the phoneme /h/ was actually a better predictor of an incoming question than the phoneme /w/. This, however, was based on the fact that there were fewer instances overall of turns starting with /h/ and many fillers *well* in the cluster of /w/ phonemes. The hypothesis can be addressed in the future in a controlled setting by assessing the differences between hearing /w/ and /h/ at the beginning of a turn in regard to question recognition. For now, we gained support that phoneme /w/ does contribute to question recognition independently from context.

It is worth noting that we tested participants with different English dialect backgrounds. We found no differences in the results based on the type of English spoken by participants. It is also important to take into account that the participants in the study were not all monolingual and 9 of them were raised bilingually from birth. It is possible to argue that the fact that an individual has information on two languages and accordingly different phonemes for question words might create some differences. Possibly, this can be observed exactly for less competent second language speakers than for bilingual natives. It would be interesting to explore whether the patterns change for L2 speakers and whether both cues, context and first phoneme, are effective in question recognition.

Finally, the findings in this study regard only English language. It would be invaluable to conduct similar investigations in other languages to explore further the benefit of the first phoneme of the question. Even though typological data suggests that similarity of question words within a language is a plausible universal cue for question recognition, this should be reconciled in an experimental setting as we did with English.

5. Conclusion

In the present paper we set out to explore whether the first phoneme and context can serve as a cue to question recognition. We found that both of these features contribute to question recognition. Importantly, while an effect of context was clearly expected, it was less certain whether there would be an effect of the first phoneme, as there is almost no research supporting such hypothesis; if anything, some research disregards it as unlikely (Cysouw, 2004). This is the first study to support the claim of Slonimska & Roberts (in prep.) that the matching first phoneme of wh-words can be used to predict an upcoming question. Our findings, however, are limited to English language and future research should continue exploring this cue in other languages as well. Only in this way can we be certain that this is not a single-language phenomena or based on some idiosyncrasy of English but is actually a universal pattern. However, the puzzle remains - why else would question words sound so similar within so many languages (given that Slonimska & Roberts account for historical factors in their study and still find significant similarities)?

We were the first to approach this topic from two different but mutually enhancing perspectives. We assessed the hypotheses by, first, analyzing natural conversations. Thus, we could look for patterns in the ecologically valid data. The fact that the decision tree generated the same predictions as our hypotheses served as a sound basis for an experimental testing. The hypotheses were also confirmed in the controlled setting – the experiment. Even though we find differences in the strength of the effects, both of the effects are clearly there. The fact that we do find differences in regard to the results of both experimental approaches indicates how important accounting for both of them is. By looking for the phenomena in natural data, testing it in a controlled way and referring back to the real world can shed the light on the importance of many ignored features. The combination of these approaches can elucidate fine-grained details that can make the difference in the final results and thus impact the conclusions about the phenomena under investigation. It can raise new questions and, most importantly, it can inform about the phenomena in a much more valid way than by using single approach instead.

To summarize, due to using different approaches in exploring the same topic we now have a comprehensive picture of the first phoneme of wh-words as a cue to questions. Namely, languages tend to have a phonetic cue of question words as shown in the cross-cultural study by Slonimska & Roberts (in prep). This phonetic cue can help in predicting questions in real conversations as shown in the corpus analysis. Finally, we find that people actually use this cue to predict questions when presented in a semi-natural setting – namely, with context available. Thus, the property of question words sounding similar (i.e., matching first phoneme) is not a random occurrence, but is used as an early cue for question recognition.

References

- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Singmann, H. (2015). lme4: Linear mixed-effects models using Eigen and S4, 2014. *R package version*, 1-1.
- Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer. 7 http://www. fon. hum. uva. nl/praat/. Zugegriffen: 17.
- Boersma. P. (2002). Praat. a system for doing phonetics by computer. *Glot international*. *5*(9/10). 341-345
- Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*,*52*, 46-57.
- Bögels, S., Casillas, M., & Levinson, S. C. (2016, June). To plan or to listen? The trade-off between comprehension and production in conversation. In *the Eighth Annual Meeting of the Society for the Neurobiology of Language*.
- Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015, August). The significance of silence. Long gaps attenuate the preference for 'yes' responses in conversation. In *the 19th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2015/goDIAL)*.
- Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never Say No… How the Brain Interprets the Pregnant Pause in Conversation. *PloS one*, *10*(12), e0145474.
- Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific reports*, *5*.
- Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., & Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, *44*(4), 387-419.
- Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.
- Cysouw, M. (2004, February). Interrogative words: an exercise in lexical typology. In *Presentation presented at the Bantu grammar: description and theory workshop, February* (Vol. 13).
- De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 515-535.

- Enfield, N. J., Stivers, T., & Levinson, S. C. (2010). Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, *42*(10), 2615-2619.
- Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PloS one*, *10*(3), e0120068.
- Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992, March). SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* (Vol. 1, pp. 517-520). IEEE.
- Holler, J., & Kendrick, K. H. (2015). Unaddressed participants' gaze in multi-person interaction: optimizing recipiency. *Front. Psychol*, *6*(98), 10-3389.
- Holler, J., Kendrick, K. H., Casillas, M., & Levinson, S. C. (2016). Turn-Taking in Human Communicative Interaction.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651-674.
- Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: a quantitative study. *Discourse Processes*, *52*(4), 255-289.
- Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Levinson, S. C. (2006). On the human" interaction engine". In *Wenner-Gren Foundation for Anthropological Research, Symposium 134* (pp. 39-69). Berg.
- Levinson, S. C. (2013). Action formation and ascription. *The handbook of conversation analysis*, 101-130.
- Levinson, S. C. (2016). Speech acts. In Y. Huang (Ed.), Oxford handbook of pragmatics. Advanced online publication. Oxford: Oxford University Press.
- Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences*, *20*(1), 6-14.
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, *6*, 731.
- Mazzoni, D., & Dannenberg, R. (2000). Audacity (software). *The Audacity Team, Pittsburg, PA, USA*.
- prosody of dialogue. *Language resources and evaluation*. *44*(4). 387-419.

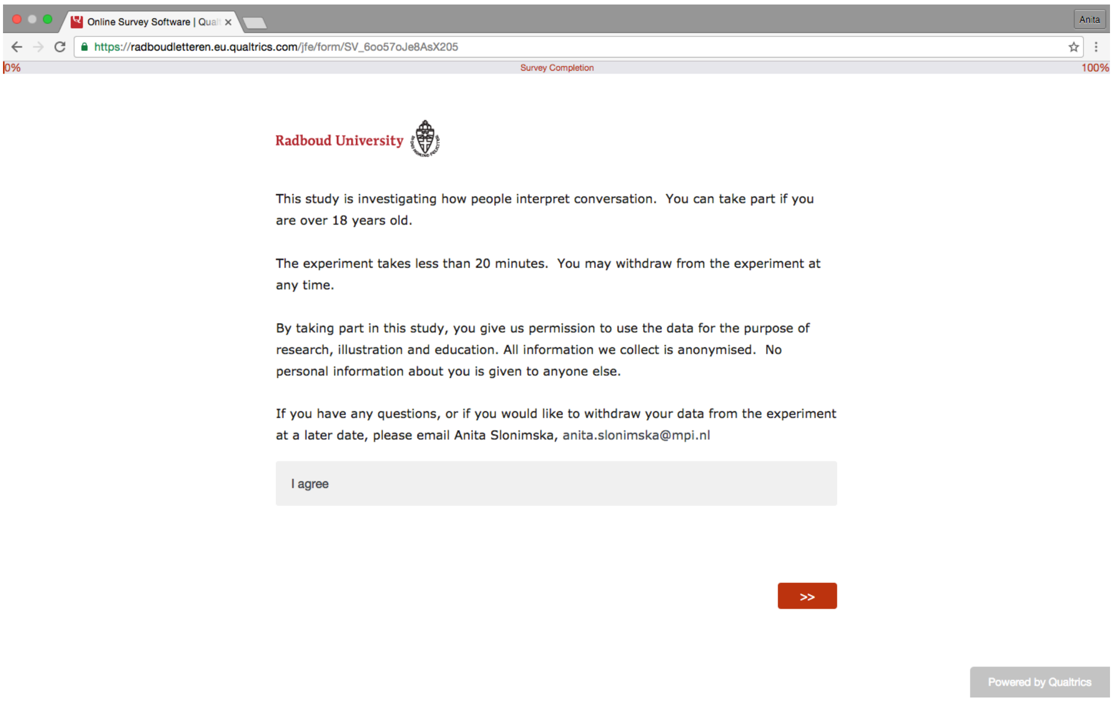
- Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, *133*(6), EL471-EL477.
- Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in psychology*, *6*, 509.
- Rossano, F. (2013). 15 Gaze in Conversation.
- Rossano, F., Brown, P., & Levinson, S. C. (2009). Gaze, questioning and culture. *Conversation analysis: Comparative perspectives*, 187-249.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, 696-735.
- Schegloff, E. A. (1979). *Identification and recognition in telephone conversation openings*.
- Schegloff, Emanuel. 1979. Identification and recognition in telephone conversation openings. In Psathas, George (ed.), Everyday language: Studies in ethnomethodology. Irvington, New York.
- Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies.*Journal of memory and language*, *29*(1), 86-102.
- Sicoli, M. A., Stivers, T., Enfield, N. J., & Levinson, S. C. (2014). Marked initial pitch in questions signals marked communicative function. *Language and speech*, 0023830914529247.
- Snow, J., & Mann, M. (2013). Qualtrics survey software: handbook for research professionals.
- Slonimska & Roberts (in prep.) A case for systematic sound symbolism in pragmatics: universals in *wh-words*.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587-10592.
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*,*14*(4), 323.
- Torreira, F., Bögels, S., & Levinson, S. C. (2015). Breathing for answering: the time course of response planning in conversation. *Front. Psychol*, *6*(284), 10-3389.

- Torsten Hothorn, Kurt Hornik and Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651--674.
- Winter. B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf

Appendix I

The experimental study administered through online software Qualtrics

The first page with a general instruction about the experiment


Online Survey Software | Qual
Anita
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205
0%
Survey Completion
100%
Radboud University

This study is investigating how people interpret conversation. You can take part if you are over 18 years old.

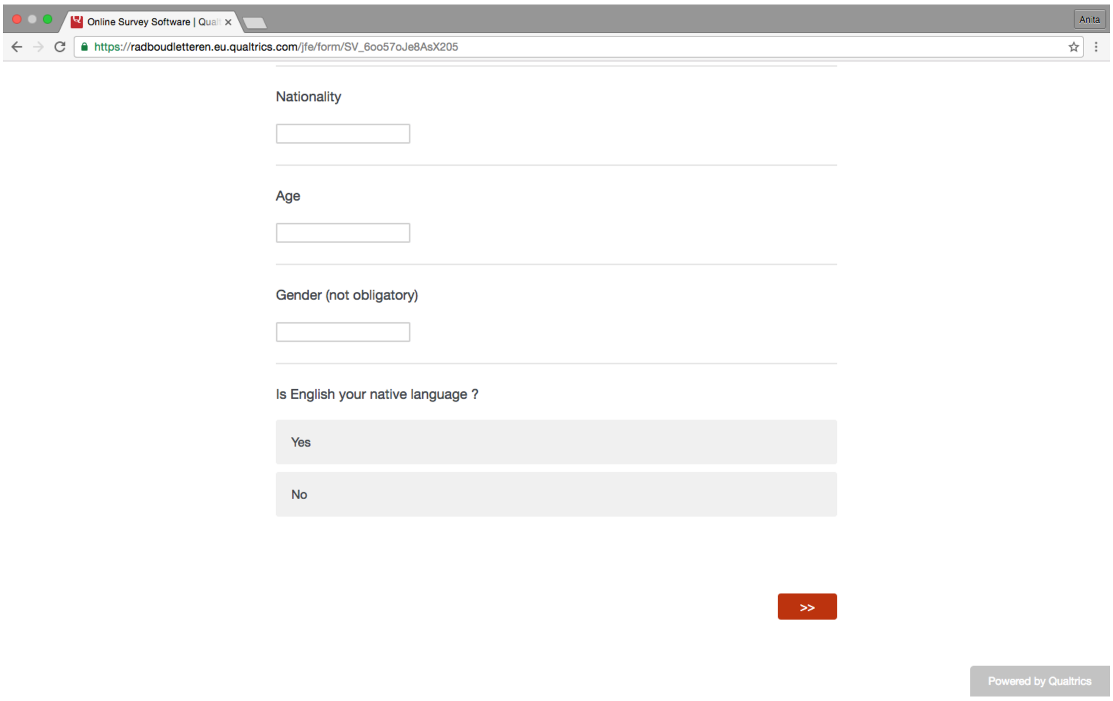
The experiment takes less than 20 minutes. You may withdraw from the experiment at any time.

By taking part in this study, you give us permission to use the data for the purpose of research, illustration and education. All information we collect is anonymised. No personal information about you is given to anyone else.

If you have any questions, or if you would like to withdraw your data from the experiment at a later date, please email Anita Slonimska, anita.slonimska@mpi.nl

I agree

General descriptives (a)


Online Survey Software | Qual
Anita
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205

Nationality

Age

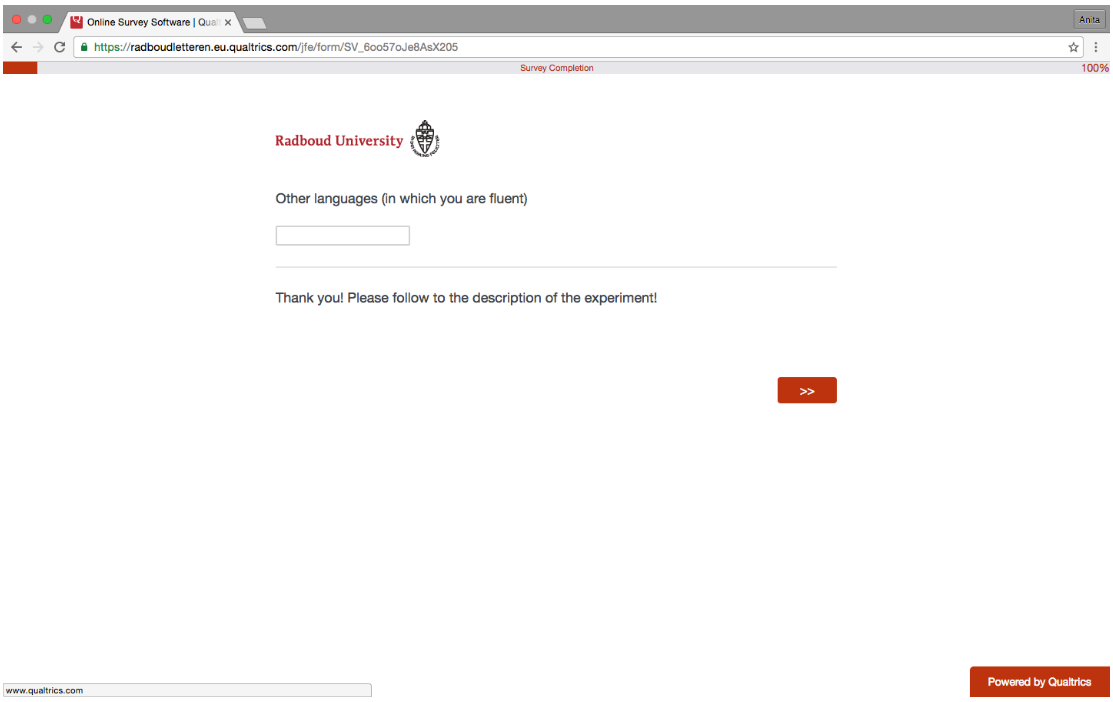
Gender (not obligatory)

Is English your native language ?

