Cognition 203 (2020) 104347

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Neural correlates of turn-taking in the wild: Response planning starts early in free interviews

Sara Bögels*

Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, the Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

ARTICLE INFO

Keywords:
Conversation
Turn-taking
Pragmatics
Language production
EEG

ABSTRACT

Conversation is generally characterized by smooth transitions between turns, with only very short gaps. This entails that responders often begin planning their response before the ongoing turn is finished. However, controversy exists about whether they start planning as early as they can, to make sure they respond on time, or as late as possible, to minimize the overlap between comprehension and production planning. Two earlier EEG studies have found neural correlates of response planning (positive ERP and alpha decrease) as soon as listeners could start planning their response, already midway through the current turn. However, in these studies, the questions asked were highly controlled with respect to the position where planning could start (e.g., very early) and required short and easy responses. The present study measured participants' EEG while an experimenter interviewed them in a spontaneous interaction. Coding the questions in the interviews showed that, under these natural circumstances, listeners can, in principle, start planning a response relatively early, on average after only about one third of the question has passed. Furthermore, ERP results showed a large positivity, interpreted before as an early neural signature of response planning, starting about half a second after the start of the word that allowed listeners to start planning a response. A second neural signature of response planning, an alpha decrease, was not replicated as reliably. In conclusion, listeners appear to start planning their response early during the ongoing turn, also under natural circumstances, presumably in order to keep the gap between turns short and respond on time. These results have several important implications for turn-taking theories, which need to explain how interlocutors deal with the overlap between comprehension and production, how they manage to come in on time, and the sources that lead to variability between conversationalists in the start of planning.

1. Introduction

Conversations with others are one of our everyday activities and they appear effortless. However, multiple cognitive processes have to be coordinated during conversation, such as understanding what the other person wants to say and planning and producing your own turn. That conversation is indeed a remarkable feat in psycholinguistic terms is made apparent by the following puzzle. Conversational corpora show that the amount of time between two turns (about 200 ms in the most frequent case, e.g., Heldner & Edlund, 2010; Levinson & Torreira, 2015) appears to be far too small for the upcoming speaker to plan even one word (which takes minimally 600 ms in picture naming paradigms, e.g., Indefrey, 2011; Indefrey & Levelt, 2004). Thus, the two processes of language understanding and production planning have to be at least partly overlapping. As an example (taken from the data used in the present study, see Table 1 for more examples), consider person B, who has just told person A that she does not have a side job next to her study. Person A could then ask:

Example 1 (original Dutch with English translations):

A: *Zou je dat wel willen, of vind je het wel prima zo*?

"Would you actually want that, or do you find it fine like this?"

[158 ms]

B: *Nee ik vind het prima zo, ik heb het eh druk zat.*

"No I find it fine like this, I am uh busy enough."

In order for person B to answer this question so quickly, she needs to start planning her response while she is still listening to person A's

* Donders Institute for Brain, Cognition and Behaviour, P.O. Box 9104, 6500 HE Nijmegen, the Netherlands.
E-mail address: s.bogels@donders.ru.nl.

https://doi.org/10.1016/j.cognition.2020.104347
Received 18 July 2019; Received in revised form 20 May 2020; Accepted 30 May 2020

Table 1
Examples of different types of questions asked by the interviewers. Questions come from different interviews. Questions and answers are given in the original Dutch followed by English translations. Answer words in the question are indicated in bold, control words are underlined. Response time is measured from question offset. # indicates at what position the question occurred in the interview.

Type	Context	Question	Response	Response time	#
Polar, scripted, early	- (first question)	*Dus je bent **student** hier op de Radboud Universiteit?* 'So you are a **student** here at the Radboud University?'	*Hmhm.* 'Hmhm.'	217 ms	1
Open, non-scripted, late	Participant has just answered some informational questions about her study (Pedagogics).	*En hoe hoe **bevalt** de studie?* 'And how how do you **like** the study?'	*Ja ik vind het een eh een leuke studie, ik had wel verwacht dat ie wat moeilijker zou zijn.* 'Yes, I think it is an uh nice study, I did expect it to be more difficult.'	653 ms	6
Polar, non-scripted, late	Participant told interviewer she usually bikes to University.	*En ook met die **sneeuw** nog?* 'And also with that **snow** still?'	*Eh ja, de fietspaden zijn nu wel vrij dus dat eh…* 'Uh, yes, the bike paths are free now, so that uh…'	800 ms	24
Polar, non-scripted, early	Participant told the interviewer that she will move out of her parents' home soon.	*Want het begon **vervelend** te worden bij je ouders thuis of vond je het reizen vooral vervelend?* 'Because it started to get **annoying** with your parents at home or did you mainly dislike the travelling?'	*Nee 't was helemaal niet vervelend thuis maar 't reizen vond ik wel eh, vervelend (…).* 'No, it was not annoying at home at all, but the travelling I did uh, dislike (…).'	− 227 ms (overlap)	22