Yes

No

General descriptives (b)

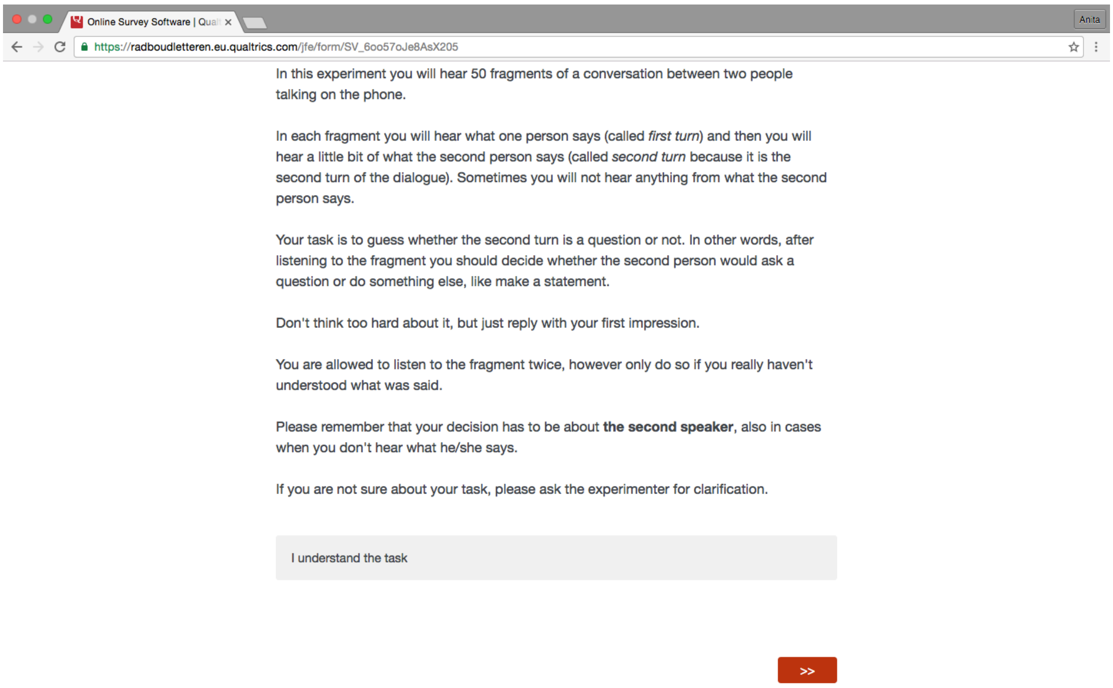

Online Survey Software | Qual
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205
Survey Completion
100%
Radboud University

Other languages (in which you are fluent)

Thank you! Please follow to the description of the experiment!

>>
www.qualtrics.com
Powered by Qualtrics

Description of the task


Online Survey Software | Qual
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205

In this experiment you will hear 50 fragments of a conversation between two people talking on the phone.

In each fragment you will hear what one person says (called *first turn*) and then you will hear a little bit of what the second person says (called *second turn* because it is the second turn of the dialogue). Sometimes you will not hear anything from what the second person says.

Your task is to guess whether the second turn is a question or not. In other words, after listening to the fragment you should decide whether the second person would ask a question or do something else, like make a statement.

Don't think too hard about it, but just reply with your first impression.

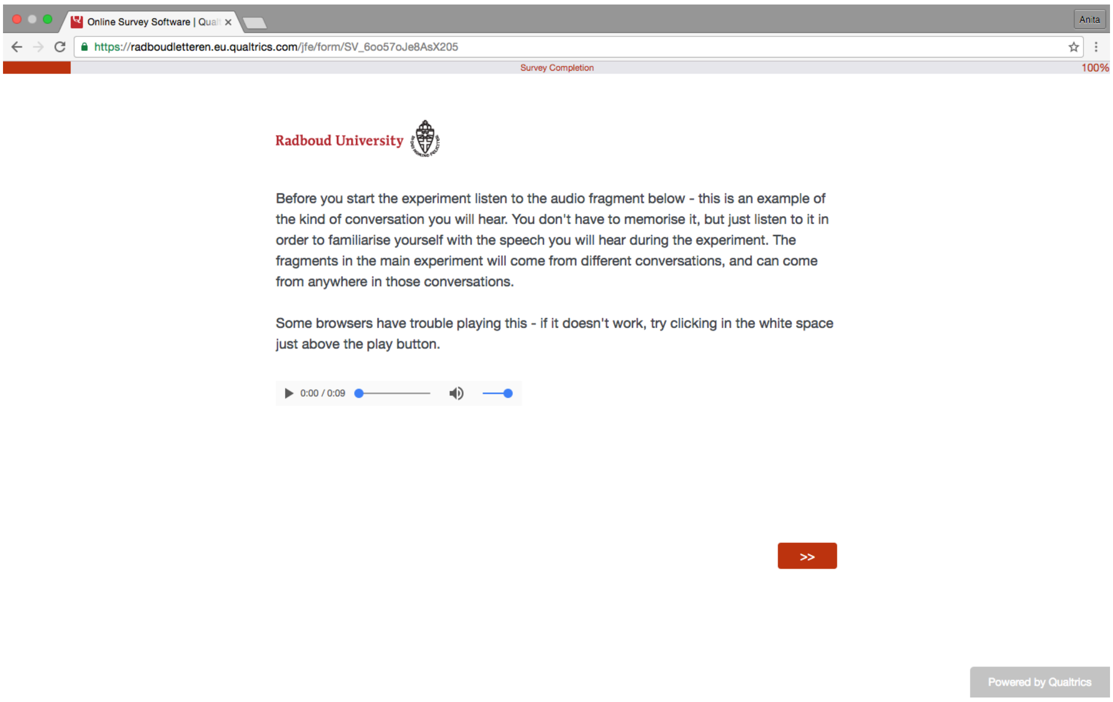
You are allowed to listen to the fragment twice, however only do so if you really haven't understood what was said.

Please remember that your decision has to be about **the second speaker**, also in cases when you don't hear what he/she says.

If you are not sure about your task, please ask the experimenter for clarification.

I understand the task

Familiarization sample

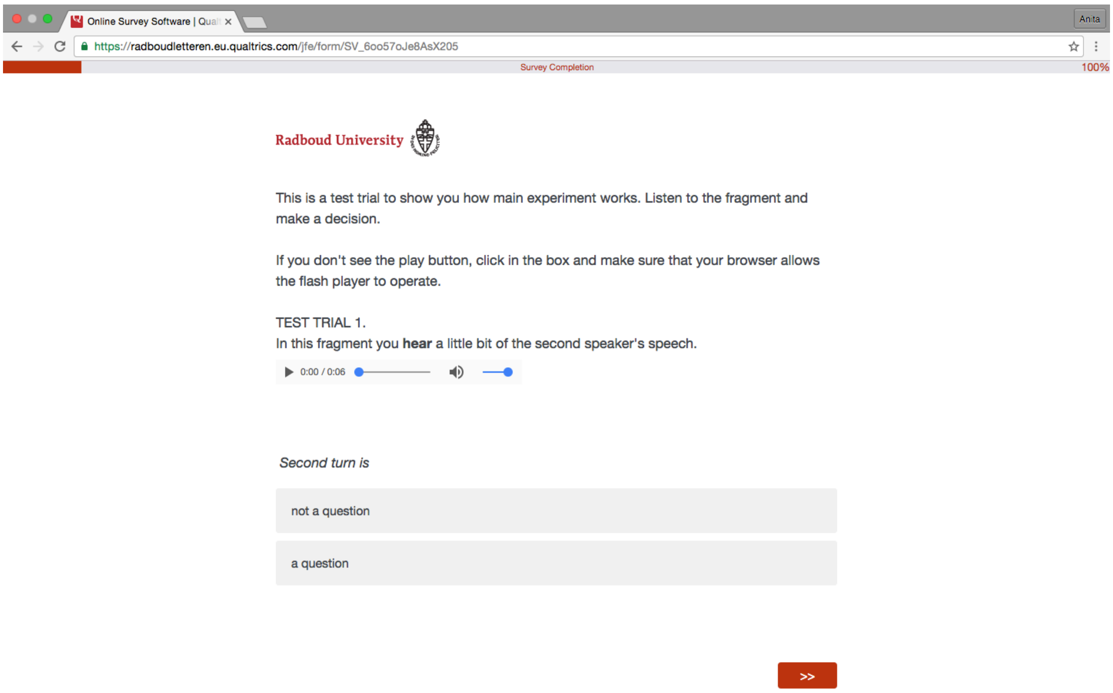


Online Survey Software | Qual
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205
Survey Completion
100%
Radboud University

Before you start the experiment listen to the audio fragment below - this is an example of the kind of conversation you will hear. You don't have to memorise it, but just listen to it in order to familiarise yourself with the speech you will hear during the experiment. The fragments in the main experiment will come from different conversations, and can come from anywhere in those conversations.

Some browsers have trouble playing this - if it doesn't work, try clicking in the white space just above the play button.

Test trial 1



Online Survey Software | Qual
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205
Survey Completion
100%
Radboud University

This is a test trial to show you how main experiment works. Listen to the fragment and make a decision.

If you don't see the play button, click in the box and make sure that your browser allows the flash player to operate.

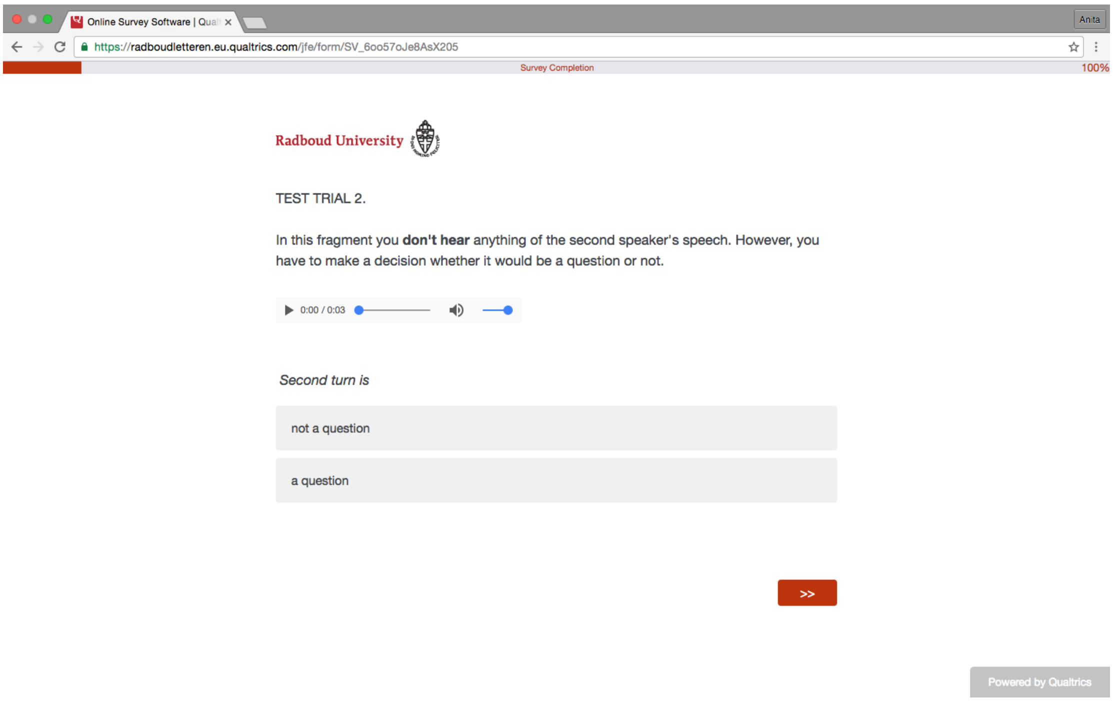
TEST TRIAL 1.
In this fragment you **hear** a little bit of the second speaker's speech.

Second turn is

not a question

a question

Test trial 2


Online Survey Software | Qual
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205
Survey Completion
100%
Radboud University
0:00 / 0:03
>>
Powered by Qualtrics

TEST TRIAL 2.

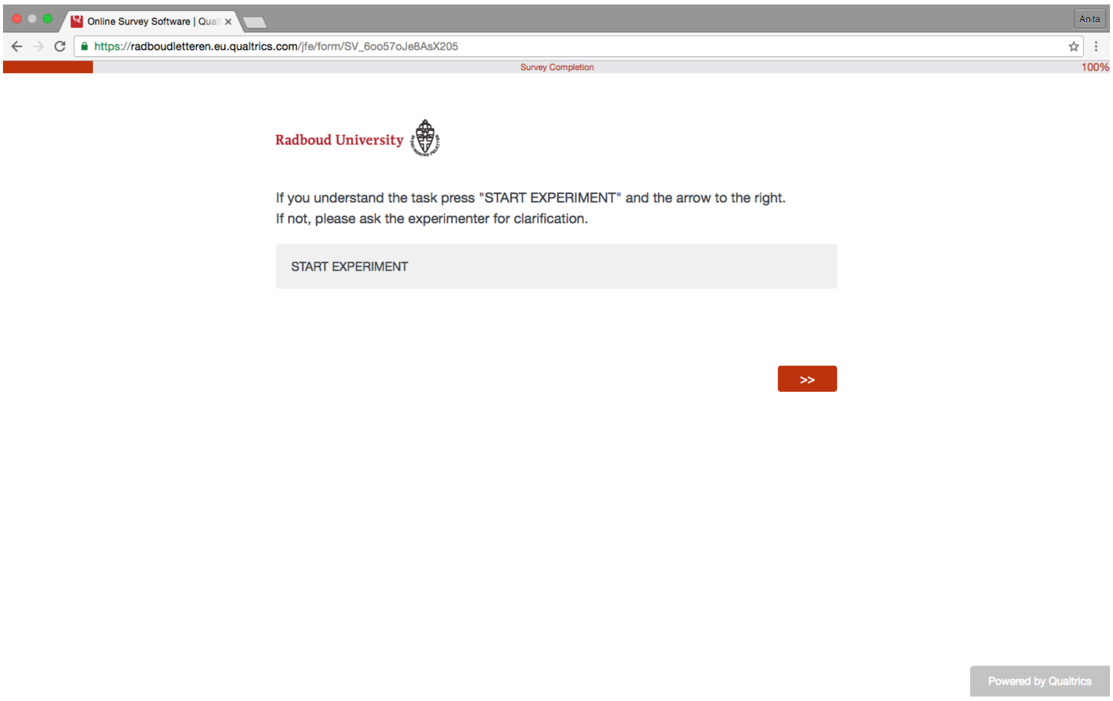
In this fragment you **don't hear** anything of the second speaker's speech. However, you have to make a decision whether it would be a question or not.

Second turn is

not a question

a question

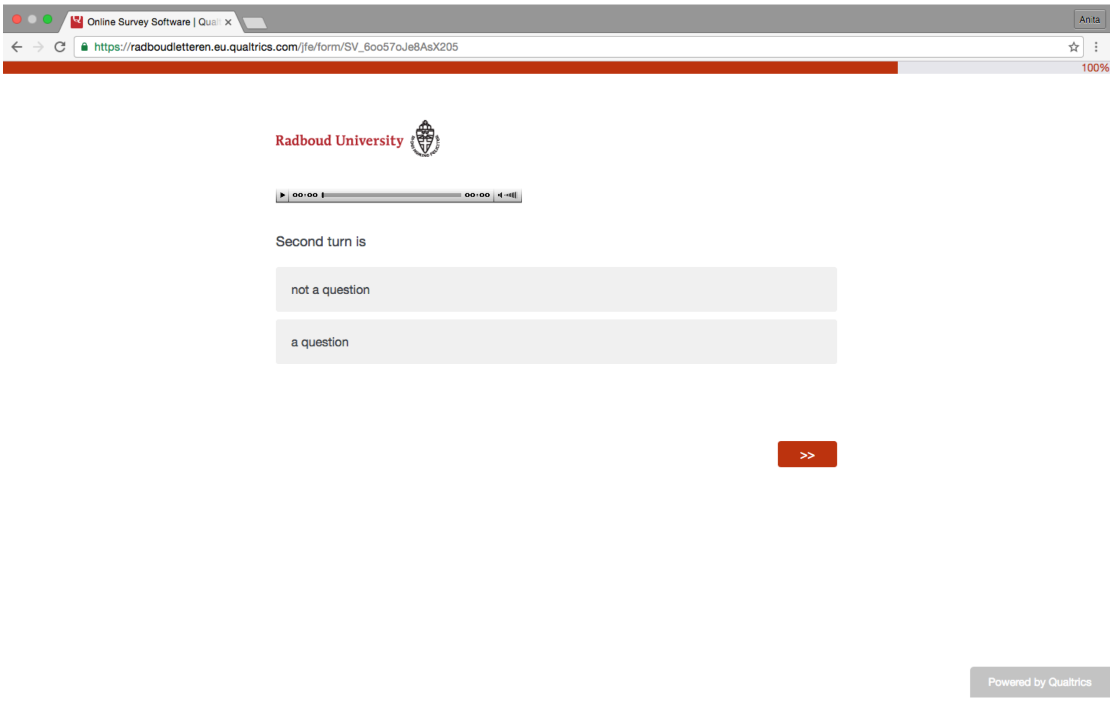
Page: Start experiment


Online Survey Software | Qual
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205
Survey Completion
100%
Radboud University
>>
Powered by Qualtrics

If you understand the task press "START EXPERIMENT" and the arrow to the right.
If not, please ask the experimenter for clarification.

START EXPERIMENT

The first trial of the experiment

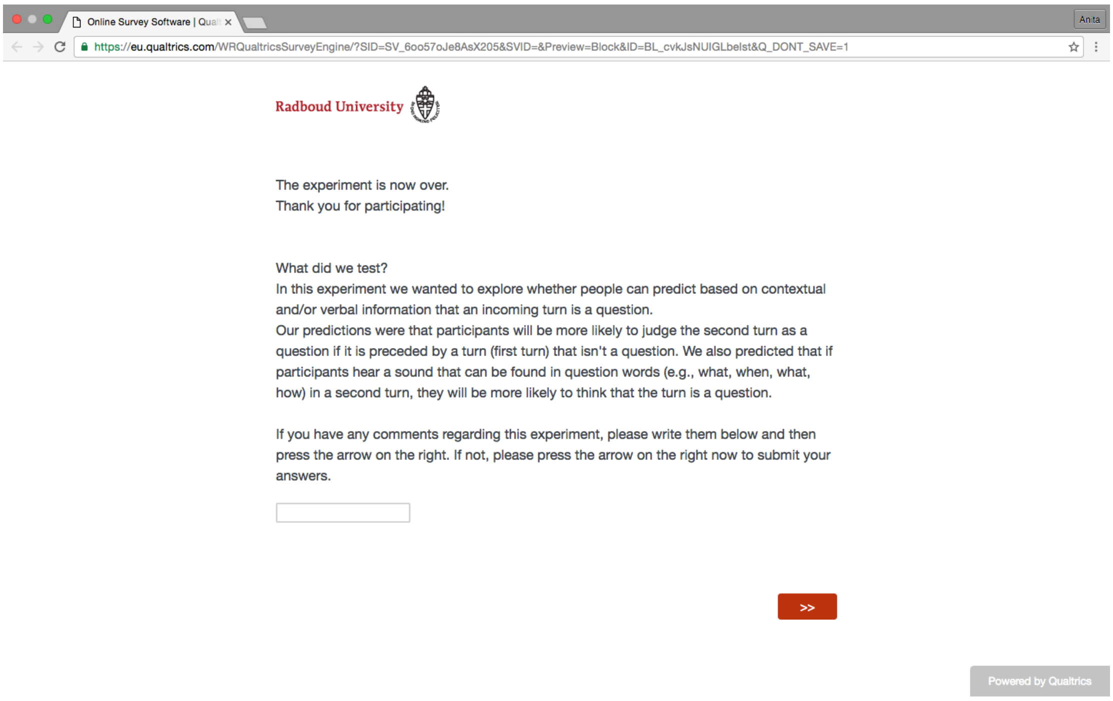

Online Survey Software | Qual
Anita
https://radboudletteren.eu.qualtrics.com/jfe/form/SV_6oo57oJe8AsX205
100%
Radboud University
00:00
00:00
>>
Powered by Qualtrics

Second turn is

not a question

a question

The final page of the experiment: debriefing


Online Survey Software | Qual
Anita
https://eu.qualtrics.com/WRQualtricsSurveyEngine/?SID=SV_6oo57oJe8AsX205&SVID=&Preview=Block&ID=BL_cvkJsNUIGLbelst&Q_DONT_SAVE=1
Radboud University
>>
Powered by Qualtrics

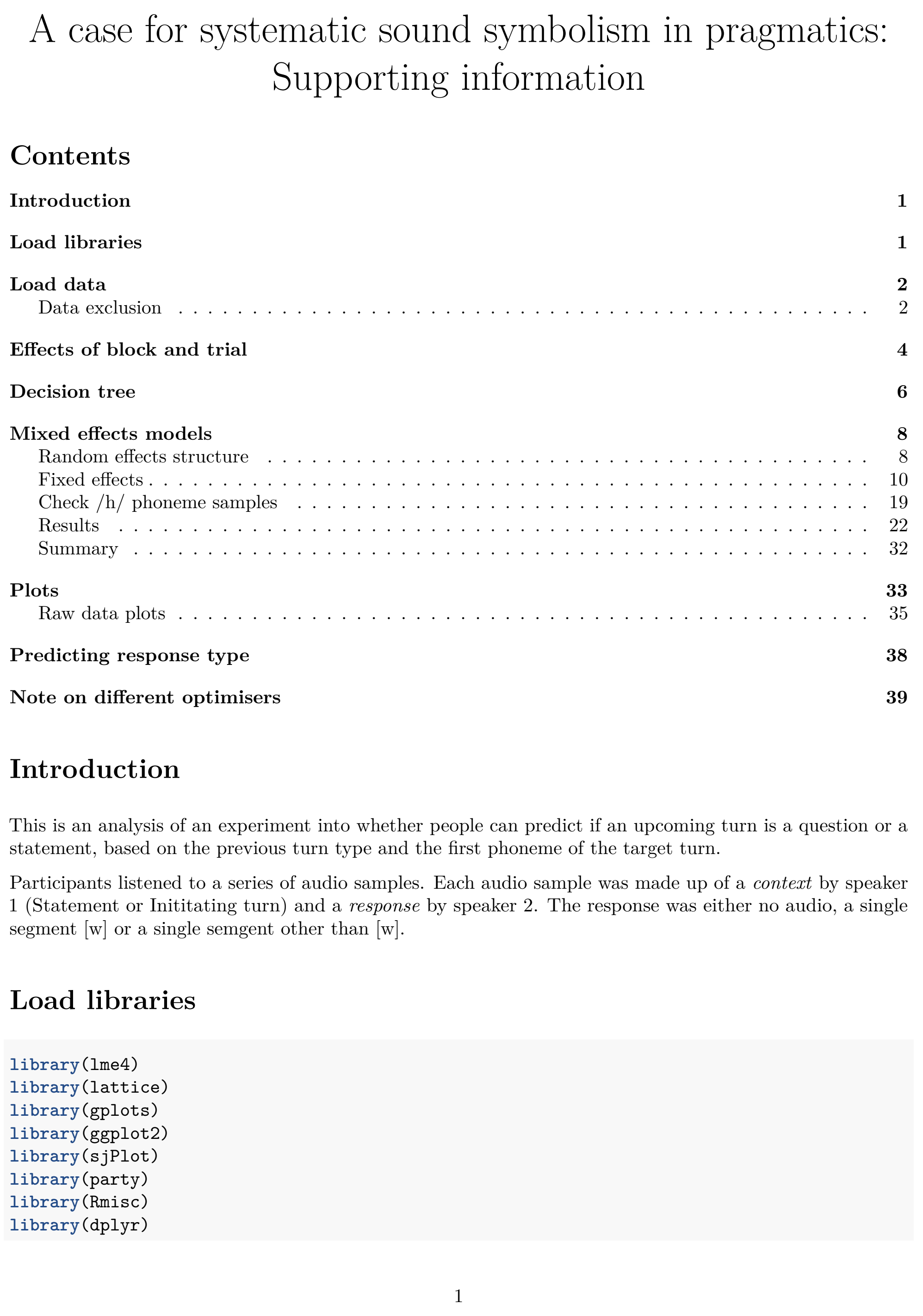
The experiment is now over.
Thank you for participating!

What did we test?
In this experiment we wanted to explore whether people can predict based on contextual and/or verbal information that an incoming turn is a question.
Our predictions were that participants will be more likely to judge the second turn as a question if it is preceded by a turn (first turn) that isn't a question. We also predicted that if participants hear a sound that can be found in question words (e.g., what, when, what, how) in a second turn, they will be more likely to think that the turn is a question.

If you have any comments regarding this experiment, please write them below and then press the arrow on the right. If not, please press the arrow on the right now to submit your answers.

Appendix II

Supplementary material of the mixed effect models analysis



A case for systematic sound symbolism in pragmatics: Supporting information

Contents

Introduction

This is an analysis of an experiment into whether people can predict if an upcoming turn is a question or a statement, based on the previous turn type and the first phoneme of the target turn.

Participants listened to a series of audio samples. Each audio sample was made up of a *context* by speaker 1 (Statement or Inititating turn) and a *response* by speaker 2. The response was either no audio, a single segment [w] or a single semgent other than [w].

Load libraries

```
library(lme4)
library(lattice)
library(gplots)
library(ggplot2)
library(sjPlot)
library(party)
library(Rmisc)
library(dplyr)
library("lme4")
library("optimx")
#library("nloptr")
```

Function for converting from logit scale

```
logit2per = function(X){
  return(exp(X)/(1+exp(X)))
}
```

Load data

```
d = read.csv("../Data/Lab_Processed.csv")
```

Each row in the data is a single response from a participant to a single sample. The key variables are:

- *partID*: identifies participants
- *contextSample*: The name of the audio sample used for the context.
- *responseSample*: The name of the audio sample used for the response.
- *responsePhoneme*: The first segment of the response.
- *responseType*: Whether the first segment of the response came from a question or statement.
- *answer*: The participant's response to "Is the next turn a question?"

Make *answer* a binary variable.

```
d$answer = d$answer=="Yes"
d$lastAnswer = d$lastAnswer=="Yes"
```

Relevel response phoneme and context.

```
d$responsePhoneme = relevel(d$responsePhoneme, 'wh')
d$context = relevel(d$context, 'ST')
```

Center trial number, so that the intercept will reflect probabilities in the middle of the experiment.

```
d$trialNumber.center = d$trialNumber - 25
# Scale between -1 and 1
d$trialNumber.center = d$trialNumber.center /
  max(d$trialNumber.center)
```

Data exclusion

We exclude participant 13 because they took much longer than other participants.

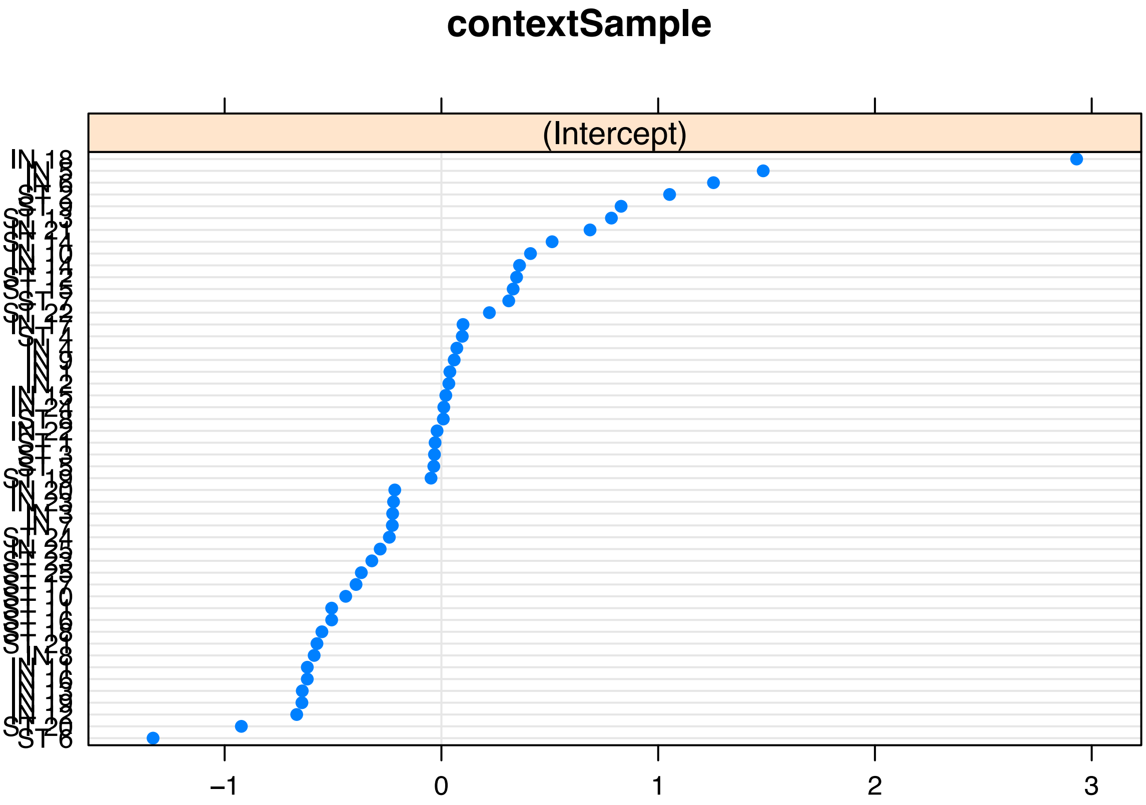
```
d = d[as.character(d$partID)!="13",]
```

Are there any samples that look like outliers? Make a basic model:

```
m3 = glmer(
  answer ~ 1 + context + responsePhoneme +
    (1 | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = glmerControl(optimizer="bobyqa", optCtrl = list(maxfun=2e4))
)
```

Then look at the random effects.

```
dotplot(ranef(m3))[[2]]
```



The sample “IN 18” is an outlier. However, models have convergence problems when leaving it out.

The data has 1200 observations:

```
# Number of observations per participant
table(d$partID)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 14 15 16 17 18 19 20 21 22 23 24 25
## 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50 50
```

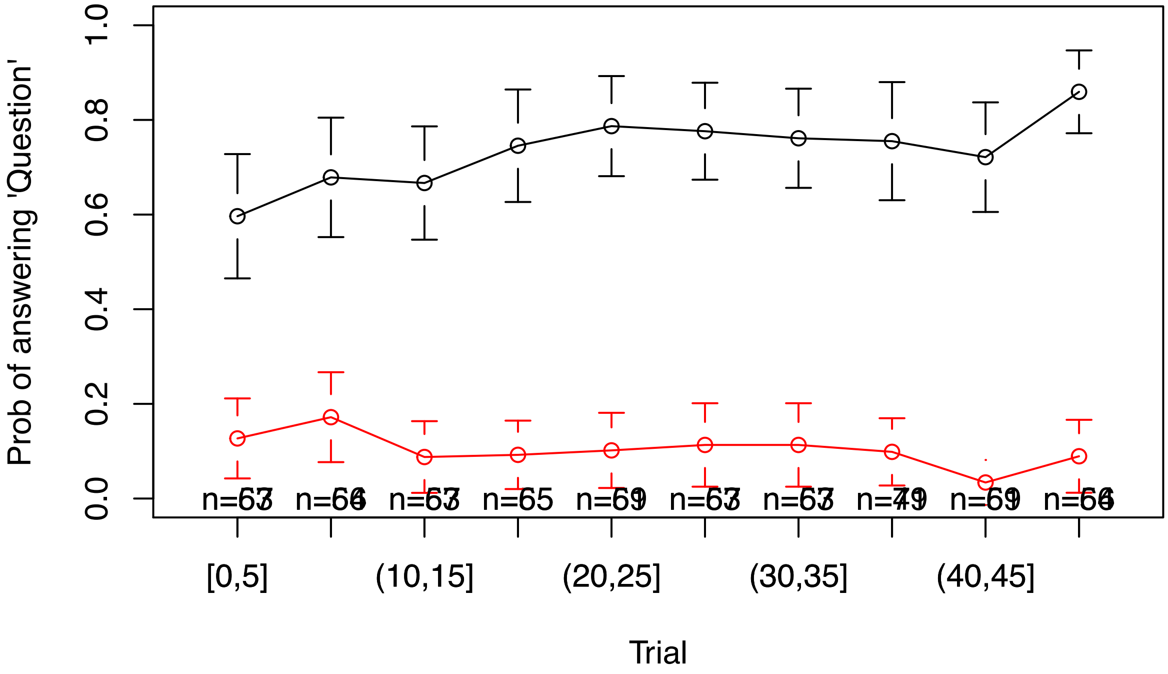
```
table(d$context, d$responsePheneme )
```

```
##
##       wh none other
##   ST 240  120   240
##   IN 240  120   240
```

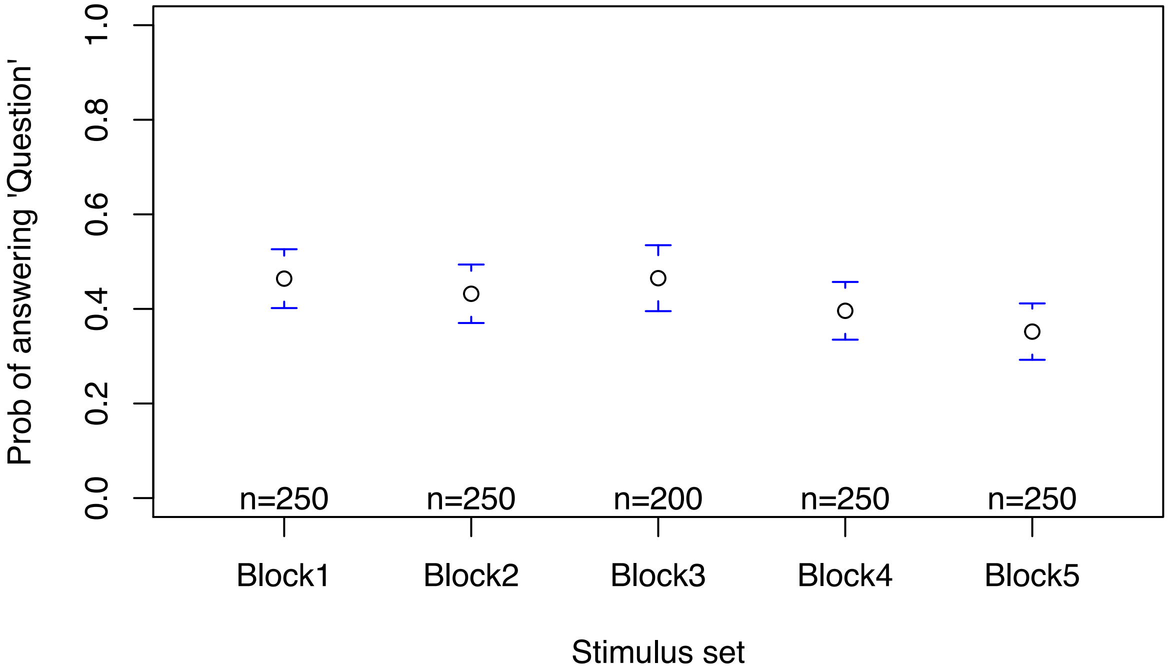
Effects of block and trial

```
plotmeans(answer ~ cut(trialNumber,seq(0,50,length.out = 11), include.lowest = T),
          ylab = "Prob of answering 'Question'",
          xlab = 'Trial',
          data = d[d$context=="ST",],ylim=c(0,1),
          col = 1, barcol = 1)
plotmeans(answer ~ cut(trialNumber,seq(0,50,length.out = 11), include.lowest = T),
          ylab = "Prob of answering 'Question'",
          xlab = 'Trial',
          data = d[d$context=="IN",],ylim=c(0,1),
          col = 2, barcol = 2, add=T)
```