question. Such overlap may be problematic given that language production (planning) and comprehension rely on largely the same cognitive and neuronal architecture (e.g., Segaert, Menenti, Weber, Petersson, & Hagoort, 2012) and both processes have been shown to require central attention (e.g., Kubose et al., 2006; Shitova, Roelofs, Coughler, & Schriefers, 2017). Indeed, dual tasking with two linguistic tasks leads to more interference than when one of the tasks is non-linguistic (Fairs, Bögels, & Meyer, 2018), and production planning in turn-taking happening in overlap with concurrent speech input leads to increased processing load as compared to planning 'in the clear', as measured by pupillary responses (Barthel & Sauppe, 2019). Thus, one might predict that interlocutors attempt to keep the amount of overlap between comprehension and production minimal, that is, start planning as late as possible. This idea will hereafter be referred to as the 'late planning hypothesis'. In our example, person B should minimize the overlap between comprehension and production planning and thus only start planning her response when the end of the question is approaching (for example during listening to the last three words: 'fine like this'). This hypothesis thus assumes that listeners can accurately estimate the end of the current speaker's turn at least some hundreds of milliseconds in advance, to be able to launch planning exactly on time. While some anticipation indeed appears possible (Magyari, Bastiaansen, de Ruiter, & Levinson, 2014; Magyari & De Ruiter, 2012), other findings suggest that anticipation in turn-taking does not help to estimate turn-ends (Corps, Crossley, Gambi, & Pickering, 2018).

While overlap might thus be costly, there are also indications that it is very important for turns to arrive 'on time', that is, early enough after the previous turn. First, the fast timing of turn-taking appears to be universal and is found in diverse languages over the world (Stivers et al., 2009). The turn-taking system even has been argued to be 'older' than the language system, both ontogenetically and phylogenetically (Levinson, 2016). Second, when responses are late (i.e., given after more than about 800 ms) they are more likely to be 'dispreferred', that is, going against the expectation expressed in the preceding turn (Kendrick & Torreira, 2015; Levinson, 1983), for example in the case of a rejection of an invitation. Listeners indeed come to expect a dispreferred response more after a longer pause (Bögels, Kendrick, & Levinson, 2015, 2019). This renders it relevant to respond on time (if the response is preferred), otherwise your interlocutor might start to get the wrong impression. In example 1 above, if B would take too long to respond, A might think that B did not understand or hear the question, or would rather not answer it. If indeed coming in on time is so important, one might argue that planning should actually start as early as possible to make sure that one is ready to speak at the moment when it is necessary. The larger amount of overlap between comprehension and production planning might then be taken for granted. This will be termed the 'early planning hypothesis'. Applying this hypothesis to example 1, upon recognizing the word 'want', person B probably has a good idea of the type of answer A expects and can start planning her response. Note that some amount of prediction or anticipation of content on the basis of the sentence and wider context presumably plays a role here. Because B knows she has just told A that she does not have a side job, she can anticipate at 'want' that B is asking about wanting a side job rather than something else. Furthermore, the early planning hypothesis requires a second mechanism to determine when to actually start speaking once a turn has been planned, in order not to start in overlap with the ongoing turn. Here, final (prosodic) cues might play an important role (Bögels & Torreira, 2015; see also Section 4.4.1 in the Discussion below).

These two opposing hypotheses described above, late and early planning, are focused on minimizing overlap versus making sure one responds on time, respectively (see, e.g., Bögels & Levinson, 2017; Corps, Gambi, & Pickering, 2018 for further reviews of the two positions). When trying to disentangle these two hypotheses, one should first consider what 'as soon as possible' means. In the example above, the word 'want' (the position at which early planning could start) is

followed by several more words before the end of the turn, so the two hypotheses seem to make quite different predictions here. However, that may not be the case in conversation generally. If the early planning position generally occurs very close to the end of the turn, the distinction between 'as early' and 'as late' as possible might become an artificial one. Most of the recent experiments on this topic have manipulated turns carefully to artificially create a clear distinction between 'early' and 'late' planning. Some of the evidence for the early planning hypothesis comes from EEG studies showing neural signatures of early planning (Bögels, Casillas, & Levinson, 2018; Bögels, Magyari, & Levinson, 2015; see below). The present study attempts to extend this research into more natural, ecologically valid circumstances. It investigates natural interactions in the form of short interviews to see (1) when is the earliest time that listeners *can* start planning in conversation, and (2) whether previously found neural signatures of early response planning under controlled circumstances can be replicated in natural dialogue.