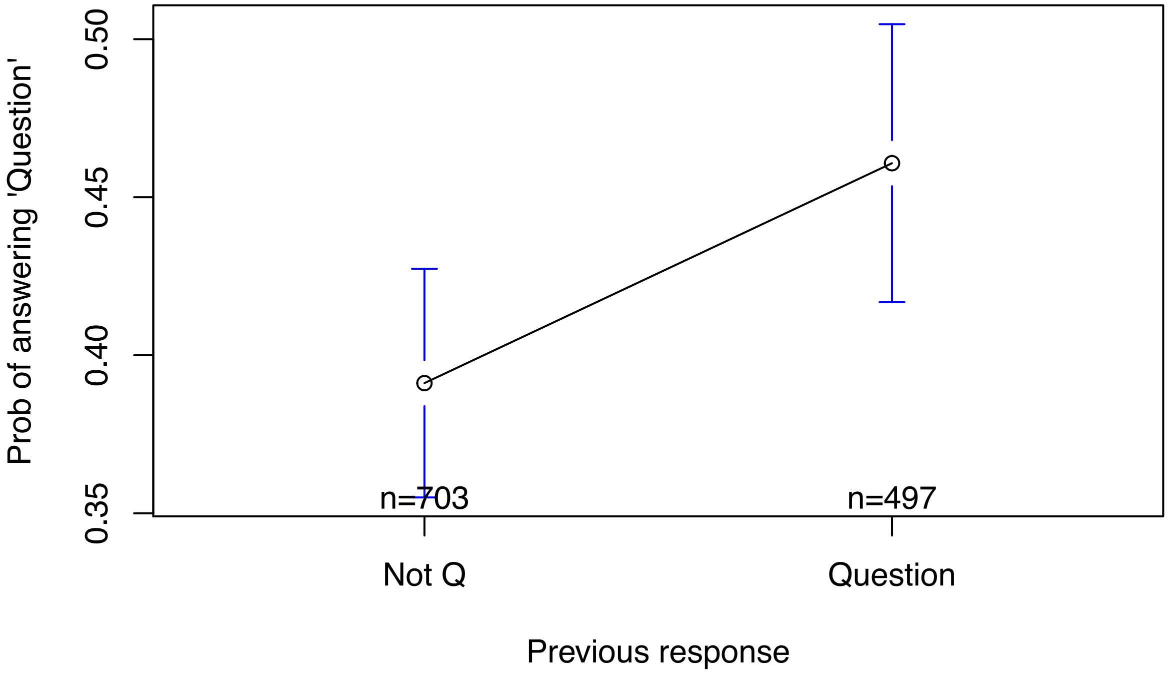
```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, :
## zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, :
## zero-length arrow is of indeterminate angle and so skipped

## Warning in axis(1, at = 1:length(means), labels = legends, ...): "add" is
## not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "add" is not a
## graphical parameter
```



```
plotmeans(d$answer ~ d$blockName,
          ylab = "Prob of answering 'Question'",
          xlab = 'Stimulus set',
          connect=F,
          ylim=c(0,1))
```



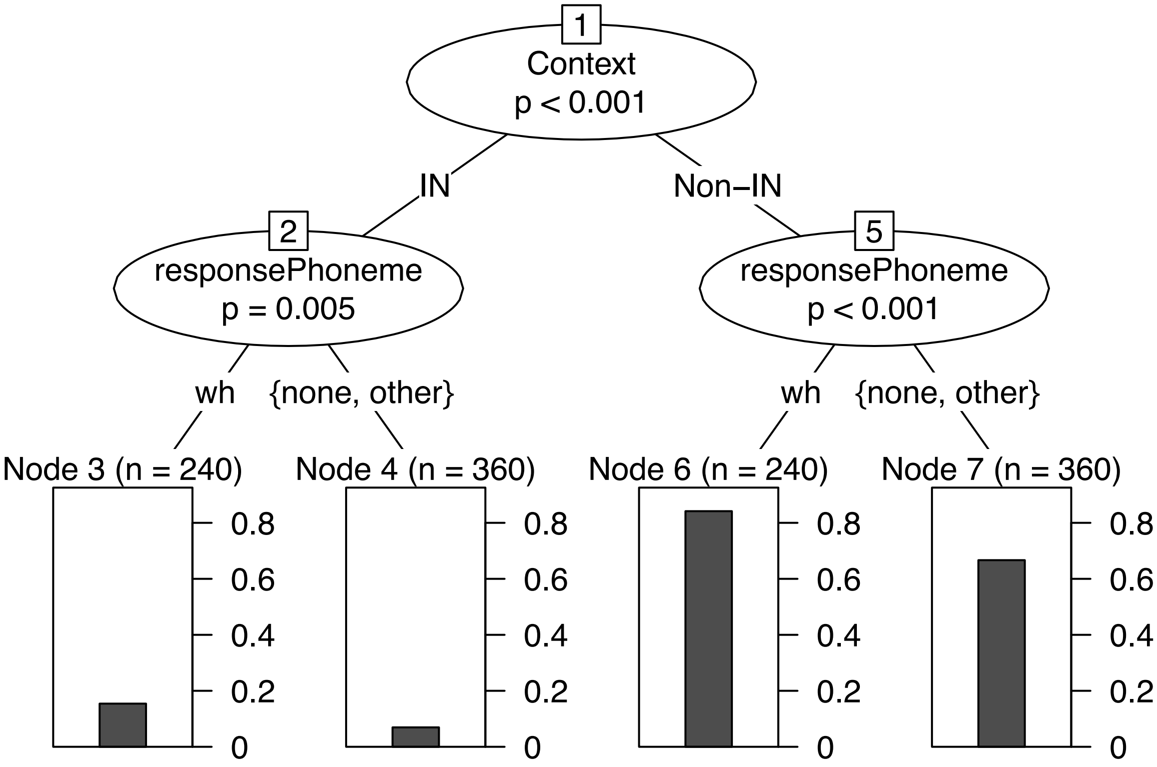
```r
plotmeans(answer ~ lastAnswer,
          ylab = "Prob of answering 'Question'",
          xlab = "Previous response",
          legends = c("Not Q", "Question"),
          data = d)
```



Decision tree

In order to get an idea of the structure of the data, we make a binary decision tree based on the data. We try to predict the participant's response by context and the type of turn the response was taken from.

```
d$Context = factor(d$context,labels = c("Non-IN","IN"))

cx.simple = ctree(answer ~
             Context +
               responsePhoneme + responseType, data = d)
plot(cx.simple, terminal_panel=node_barplot(cx.simple))
```

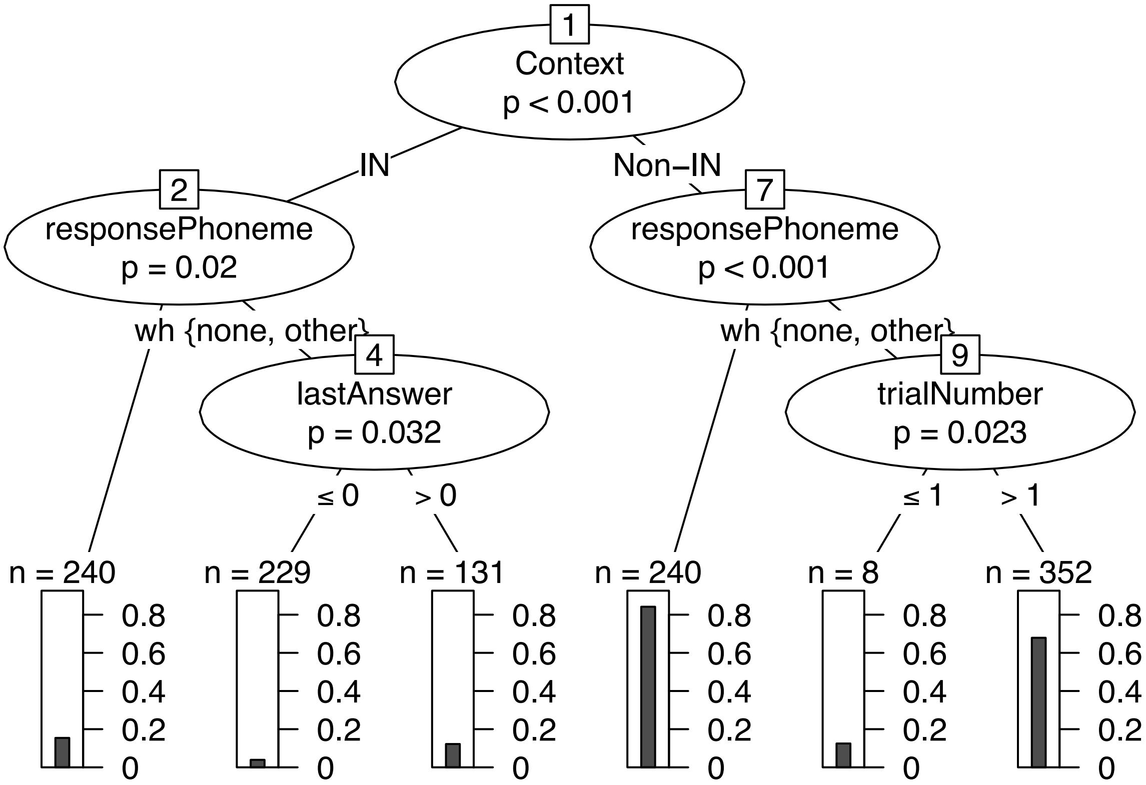


And here is a more detailed analysis:

```
cx = ctree(answer ~
             Context +responsePhoneme + responseType +
             Age + Sex + EnglishType +
             response.sex + context.sex +
             trialNumber + lastAnswer +
             blockName,
           data = d,
           controls = ctree_control(mincriterion = 0.95))
```

Plot the decision tree:

```
plot(cx, terminal_panel=node_barplot(cx, id=F))
```



Context is the most important factor, followed by first phoneme of the response.

Mixed effects models

Make a series of mixed effects models. We can fix this using the "nlminb" optimiser for both phases of the convergence and letting the algorithm run longer:

```
nlminbw   <- lme4:::nlminbwrap
gcontrol = glmerControl(optimizer="nlminbw",optCtrl = list(maxfun=2e4))
```

(Note that several convergence algorithms were tested, and the three best fitting solutions had essentially no differences in fixed effect estimates)

Random effects structure

We have a good idea of what the random effects structure should be, but first we check whether there are significant differences by participant etc.

```
mA0 = glmer(
  answer ~ 1 +
    (1 | partID),
  data = d,
  family = binomial,
  control = glmerControl(optimizer='bobyqa',optCtrl=list(maxfun=2e4))
)


mA0b = glmer(
  answer ~ 1 +
    (1 | blockName/partID) ,
  data = d,
  family = binomial,
  control = glmerControl(optimizer='bobyqa',optCtrl=list(maxfun=2e4))
)
ltrf = anova(mA0,mA0b)
ltrf
```

```
## Data: d
## Models:
## mA0: answer ~ 1 + (1 | partID)
## mA0b: answer ~ 1 + (1 | blockName/partID)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mA0   2 1636.7 1646.9 -816.35   1632.7
## mA0b  3 1637.6 1652.8 -815.78   1631.6 1.1313      1     0.2875
```

There is no significant improvement in the model when taking stimulus set into account. Because it complicates the analysis, we'll leave it out.

```
## Mixed effect models summary
##
##  ../results/lmerTests/lmerTestSummary.txt
```

```
mA1 = glmer(
  answer ~ 1 +
    (1 | partID) +
    (1 | contextSample),
  data = d,
  family = binomial,
  control = gcontrol
)

mA2 = glmer(
  answer ~ 1 +
    (1 | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

mA3 = glmer(
  answer ~ 1 +
    (1 + context| partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

mA4 = glmer(
  answer ~ 1 +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

ltrf2 = anova(mA0, mA1, mA2, mA3, mA4)
ltrf2
```

```
## Data: d
## Models:
## mA0: answer ~ 1 + (1 | partID)
## mA1: answer ~ 1 + (1 | partID) + (1 | contextSample)
## mA2: answer ~ 1 + (1 | partID) + (1 | contextSample) + (1 | responseSample)
## mA3: answer ~ 1 + (1 + context | partID) + (1 | contextSample) + (1 |
## mA3:     responseSample)
## mA4: answer ~ 1 + (1 + context + responsePhoneme | partID) + (1 |
## mA4:     contextSample) + (1 | responseSample)
##     Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## mA0  2 1636.7 1646.9 -816.35   1632.7
## mA1  3 1116.4 1131.7 -555.21   1110.4 522.281      1  < 2.2e-16 ***
## mA2  4 1094.7 1115.1 -543.37   1086.7  23.682      1  1.136e-06 ***
## mA3  6 1059.5 1090.0 -523.76   1047.5  39.222      2  3.041e-09 ***
## mA4 13 1054.1 1120.3 -514.06   1028.1  19.393      7   0.007041 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All proposed random effects significantly improve the fit of the model, except for the random slope for responsePhoneme by participant.

Fixed effects

We are most interested in the effects of context and response type, but we need to check some other possible confounding variables.

Trial

```
m0 = glmer(
  answer ~ 1 +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)

trial = glmer(
  answer ~ 1 + trialNumber.center +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

trialQ = glmer(
  answer ~ 1 + trialNumber.center + I(trialNumber.center^2) +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)


lttr = anova(m0,trial, trialQ)
lttr
```

```
## Data: d
## Models:
## m0: answer ~ 1 + (1 + context + responsePhoneme | partID) + (1 |
## m0:     contextSample) + (1 | responseSample)
## trial: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trial:     partID) + (1 | contextSample) + (1 | responseSample)
## trialQ: answer ~ 1 + trialNumber.center + I(trialNumber.center^2) + (1 +
## trialQ:     context + responsePhoneme | partID) + (1 | contextSample) +
## trialQ:     (1 | responseSample)
##        Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## m0     13 1054.1 1120.3 -514.06   1028.1
## trial  14 1051.3 1122.6 -511.66   1023.3 4.796      1    0.02853 *
## trialQ 15 1052.5 1128.8 -511.23   1022.5 0.862      1    0.35318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A significant effect of trial, but no significant quadratic term.

Previous answer

```
prevAns = glmer(
  answer ~ 1 + trialNumber.center + lastAnswer +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)
ltpa = anova(trial,prevAns)
ltpa
```

```
## Data: d
## Models:
## trial: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trial:     partID) + (1 | contextSample) + (1 | responseSample)
## prevAns: answer ~ 1 + trialNumber.center + lastAnswer + (1 + context +
## prevAns:     responsePhoneme | partID) + (1 | contextSample) + (1 | responseSample)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## trial   14 1051.3 1122.6 -511.66   1023.3
## prevAns 15 1051.6 1127.9 -510.80   1021.6 1.7284      1     0.1886
```

No significant effect of previous answer.

Sex of speakers in samples

```
contS = glmer(
  answer ~ 1 + trialNumber.center +
    context.sex +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)

respS = glmer(
  answer ~ 1 + trialNumber.center +
    context.sex + response.sex +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)
```

```r
contXrespS = glmer(
  answer ~ 1 + trialNumber.center +
    context.sex * response.sex +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)


ltsx = anova(trial,contS, respS, contXrespS)
ltsx
```

```
## Data: d
## Models:
## trial: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trial:     partID) + (1 | contextSample) + (1 | responseSample)
## contS: answer ~ 1 + trialNumber.center + context.sex + (1 + context +
## contS:     responsePhoneme | partID) + (1 | contextSample) + (1 | responseSample)
## respS: answer ~ 1 + trialNumber.center + context.sex + response.sex +
## respS:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## respS:     (1 | responseSample)
## contXrespS: answer ~ 1 + trialNumber.center + context.sex * response.sex +
## contXrespS:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## contXrespS:     (1 | responseSample)
##            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## trial      14 1051.3 1122.6 -511.66   1023.3
## contS      15 1051.8 1128.1 -510.90   1021.8 1.5291      1     0.2162
## respS      16 1053.8 1135.2 -510.89   1021.8 0.0194      1     0.8892
## contXrespS 17 1055.8 1142.3 -510.87   1021.8 0.0235      1     0.8783
```

No significant effects of the sex of the speakers in the samples.

Sex of participants

```r
sex = glmer(
  answer ~ 1 + trialNumber.center + Sex +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)

ltsxp = anova(trial,sex)
ltsxp
```

```
## Data: d
## Models:
## trial: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trial:     partID) + (1 | contextSample) + (1 | responseSample)
## sex: answer ~ 1 + trialNumber.center + Sex + (1 + context + responsePhoneme |
## sex:     partID) + (1 | contextSample) + (1 | responseSample)
##       Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## trial 14 1051.3 1122.6 -511.66   1023.3
## sex   15 1053.3 1129.7 -511.65   1023.3  0.02      1     0.8874
```

No significant effect of the sex of the participant.

Age of participants

(does't converge with nlminb, so using bobyqa)

```
age = glmer(
  answer ~ 1 + trialNumber.center + Age +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = glmerControl(optimizer="bobyqa")
)

ltag = anova(trial,age)
ltag
```

```
## Data: d
## Models:
## trial: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trial:     partID) + (1 | contextSample) + (1 | responseSample)
## age: answer ~ 1 + trialNumber.center + Age + (1 + context + responsePhoneme |
## age:     partID) + (1 | contextSample) + (1 | responseSample)
##       Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## trial 14 1051.3 1122.6 -511.66   1023.3
## age   15 1052.5 1128.9 -511.25   1022.5 0.8138      1      0.367
```

No significant effect of age of partcipant.

Type of English spoken

```
Etype = glmer(
  answer ~ 1 + trialNumber.center + EnglishType +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)

lten = anova(trial,Etype)
lten
```

```
## Data: d
## Models:
## trial: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trial:     partID) + (1 | contextSample) + (1 | responseSample)
## Etype: answer ~ 1 + trialNumber.center + EnglishType + (1 + context +
## Etype:     responsePhoneme | partID) + (1 | contextSample) + (1 | responseSample)
##       Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## trial 14 1051.3 1122.6 -511.66   1023.3
## Etype 16 1053.2 1134.7 -510.61   1021.2 2.0944      2     0.3509
```

No significant effec of the type of English the participant speaks.

Effects of Context and Response

The only significant confounding variable is trial, so that forms the baseline.

```r
context = glmer(
  answer ~ 1 + trialNumber.center +
    context +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)

rPhon = glmer(
  answer ~ 1 + trialNumber.center +
    context + responsePhoneme +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

conXrPh = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

anova(trial, context,rPhon, conXrPh)
```

```
## Data: d
## Models:
## trial: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trial:     partID) + (1 | contextSample) + (1 | responseSample)
## context: answer ~ 1 + trialNumber.center + context + (1 + context + responsePhoneme |
## context:     partID) + (1 | contextSample) + (1 | responseSample)
## rPhon: answer ~ 1 + trialNumber.center + context + responsePhoneme +
## rPhon:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## rPhon:     (1 | responseSample)
## conXrPh: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## conXrPh:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## conXrPh:     (1 | responseSample)
##         Df     AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## trial   14 1051.32 1122.6 -511.66  1023.32
## context 15 1007.58 1083.9 -488.79   977.58 45.742      1  1.349e-11 ***
## rPhon   17  997.75 1084.3 -481.88   963.75 13.828      2  0.0009938 ***
## conXrPh 19 1000.41 1097.1 -481.20   962.41  1.344      2  0.5106922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction between Sex and responses

```
Sex = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    Sex +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

SexXresp = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    Sex*responsePhoneme +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)

SexXcon = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    Sex*responsePhoneme +
    Sex:context +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)

SxXcoXre = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    Sex*responsePhoneme*context +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d,
  family = binomial,
  control = gcontrol
)
```

```
ltsxx = anova(conXrPh, Sex, SexXresp, SexXcon, SxXcoXre)
ltsxx
```

```
## Data: d
## Models:
## conXrPh: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## conXrPh:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## conXrPh:     (1 | responseSample)
## Sex: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## Sex:     Sex + (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## Sex:     (1 | responseSample)
## SexXresp: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## SexXresp:     Sex * responsePhoneme + (1 + context + responsePhoneme |
## SexXresp:     partID) + (1 | contextSample) + (1 | responseSample)
## SexXcon: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## SexXcon:     Sex * responsePhoneme + Sex:context + (1 + context + responsePhoneme |
## SexXcon:     partID) + (1 | contextSample) + (1 | responseSample)
## SxXcoXre: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## SxXcoXre:     Sex * responsePhoneme * context + (1 + context + responsePhoneme |
## SxXcoXre:     partID) + (1 | contextSample) + (1 | responseSample)
##          Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## conXrPh  19 1000.4 1097.1 -481.20   962.41
## Sex      20 1002.4 1104.2 -481.20   962.40 0.0049      1     0.9439
## SexXresp 22 1003.1 1115.1 -479.57   959.14 3.2639      2     0.1955
## SexXcon  23 1004.6 1121.7 -479.30   958.59 0.5471      1     0.4595
## SxXcoXre 25 1008.2 1135.5 -479.13   958.25 0.3400      2     0.8437
```

No effect by sex of participant.

Interaction with trial

```
trialXCon = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    trialNumber.center:context +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

trialXph = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    trialNumber.center:context +
    trialNumber.center:responsePhoneme +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)

trXcoXph = glmer(
  answer ~ 1 + trialNumber.center *
    context * responsePhoneme +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d,
  family = binomial,
  control = gcontrol
)
```

```
lttrx = anova(conXrPh, trialXCon, trialXph, trXcoXph)
lttrx
```

```
## Data: d
## Models:
## conXrPh: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## conXrPh:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## conXrPh:     (1 | responseSample)
## trialXCon: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## trialXCon:     trialNumber.center:context + (1 + context + responsePhoneme |
## trialXCon:     partID) + (1 | contextSample) + (1 | responseSample)
## trialXph: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## trialXph:     trialNumber.center:context + trialNumber.center:responsePhoneme +
## trialXph:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## trialXph:     (1 | responseSample)
## trXcoXph: answer ~ 1 + trialNumber.center * context * responsePhoneme +
## trXcoXph:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## trXcoXph:     (1 | responseSample)
##           Df     AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## conXrPh   19 1000.41 1097.1 -481.20   962.41
## trialXCon 20  989.60 1091.4 -474.80   949.60 12.8092      1  0.0003449 ***
## trialXph  22  993.25 1105.2 -474.62   949.25  0.3540      2  0.8377824
## trXcoXph  24  996.63 1118.8 -474.32   948.63  0.6147      2  0.7353795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant interaction between trial and context, but no reliable further interaction.

Effect of stimulus set

Below we adjust the random effects so that participants are nested within stimulus set (the variable *blockName*).

```
stimSet = glmer(
     answer ~ 1 + trialNumber.center +
         context * responsePhoneme +
         trialNumber.center:context +
         (1 + context  + responsePhoneme| blockName/partID) +
         (1 | contextSample) +
         (1 | responseSample),
     data = d,
     family = binomial,
     control = gcontrol
 )
```

```
anova(trialXCon,stimSet)
```

```
## Data: d
## Models:
## trialXCon: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## trialXCon:     trialNumber.center:context + (1 + context + responsePhoneme |
## trialXCon:     partID) + (1 | contextSample) + (1 | responseSample)
## stimSet: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## stimSet:     trialNumber.center:context + (1 + context + responsePhoneme |
## stimSet:     blockName/partID) + (1 | contextSample) + (1 | responseSample)
##           Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## trialXCon 20  989.6 1091.4 -474.80   949.60
## stimSet   30 1002.3 1155.0 -471.16   942.32 7.2839     10     0.6984
```

There is no significant improvement in the model, and in any case the stimuli sets are counterbalanced experimentally, so we don't include it.

In any case, the qualitative results are the same, and the estimates are very similar, suggesting that stimulus set does not have an impact on the main findings.

```
cbind(without=fixef(trialXCon),withRForStimSet=fixef(stimSet))
```

```
##                                  without withRForStimSet
## (Intercept)                    2.1405023      2.12652608
## trialNumber.center             0.7511443      0.77419202
## contextIN                     -4.4088146     -4.37373392
## responsePhonemenone           -1.3028431     -1.27113955
## responsePhonemeother          -1.2297526     -1.21010008
## contextIN:responsePhonemenone -0.4728529     -0.52928626
## contextIN:responsePhonemeother  0.2271069      0.03170117
## trialNumber.center:contextIN  -1.2321227     -1.25867272
```

Check /h/ phoneme samples

Only one stimuli set had a /h/ response phoneme, so we re-run the main analysis without those trials.

```
takeOutSet = d[d$response.firstO=='h',]$setNum[1]

trialH = glmer(
  answer ~ 1 + trialNumber.center +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d[d$setNum != takeOutSet,],
  family = binomial,
  control = gcontrol
)

contextH = glmer(
  answer ~ 1 + trialNumber.center +
    context +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample) ,
  data = d[d$setNum != takeOutSet,],
  family = binomial,
  control = gcontrol
)

rPhonH = glmer(
  answer ~ 1 + trialNumber.center +
    context + responsePhoneme +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d[d$setNum != takeOutSet,],
  family = binomial,
  control = gcontrol
)

conXrPhH = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    (1 + context + responsePhoneme | partID) +
    (1 | contextSample) +
    (1 | responseSample),
  data = d[d$setNum != takeOutSet,],
  family = binomial,
  control = gcontrol
)

trialXConH = glmer(
  answer ~ 1 + trialNumber.center +
    context * responsePhoneme +
    trialNumber.center:context +
    (1 + context + responsePhoneme | partID) +
```

```r
    (1 | contextSample) +
    (1 | responseSample),
  data = d[d$setNum != takeOutSet,],
  family = binomial,
  control = gcontrol
)

anova(trialH, contextH,rPhonH, conXrPhH, trialXConH)
```

```
## Data: d[d$setNum != takeOutSet, ]
## Models:
## trialH: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trialH:     partID) + (1 | contextSample) + (1 | responseSample)
## contextH: answer ~ 1 + trialNumber.center + context + (1 + context + responsePhoneme |
## contextH:     partID) + (1 | contextSample) + (1 | responseSample)
## rPhonH: answer ~ 1 + trialNumber.center + context + responsePhoneme +
## rPhonH:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## rPhonH:     (1 | responseSample)
## conXrPhH: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## conXrPhH:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## conXrPhH:     (1 | responseSample)
## trialXConH: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## trialXConH:     trialNumber.center:context + (1 + context + responsePhoneme |
## trialXConH:     partID) + (1 | contextSample) + (1 | responseSample)
##            Df     AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## trialH     14 1022.03 1092.7 -497.02   994.03
## contextH   15  978.65 1054.4 -474.32   948.65 45.3814      1  1.622e-11
## rPhonH     17  969.63 1055.5 -467.82   935.63 13.0146      2  0.0014925
## conXrPhH   19  972.25 1068.2 -467.13   934.25  1.3801      2  0.5015480
## trialXConH 20  962.34 1063.3 -461.17   922.34 11.9158      1  0.0005566
##
## trialH
## contextH   ***
## rPhonH     **
## conXrPhH
## trialXConH ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(trialXConH)$coef
```

```
##                                   Estimate Std. Error    z value
## (Intercept)                      2.0887338  0.3635915  5.7447265
## trialNumber.center               0.7216970  0.2021668  3.5698089
## contextIN                       -4.3725230  0.5135860 -8.5137119
## responsePhonemenone             -1.2261652  0.6124030 -2.0022194
## responsePhonemeother            -1.1943781  0.3375941 -3.5379117
## contextIN:responsePhonemenone   -0.4386208  0.6676415 -0.6569705
## contextIN:responsePhonemeother   0.2484692  0.4676457  0.5313193
## trialNumber.center:contextIN    -1.2000509  0.3534447 -3.3953006
##                                     Pr(>|z|)
## (Intercept)                     9.206970e-09
## trialNumber.center              3.572417e-04
## contextIN                       1.684522e-17
## responsePhonemenone            4.526114e-02
## responsePhonemeother           4.033049e-04
## contextIN:responsePhonemenone  5.111999e-01
## contextIN:responsePhonemeother 5.951976e-01
## trialNumber.center:contextIN   6.855328e-04
```

There are no qualitative differences when removing these trials.

Results

Model comparison

```
mainResults = anova(m0, trial, context,rPhon, conXrPh, trialXCon)
mainResults
```

```
## Data: d
## Models:
## m0: answer ~ 1 + (1 + context + responsePhoneme | partID) + (1 |
## m0:     contextSample) + (1 | responseSample)
## trial: answer ~ 1 + trialNumber.center + (1 + context + responsePhoneme |
## trial:     partID) + (1 | contextSample) + (1 | responseSample)
## context: answer ~ 1 + trialNumber.center + context + (1 + context + responsePhoneme |
## context:     partID) + (1 | contextSample) + (1 | responseSample)
## rPhon: answer ~ 1 + trialNumber.center + context + responsePhoneme +
## rPhon:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## rPhon:     (1 | responseSample)
## conXrPh: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## conXrPh:     (1 + context + responsePhoneme | partID) + (1 | contextSample) +
## conXrPh:     (1 | responseSample)
## trialXCon: answer ~ 1 + trialNumber.center + context * responsePhoneme +
## trialXCon:     trialNumber.center:context + (1 + context + responsePhoneme |
## trialXCon:     partID) + (1 | contextSample) + (1 | responseSample)
##           Df     AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m0        13 1054.12 1120.3 -514.06  1028.12
## trial     14 1051.32 1122.6 -511.66  1023.32  4.796      1  0.0285265 *
## context   15 1007.58 1083.9 -488.79   977.58 45.742      1  1.349e-11 ***
## rPhon     17  997.75 1084.3 -481.88   963.75 13.828      2  0.0009938 ***
## conXrPh   19 1000.41 1097.1 -481.20   962.41  1.344      2  0.5106922
## trialXCon 20  989.60 1091.4 -474.80   949.60 12.809      1  0.0003449 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fixed effects

Model estimates:

```r
finalModel = trialXCon
save(finalModel, file="../results/FinalModel.Rdat")
summary(finalModel)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: answer ~ 1 + trialNumber.center + context * responsePhoneme +
##     trialNumber.center:context + (1 + context + responsePhoneme |
##     partID) + (1 | contextSample) + (1 | responseSample)
##    Data: d
## Control: gcontrol
##
##      AIC      BIC   logLik deviance df.resid
##    989.6   1091.4   -474.8    949.6     1180
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2698 -0.2894 -0.1330  0.4098  5.5016
##
## Random effects:
##  Groups         Name                 Variance Std.Dev. Corr
##  responseSample (Intercept)          0.2651   0.5149
##  contextSample  (Intercept)          1.0303   1.0150
##  partID         (Intercept)          0.5637   0.7508
##                 contextIN            1.1434   1.0693   -0.67
##                 responsePhonemenone  0.4270   0.6535    0.34 -0.53
##                 responsePhonemeother 0.3828   0.6187   -0.45 -0.32  0.43
## Number of obs: 1200, groups:
## responseSample, 51; contextSample, 50; partID, 24
##
## Fixed effects:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      2.1405     0.3620   5.913 3.37e-09 ***
## trialNumber.center               0.7511     0.2008   3.740 0.000184 ***
## contextIN                       -4.4088     0.5107  -8.632  < 2e-16 ***
## responsePhonemenone             -1.3028     0.6326  -2.060 0.039440 *
## responsePhonemeother            -1.2298     0.3389  -3.629 0.000285 ***
## contextIN:responsePhonemenone   -0.4729     0.6773  -0.698 0.485085
## contextIN:responsePhonemeother   0.2271     0.4609   0.493 0.622169
## trialNumber.center:contextIN    -1.2321     0.3501  -3.519 0.000433 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##                     (Intr) trlNm. cntxIN rspnsPhnmn rspnsPhnmt
## trlNmbr.cnt          0.107
## contextIN           -0.675 -0.093
## rspnsPhnmnn         -0.247 -0.027  0.128
## rspnsPhnmth         -0.593 -0.105  0.278  0.323
## cntxtIN:rspnsPhnmn  0.213  0.011 -0.279 -0.255     -0.221
## cntxtIN:rspnsPhnmt  0.264  0.045 -0.383 -0.150     -0.441
## trlNmbr.:IN        -0.071 -0.574  0.093  0.019      0.071
##                    cntxtIN:rspnsPhnmn cntxtIN:rspnsPhnmt
## trlNmbr.cnt
## contextIN
## rspnsPhnmnn
## rspnsPhnmth
## cntxtIN:rspnsPhnmn
## cntxtIN:rspnsPhnmt  0.333
## trlNmbr.:IN         0.012              0.033
```

Relevel the response phoneme to see other comparisons:

```
d2 = d
d2$responsePhoneme = relevel(d2$responsePhoneme,"other")
fm2 = update(finalModel, data=d2)
summary(fm2)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: answer ~ 1 + trialNumber.center + context * responsePhoneme +
##     trialNumber.center:context + (1 + context + responsePhoneme |
##     partID) + (1 | contextSample) + (1 | responseSample)
##    Data: d2
## Control: gcontrol
##
##      AIC      BIC   logLik deviance df.resid
##    989.6   1091.4   -474.8    949.6     1180
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2698 -0.2894 -0.1330  0.4098  5.5015
##
## Random effects:
##  Groups         Name                Variance Std.Dev. Corr
##  responseSample (Intercept)         0.2651   0.5149
##  contextSample  (Intercept)         1.0302   1.0150
##  partID         (Intercept)         0.5273   0.7261
##                 contextIN           1.1434   1.0693   -0.97
##                 responsePhonemewh   0.3828   0.6187   -0.39  0.32
##                 responsePhonemenone 0.4654   0.6822    0.34 -0.21  0.50
## Number of obs: 1200, groups:
## responseSample, 51; contextSample, 50; partID, 24
##
## Fixed effects:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    0.91078    0.31669   2.876 0.004028 **
## trialNumber.center             0.75116    0.20084   3.740 0.000184 ***
## contextIN                     -4.18171    0.54110  -7.728 1.09e-14 ***
## responsePhonemewh              1.22977    0.33890   3.629 0.000285 ***
## responsePhonemenone           -0.07309    0.61348  -0.119 0.905169
## contextIN:responsePhonemewh   -0.22715    0.46086  -0.493 0.622101
## contextIN:responsePhonemenone -0.69994    0.68070  -1.028 0.303822
## trialNumber.center:contextIN  -1.23213    0.35014  -3.519 0.000433 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##                      (Intr) trlNm. cntxIN rspnsPhnmw rspnsPhnmn
## trlNmbr.cnt           0.011
## contextIN            -0.592 -0.050
## rspnsPhnmwh          -0.392  0.105  0.113
## rspnsPhnmnn          -0.151  0.030  0.056  0.219
## cntxtIN:rspnsPhnmw  0.170 -0.045 -0.490 -0.441     -0.089
## cntxtIN:rspnsPhnmn  0.121 -0.019 -0.312 -0.079     -0.200
## trlNmbr.:IN          -0.006 -0.574  0.115 -0.071     -0.020
##                      cntxtIN:rspnsPhnmw cntxtIN:rspnsPhnmn
## trlNmbr.cnt
## contextIN
## rspnsPhnmwh
## rspnsPhnmnn
## cntxtIN:rspnsPhnmw
## cntxtIN:rspnsPhnmn  0.346
## trlNmbr.:IN          -0.033             -0.010
```

```
write.csv(as.data.frame(summary(fm2)$coef),
          "../results/FinalModelCoefficients_relevel.csv")
```

Confidence intervals (through Wald method):

```
CI = confint(finalModel,parm="beta_", method="Wald")
cx = summary(finalModel)$coef
cx = cbind(cx[,1],CI,cx[,2:4])
cx2 = cx
for(i in 1:5){cx2[,i] = round(cx2[,i],3)}
cx2 = as.data.frame(cx2)
names(cx2)[1] = "estimate.logit"
cx2$esimate.odds = exp(cx2[,1])
cx2$esimate.odds.lower = exp(cx2[,2])
cx2$esimate.odds.upper = exp(cx2[,2])

cx2
```

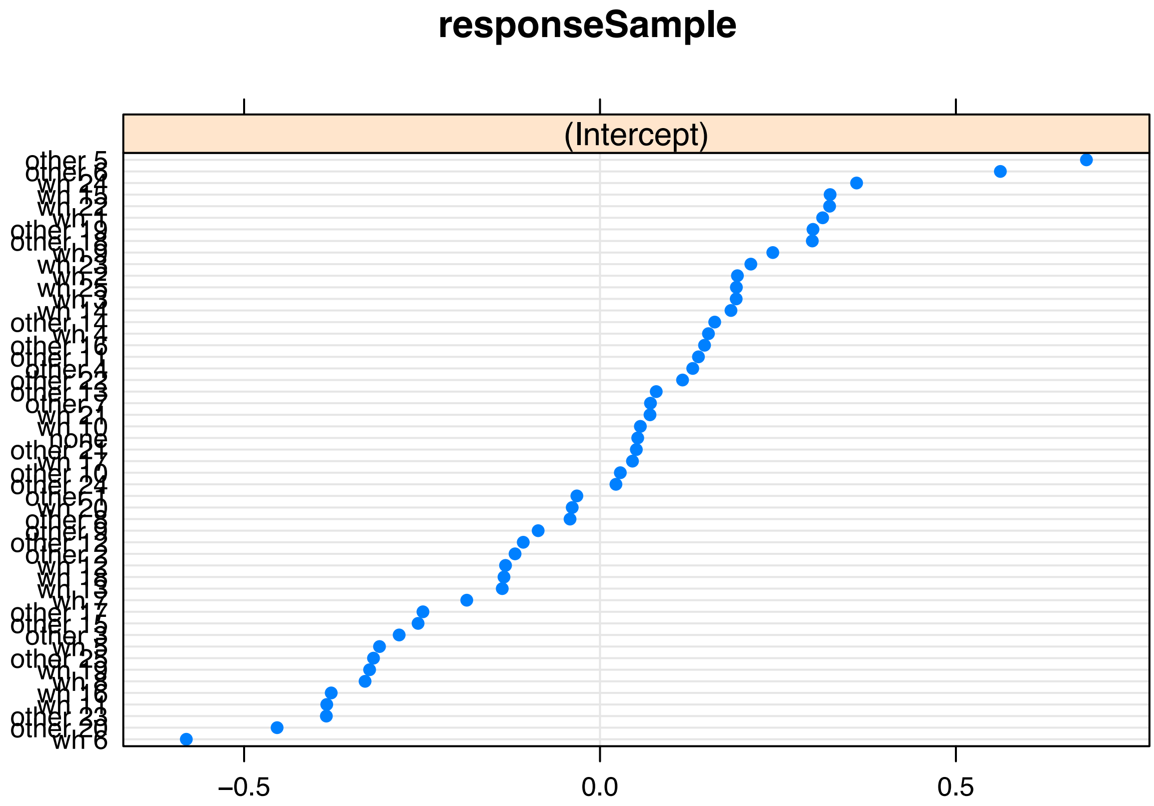
```
##                                 estimate.logit  2.5 % 97.5 % Std. Error
## (Intercept)                             2.141  1.431  2.850      0.362
## trialNumber.center                      0.751  0.358  1.145      0.201
## contextIN                              -4.409 -5.410 -3.408      0.511
## responsePhonemenone                    -1.303 -2.543 -0.063      0.633
## responsePhonemeother                   -1.230 -1.894 -0.566      0.339
## contextIN:responsePhonemenone          -0.473 -1.800  0.855      0.677
## contextIN:responsePhonemeother          0.227 -0.676  1.130      0.461
## trialNumber.center:contextIN           -1.232 -1.918 -0.546      0.350
##                                 z value     Pr(>|z|) esimate.odds
## (Intercept)                      5.913 3.366019e-09   8.50794132
## trialNumber.center               3.740 1.840016e-04   2.11911808
## contextIN                       -8.632 6.007100e-18   0.01216734
## responsePhonemenone             -2.060 3.943976e-02   0.27171542
## responsePhonemeother            -3.629 2.848789e-04   0.29229258
## contextIN:responsePhonemenone   -0.698 4.850847e-01   0.62313007
## contextIN:responsePhonemeother   0.493 6.221693e-01   1.25482987
## trialNumber.center:contextIN    -3.519 4.333312e-04   0.29170858
##                                esimate.odds.lower esimate.odds.upper
## (Intercept)                            4.18287998         4.18287998
## trialNumber.center                     1.43046562         1.43046562
## contextIN                              0.00447164         0.00447164
## responsePhonemenone                    0.07863016         0.07863016
## responsePhonemeother                   0.15046873         0.15046873
## contextIN:responsePhonemenone          0.16529889         0.16529889
## contextIN:responsePhonemeother         0.50864752         0.50864752
## trialNumber.center:contextIN           0.14690047         0.14690047
```

```
write.csv(cx, "../results/FinalModelCoefficients.csv")
```

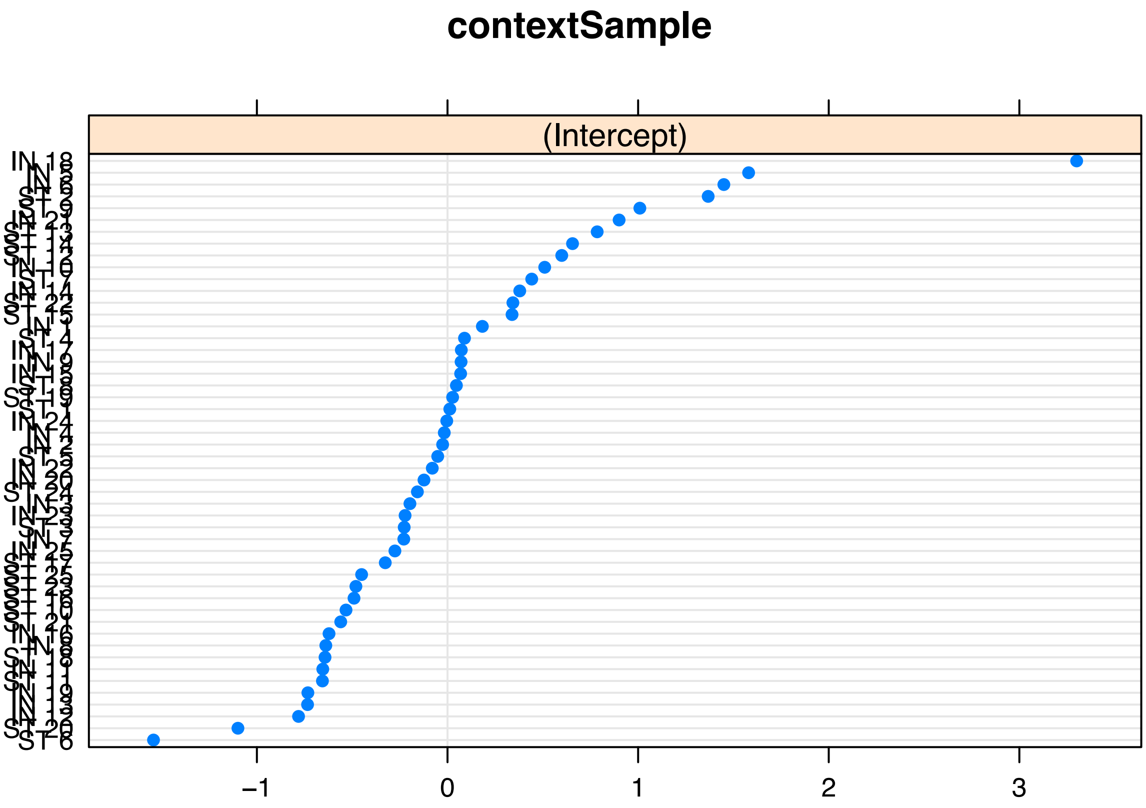
Random effects

```r
dotplot(ranef(finalModel))
```

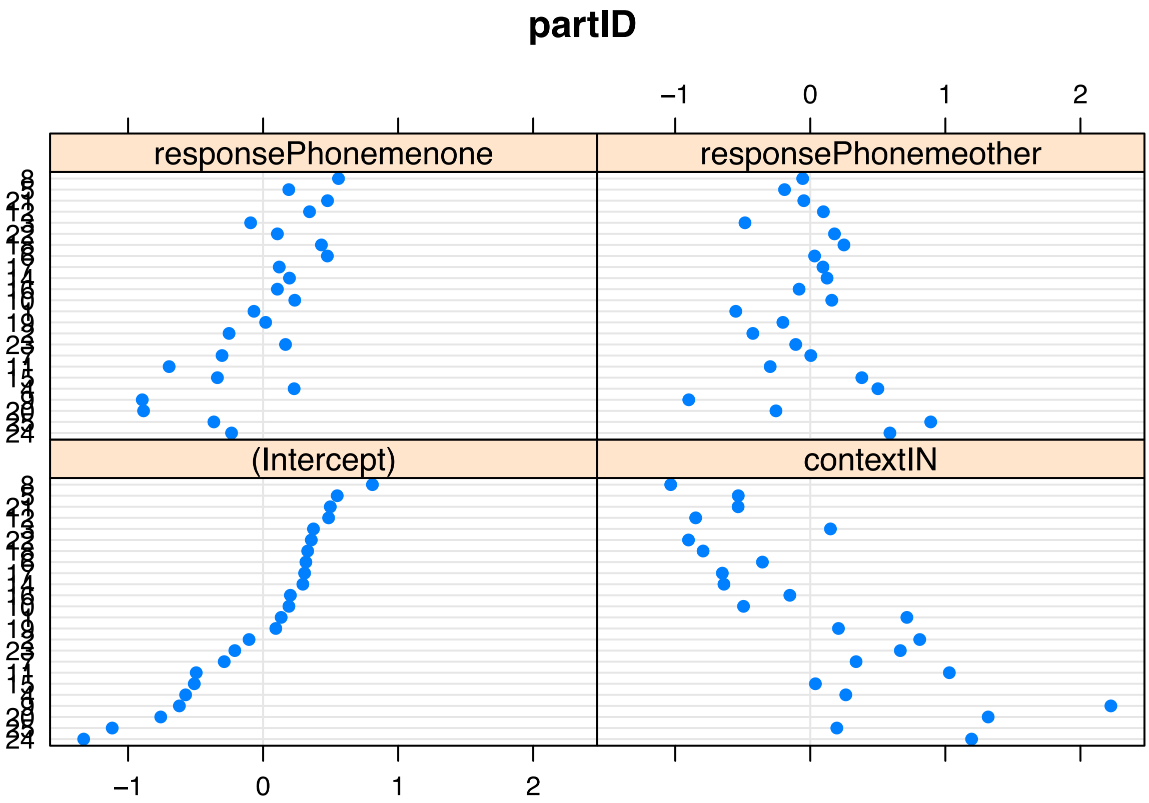
```
## $responseSample
```



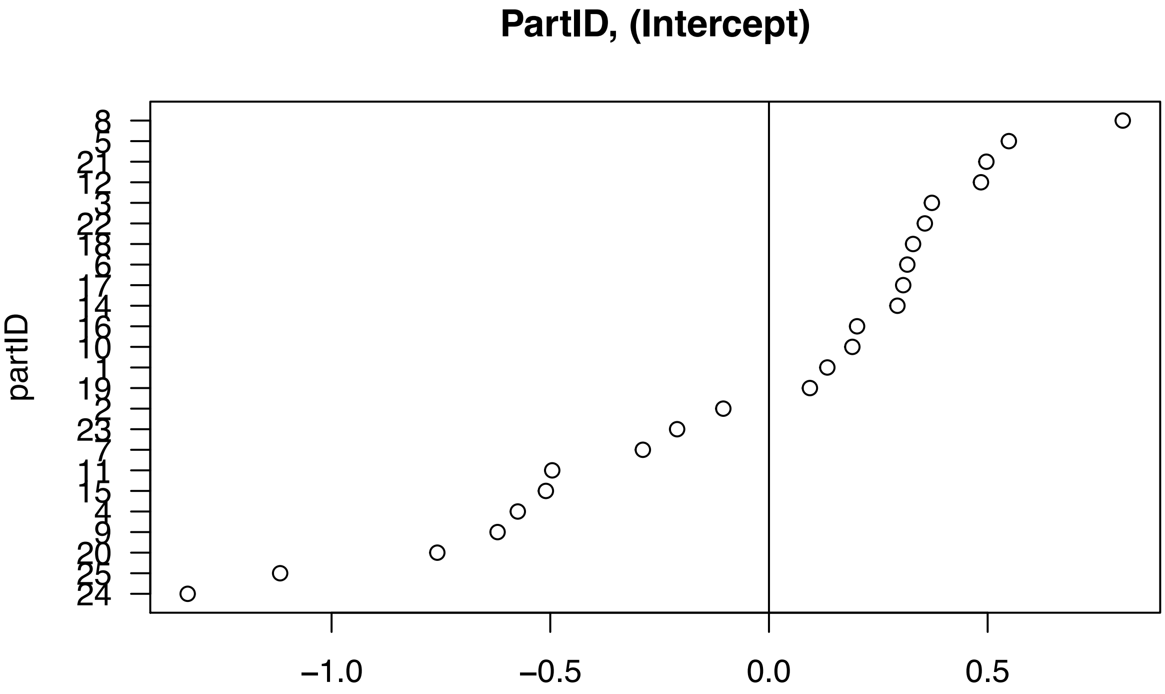
```
##
## $contextSample
```

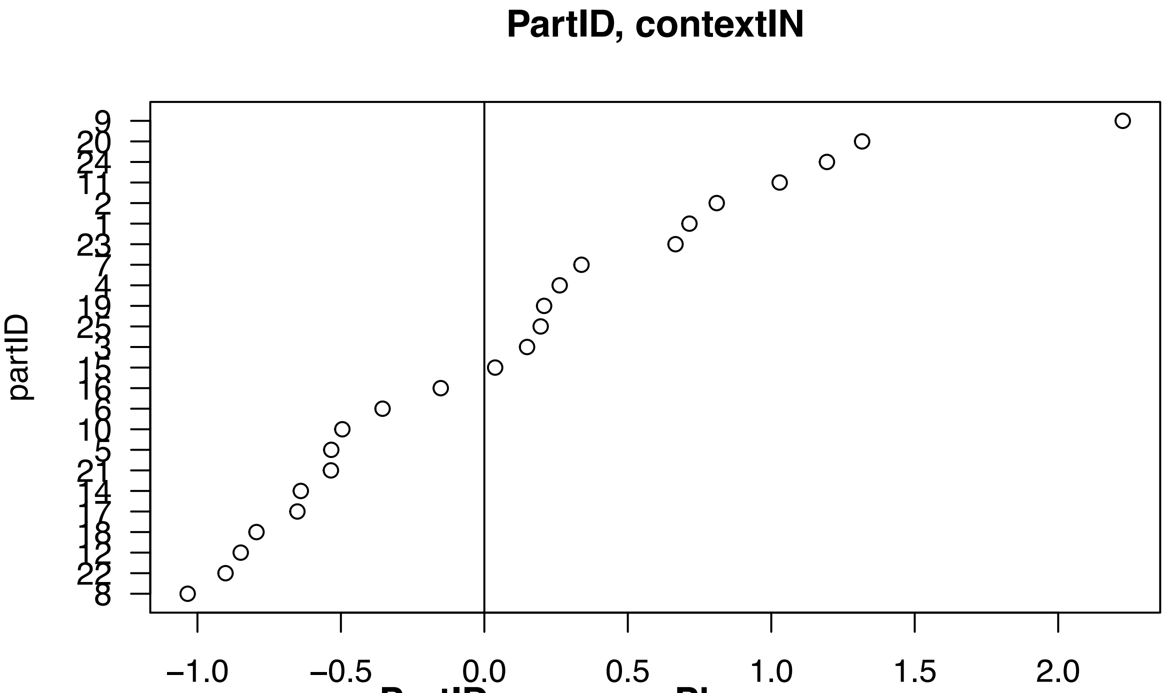

contextSample
(Intercept)
IN 18
ST 6
−1
0
1
2
3

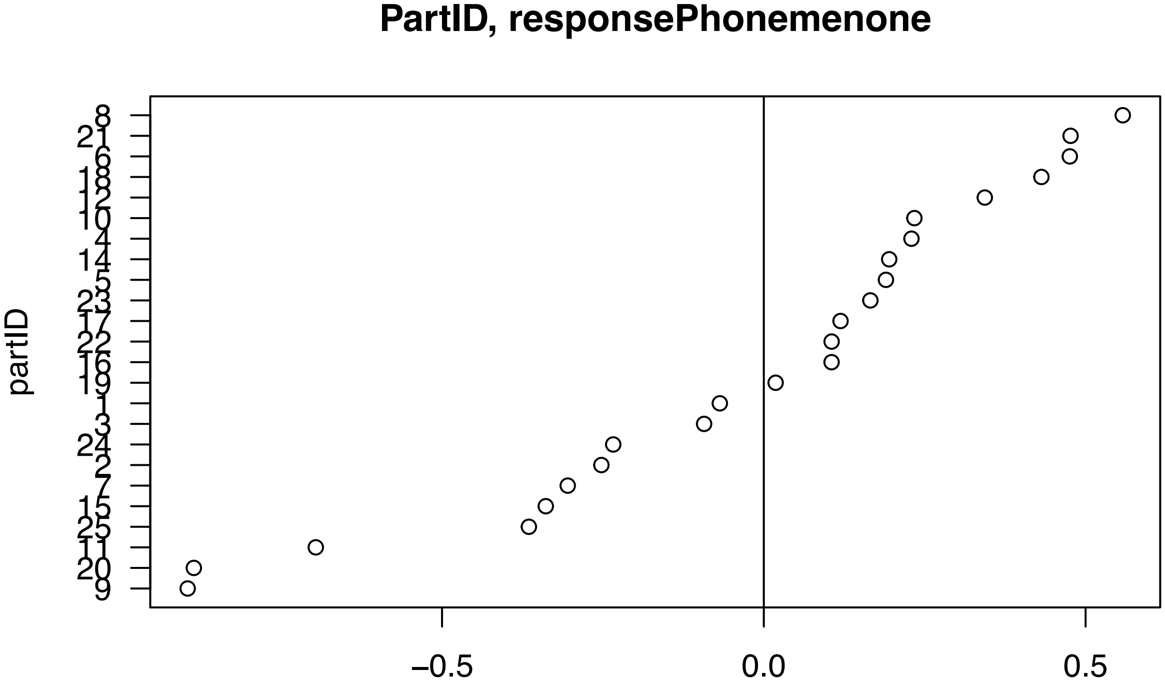
```
##
## $partID
```

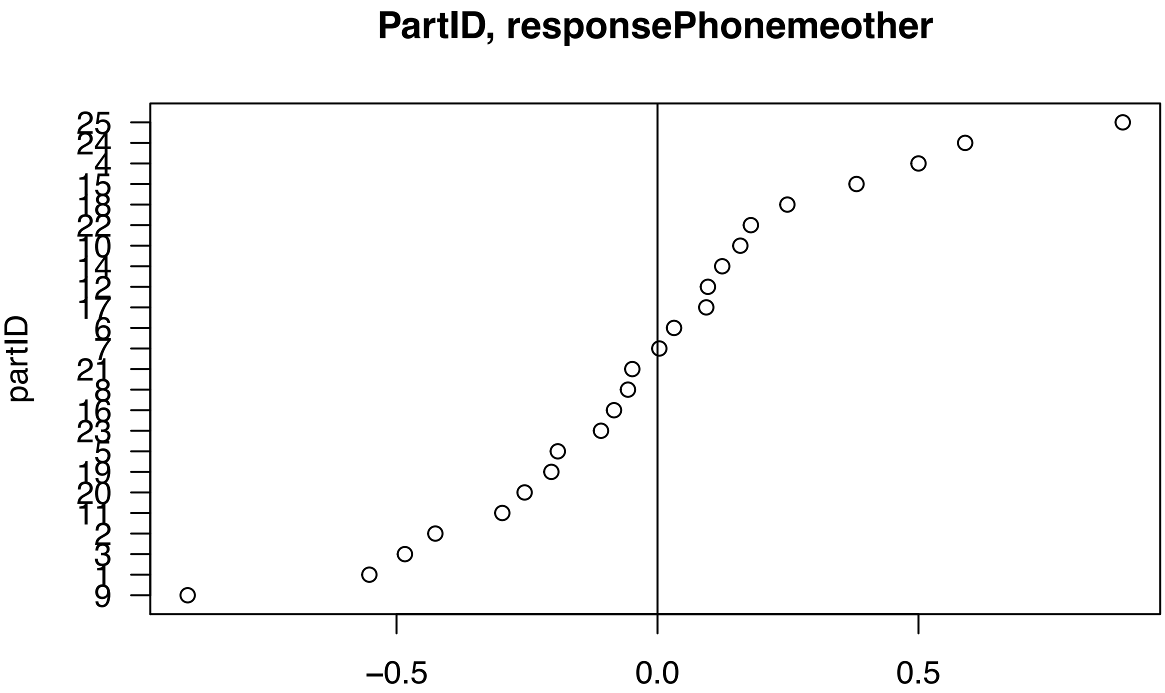

partID
responsePhonemenone
responsePhonemeother
(Intercept)
contextIN
8
24
−1
0
1
2

```
x = as.data.frame(ranef(finalModel)$partID)
rownames(x) = rownames(ranef(finalModel)$partID)
for(i in 1:ncol(x)){
  x = x[order(x[,i]),]
  plot(x[,i],1:nrow(x),
       main = paste("PartID,",colnames(x)[i]),
       yaxt='n',
       ylab='partID', xlab='')
  abline(v=0)
  axis(2,at=1:nrow(x),labels=rownames(x), las=2)
}
```



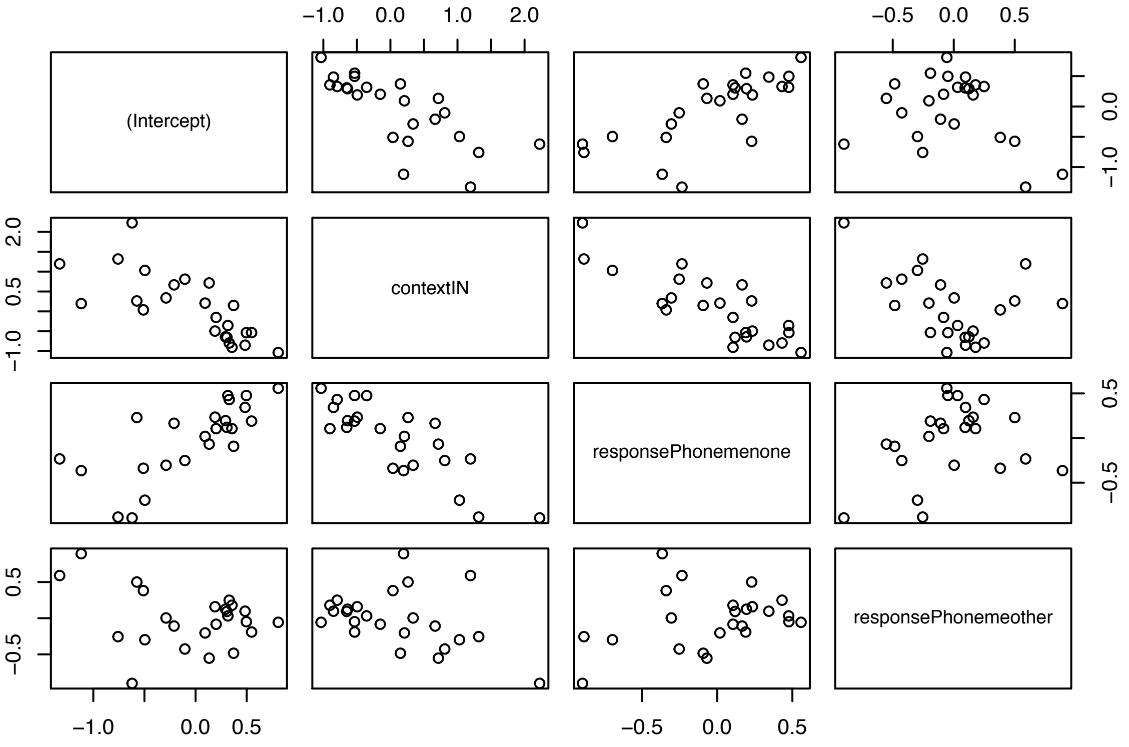

PartID, contextIN


PartID, responsePhonemenone



Correlation between random effects for poarticipants:

```r
plot(ranef(finalModel)$partID)
```



31

Summary

Here is a summary of the main results:

There was a significant main effect of context (log likelihood difference = 23 , df = 1 , Chi Squared = 45.74 , p = 1.3e-11).

There was a significant main effect of phoneme (log likelihood difference = 6.9 , df = 2 , Chi Squared = 13.83 , p = 0.00099).

There was no significant interaction between context and phoneme (log likelihood difference = 0.67 , df = 2 , Chi Squared = 1.34 , p = 0.51).

There was a significant main effect of trial (log likelihood difference = 2.4 , df = 1 , Chi Squared = 4.8 , p = 0.029).

Work out model esimates for probabilities in each condition:

```
newD = data.frame(context=c("IN","IN","IN","ST","ST",'ST'),
         responsePhoneme = c("none","other","wh",'none','other','wh'),
         trialNumber.center = c(0,0,0,0,0,0))
rownames(newD) = c("IN + none", "IN + other", "IN + wh",
                   "ST + none", "ST + other", "ST + wh")
prx = predict(finalModel,re.form=NA,newdata=newD)

t(t(logit2per(prx)))
```

```
##                  [,1]
## IN + none  0.01722517
## IN + other 0.03658105
## IN + wh    0.09378154
## ST + none  0.69797199
## ST + other 0.71315356
## ST + wh    0.89477791
```

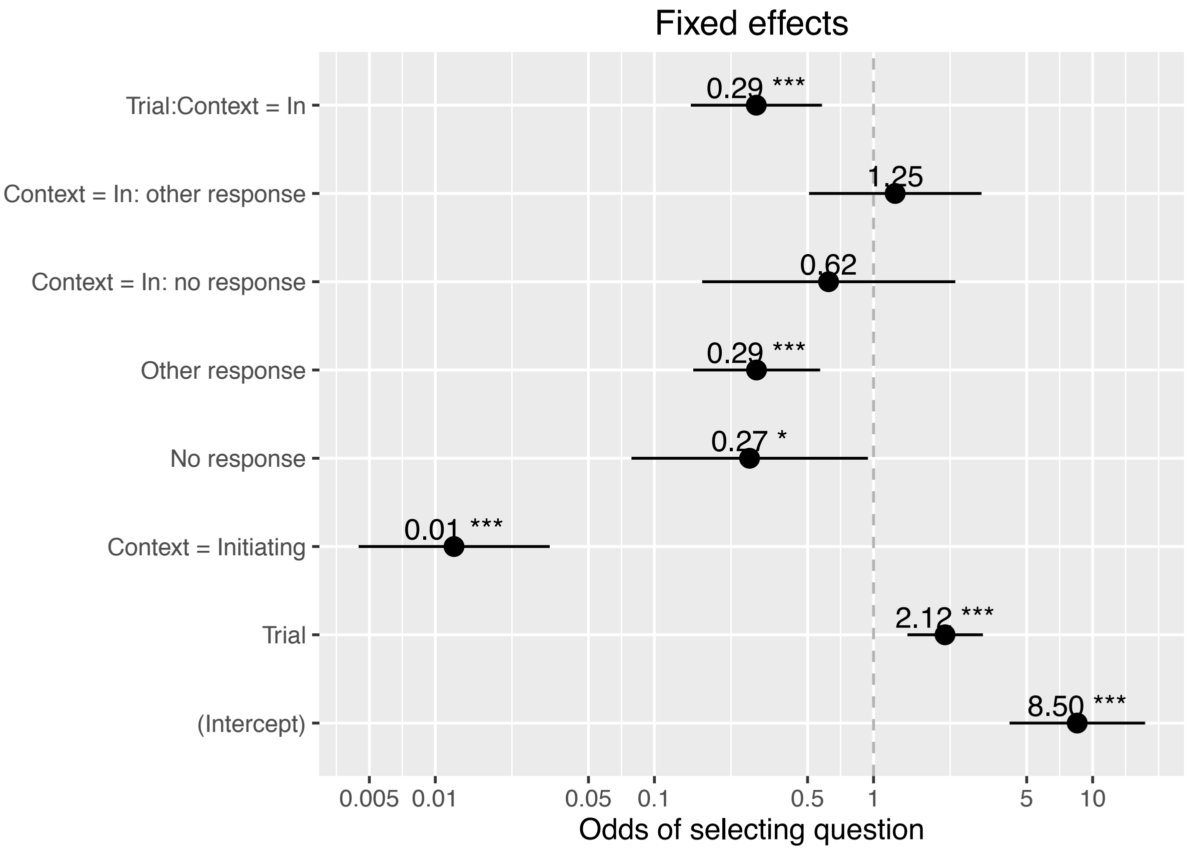
Plots

Fixed effects estimates:

```
feLabels = matrix(c(
  "(Intercept)"             ,"Intercept"      , NA,
  "trialNumber.center", "Trial",NA,
  "contextST", "Context = Statement", "context",
  "contextIN", "Context = Initiating", "context",
  "responsePhonemenone", "No response", 'rPhon',
  "responsePhonemewh", "wh response", 'rPhon',
  "responsePhonemeother","Other response", 'rPhon',
  "contextIN:responsePhonemenone", "Context = In: no response", "conXrPh",
  "contextIN:responsePhonemewh", "Context = In: wh response", "conXrPh",
  "contextIN:responsePhonemeother", "Context = In: other response", "conXrPh",
  "trialNumber.center:contextIN","Trial:Context = In",'trialXCon'
), ncol=3, byrow = T)

feLabels2 = as.vector(feLabels[match(names(fixef(finalModel)),feLabels[,1]),2])
```

```
sjp.glmer(finalModel,'fe',
          show.intercept = T,
          geom.colors = c(1,1),
          axis.title = "Odds of selecting question",
          y.offset = 0.2,
          axis.labels = feLabels2[2:length(feLabels2)]
)
```

```
## Warning: Deprecated, use tibble::rownames_to_column() instead.
```

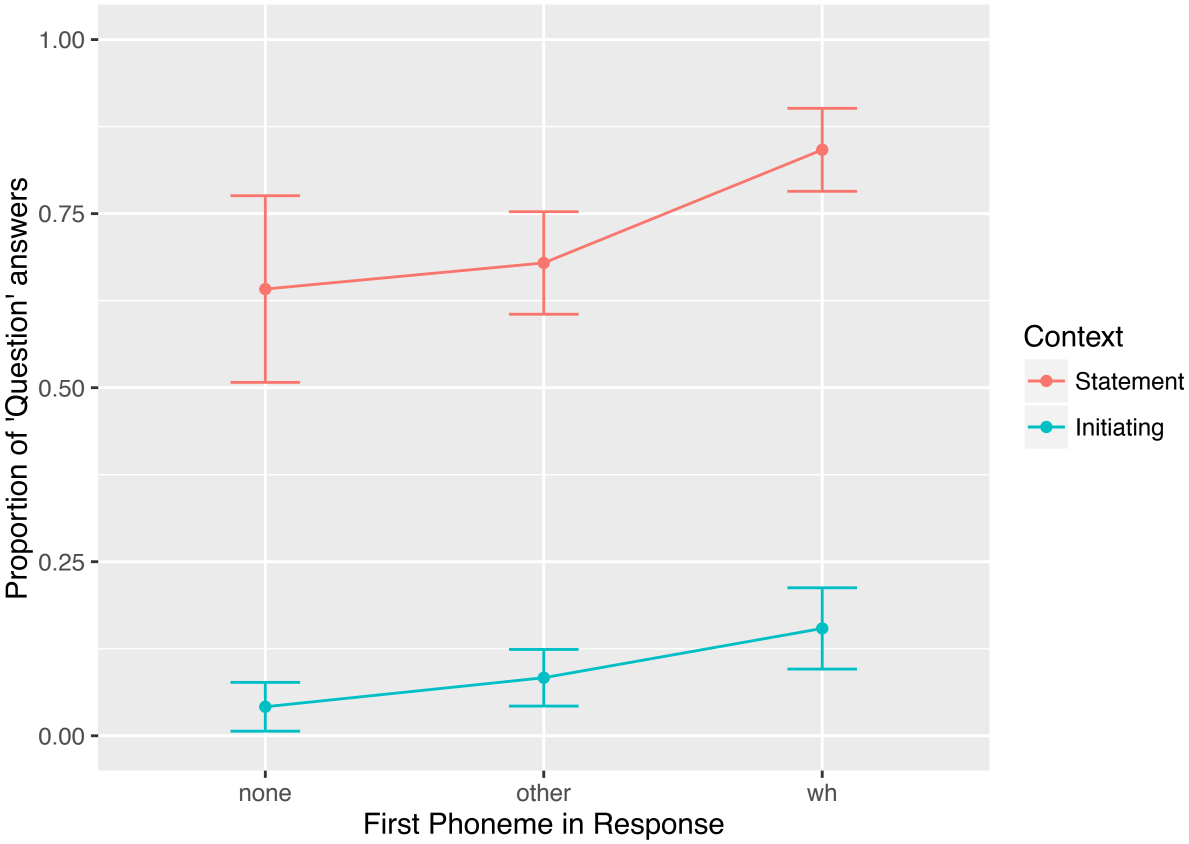

Fixed effects
Trial:Context = In
Context = In: other response
Context = In: no response
Other response
No response
Context = Initiating
Trial
(Intercept)
0.29 ***
1.25
0.62
0.29 ***
0.27 *
0.01 ***
2.12 ***
8.50 ***
0.005
0.01
0.05
0.1
0.5
1
5
10
Odds of selecting question

Raw data plots

```
d$responsePhoneme2 = relevel(relevel(d$responsePhoneme,'other'),'none')

sumStats = group_by(d, partID ,context,responsePhoneme2 ) %>%
              summarise(mean =mean(answer) )

sumStats2 = summarySE(sumStats, measurevar="mean", groupvars=c("context","responsePhoneme2"))
sumStats2$upper = sumStats2$mean + sumStats2$ci
sumStats2$lower = sumStats2$mean - sumStats2$ci

sumStats2
```

```
##   context responsePhoneme2  N       mean         sd         se         ci
## 1      ST             none 24 0.64166667 0.31748559 0.06480648 0.13406241
## 2      ST            other 24 0.67916667 0.17440375 0.03560002 0.07364424
## 3      ST               wh 24 0.84166667 0.14116493 0.02881517 0.05960872
## 4      IN             none 24 0.04166667 0.08297022 0.01693623 0.03503525
## 5      IN            other 24 0.08333333 0.09630868 0.01965893 0.04066759
## 6      IN               wh 24 0.15416667 0.13824731 0.02821961 0.05837672
##        upper       lower
## 1 0.77572907 0.507604259
## 2 0.75281091 0.605522423
## 3 0.90127539 0.782057946
## 4 0.07670192 0.006631414
## 5 0.12400092 0.042665743
## 6 0.21254339 0.095789947
```

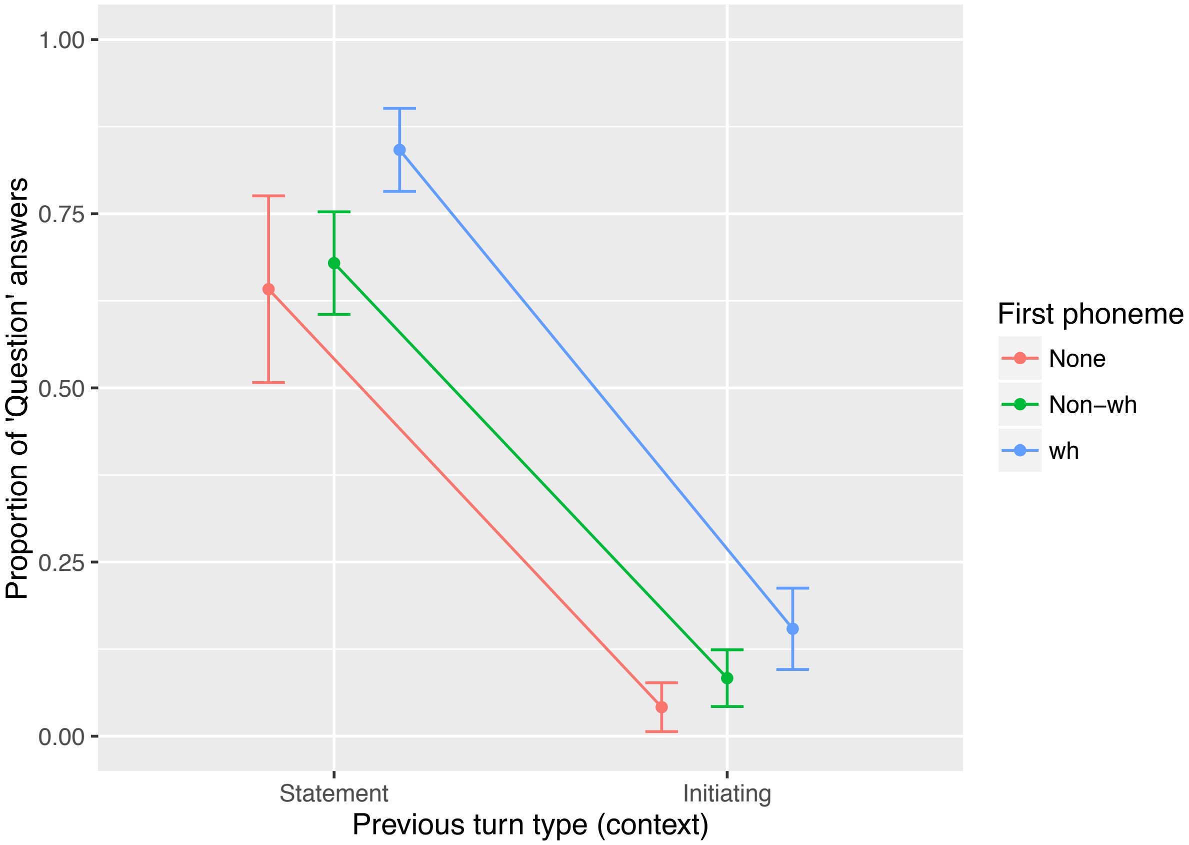
```
dodge <- position_dodge(width=0.5)

main.plot <- ggplot(sumStats2,
    aes(x = responsePhoneme2, y = mean, colour=context)) +
  geom_point() + geom_line(aes(group=context)) +
  geom_errorbar(aes(ymax=mean+ci, ymin=mean-ci), width=0.25) +
  xlab("First Phoneme in Response") +
  ylab("Proportion of 'Question' answers") +
  coord_cartesian(ylim=c(0,1)) +
  scale_color_discrete(breaks=c("ST","IN"),
                       labels=c("Statement","Initiating"),
                       name="Context")

main.plot
```



```
pdf("../results/graphs/PropQResponses_by_firstPhoneme_withPartCI.pdf",
    width = 4, height=3)
main.plot
dev.off()
```

```
## pdf
##   2
```

```
main.plot2 <- ggplot(sumStats2,
    aes(x = context, y = mean, colour=responsePhoneme2)) +
  geom_point(position=dodge) + geom_line(aes(group=responsePhoneme2), position=dodge) +
  geom_errorbar(aes(ymax=mean+ci, ymin=mean-ci), width=0.25, position=dodge) +
  xlab("Previous turn type (context)") +
  ylab("Proportion of 'Question' answers") +
  coord_cartesian(ylim=c(0,1)) +
  scale_color_discrete(breaks=c("none","other",'wh'),
                       labels=c("None","Non-wh","wh"),
                       name="First phoneme") +
  scale_x_discrete(breaks=c("ST", "IN"),
                     labels=c("Statement", "Initiating"))

main.plot2
```



```
pdf("../results/graphs/PropQResponses_by_context_withPartCI.pdf",
    width = 4, height=3)
main.plot2
dev.off()
```

```
## pdf
##   2
```

Predicting response type

```
d2 = d[d$responsePhoneme!="none",]

table(d2$answer,d2$responseType)
```

```
##
##         none other   Q
##   FALSE    0   271 267
##   TRUE     0   209 213
```

```
d2$correct = "Correct"
d2$correct[!d2$answer & d2$responseType=="Q"] = "Incorrect"
d2$correct[d2$answer & d2$responseType=="other"] = "Incorrect"
# number of "correct" responses
table(d2$correct)
```

```
##
##   Correct Incorrect
##       484       476
```

```
mOT = glmer(answer ~ 1 + context*responsePhoneme +
          (1 + context | partID) +
          (1 | contextSample) +
          (1 | responseSample) ,
           data = d2,
           family = binomial,
          control=gcontrol)

respT = glmer(answer ~ 1 + context*responsePhoneme +
                responseType +
          (1 + context | partID) +
          (1 | contextSample) +
          (1 | responseSample) ,
           data = d2,
           family = binomial,
          control=gcontrol)

respTXco = glmer(answer ~ 1 + context*responsePhoneme +
                responseType*context +
          (1 + context | partID) +
          (1 | contextSample) +
          (1 | responseSample) ,
           data = d2,
           family = binomial,
          control=gcontrol)

respTXrp = glmer(answer ~ 1 + context*responsePhoneme +
                responseType*context +
                + responseType: responsePhoneme +
          (1 + context | partID) +
          (1 | contextSample) +
          (1 | responseSample) ,
           data = d2,
           family = binomial,
          control=gcontrol)

rTXcoXrp = glmer(answer ~ 1 + context*responsePhoneme +
                responseType*context *responsePhoneme +
          (1 + context | partID) +
          (1 | contextSample) +
          (1 | responseSample) ,
           data = d2,
           family = binomial,
          control=gcontrol)

anova(mOT, respT, respTXco, respTXrp, rTXcoXrp)
```

```
## Data: d2
## Models:
## mOT: answer ~ 1 + context * responsePhoneme + (1 + context | partID) +
## mOT:     (1 | contextSample) + (1 | responseSample)
## respT: answer ~ 1 + context * responsePhoneme + responseType + (1 +
## respT:     context | partID) + (1 | contextSample) + (1 | responseSample)
## respTXco: answer ~ 1 + context * responsePhoneme + responseType * context +
## respTXco:     (1 + context | partID) + (1 | contextSample) + (1 | responseSample)
## respTXrp: answer ~ 1 + context * responsePhoneme + responseType * context +
## respTXrp:     +responseType:responsePhoneme + (1 + context | partID) +
## respTXrp:     (1 | contextSample) + (1 | responseSample)
## rTXcoXrp: answer ~ 1 + context * responsePhoneme + responseType * context *
## rTXcoXrp:     responsePhoneme + (1 + context | partID) + (1 | contextSample) +
## rTXcoXrp:     (1 | responseSample)
##          Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mOT       9 819.68 863.48 -400.84   801.68
## respT    10 821.57 870.24 -400.79   801.57 0.1050      1     0.7459
## respTXco 11 823.40 876.94 -400.70   801.40 0.1702      1     0.6800
## respTXrp 12 825.40 883.80 -400.70   801.40 0.0080      1     0.9288
## rTXcoXrp 13 826.31 889.58 -400.16   800.31 1.0823      1     0.2982
```

No effects of actual response type.

Note on different optimisers

The nlminb optimiser was used instead of the default bobyqa and Nelder-Mead optimisers. The deafult settings caused convergence problems for the model with the interaction between responsePhoneme and context, probably due to the lack of variation in some of the conditions. Several other optimisers were tried, and the main results remained qualitatively the same (main effect of context, main effect of responsePhoneme, no interaction).

For the `conXrPh` model above, 7 different optimiser settings were tried (following this approach), all but nlminbw produced convergence warnings. The optimisers returned very similar log liklihoods:

- Nelder_Mead -482.3682
- bobyqa -481.4402
- nloptwrap.NLOPT_LN_NELDERMEAD -481.2140
- nloptwrap.NLOPT_LN_BOBYQA -481.2140
- nmkbw -481.2042
- optimx.L-BFGS-B -481.2041
- nlminbw -481.2041

Below is a summary of the estimates for different fixed effects for different optimisers for the `conXrPh` model above. The estimates vary little between the runs.

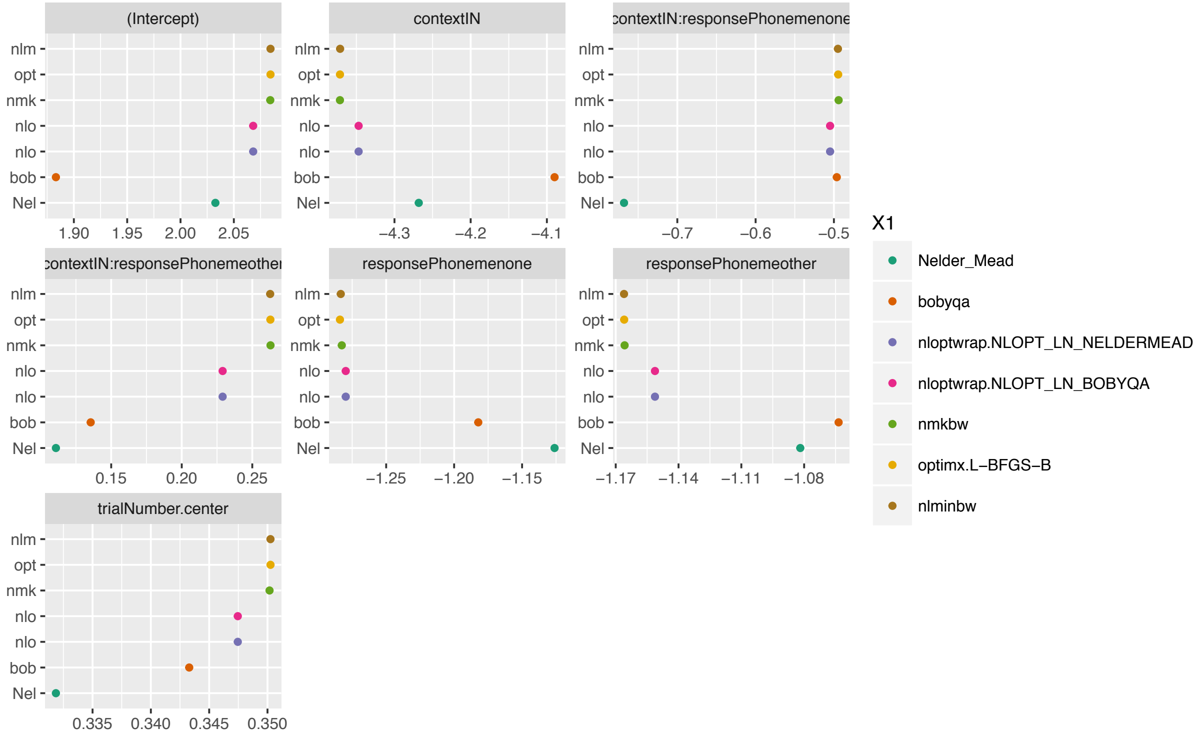


Figure 1: DifferentOptimizerResults