Recent literature presents mixed findings on the question whether production planning starts as early or as late as possible during turn-taking. At least two studies' findings appear to point to the late planning hypothesis. The first one asked participants to have a natural conversation while tracking a target on a computer screen with their computer mouse (Boiteau, Malone, Peters, & Almor, 2014). They found that performance in the tracking task decreased most during and just before speaking, suggesting that listeners start planning their turn only just before they start speaking. However, note that this study did not manipulate or measure when interlocutors could start planning their response. The second study (Sjerps & Meyer, 2015) presented participants with two rows of pictures. They first heard a recorded voice name one of the rows and they were asked to name the other row when the recorded voice was finished. Participants simultaneously performed a finger-tapping task, which again deteriorated only shortly (about 500 ms) before the recorded speech ended (and the participant's speech started). These two studies appear to suggest that at least the attention-demanding aspects of production planning start only shortly before the offset of the previous turn. However, since in the Boiteau et al. (2014) study, no measure was taken of when participants could start planning, it is uncertain whether they started as early as possible or not. Conversely, in Sjerps and Meyer (2015), participants could in principle start planning as soon as they recognized the first word of the recorded voice, because that enabled them to know which of the two rows they had to describe subsequently. However, participants might have felt compelled to look at the pictures while they were named, as they were present on the screen. This may have precluded participants to look at their own pictures and start planning. Also, the turn-taking task used in that experiment was quite different from conversation, because the participant's response was not semantically contingent on what the recorded voice said (except for knowing which row of objects to name).

A few studies have also presented findings pointing to the contrary conclusion, that production planning starts early. One study (Barthel, Sauppe, Levinson, & Meyer, 2016) asked participants to listen to a confederate who named some of the objects on the screen, after which they themselves had to name the remaining objects (note that their response was thus minimally contingent on their partner's turn). Crucially, the confederate's turn sometimes ended on the critical noun that enabled participants to start planning their own turn, and sometimes continued with another irrelevant verb. Participants looked at the objects they had to name as soon as they could recognize the last noun of the confederate, regardless of whether a (predictable) verb was still coming up, suggesting they started planning as soon as they could. However, in this study, the 'early' planning point still occurred quite close to the end of the current speaker's turn (with only one verb following). By contrast, in an EEG study (Bögels, Magyari, & Levinson, 2015), participants answered quiz questions for which response planning could start either only after hearing the last word (see example 2), or already midway through the question, usually several seconds from the end (see example 3).

2. Which character from the famous movies, is also called 007?
3. Which character, also called 007, appears in the famous movies?

In these examples, the critical word is *007* because participants could in principle start planning their answer (James Bond) after recognizing this word. The EEG results showed two neural correlates that both started about 500 ms after the start of this critical word; a positivity in the ERPs and an alpha reduction in time-frequency analyses. These effects diminished or disappeared in a control experiment where participants heard the same questions but did not have to answer them (they had to remember them). Therefore, both effects were interpreted as being related to production planning. The positivity was localized to areas implicated in language production (e.g., Broca's area and the temporal lobe), suggesting it might be a direct reflection of the start of production planning. In contrast, the alpha reduction appeared in occipital and parietal regions. Therefore, this was speculatively interpreted as a switch from attending exclusively to the spoken input (which might lead to an increase in alpha over visual areas; see, e.g., Jensen, Gelfand, Kounios, & Lisman, 2002) to spreading attention towards production planning or even mental imagery of the answer (resulting in a net decrease in occipital alpha). Note that this explanation was speculative given that reduced alpha power may be related to other processes as well. One production study (Piai, Roelofs, Rommers, Dahlslätt, & Maris, 2015) appears of potential relevance here. In this study, MEG participants were asked to plan the production of a non-word but to then utter it only after seeing another (unpronounceable) non-word (the 'go-signal'). They also found an alpha (and beta) reduction over posterior brain areas but instead of attributing this to response planning, they interpreted it as reflecting monitoring for the go-signal. Crucial to their interpretation is that the go-signal was presented visually, so an occipital alpha/beta reduction (generally indicating stronger processing) makes sense. In contrast, in the study by Bögels, Magyari, & Levinson, 2015 the 'go-signal', that is, the end of the question, was presented auditorily which would not match an *occipital* alpha reduction.

Given the exploratory nature of the results by Bögels, Magyari, and Levinson (2015), a replication of this EEG study was performed (Bögels et al., 2018), using different stimuli and a slightly different paradigm. Instead of asking general knowledge quiz questions, participants saw two pictures (e.g., a banana and a pineapple), followed by a question from a confederate (e.g., 'Which object is curved and is seen as fruit?'). Participants had to answer the question by naming one of the pictures. Again, the critical word ('curved') could appear either early or late in the question. The neural correlates of response planning found in the earlier quiz study (Bögels, Magyari, & Levinson, 2015), a positivity and an alpha reduction, were largely replicated, with a similar early timing. In addition, a beta reduction was found in one of the conditions along with the alpha reduction, which was speculatively interpreted in a similar way. An additional manipulation in the same study involved expected versus unexpected words that occurred in the confederate's question, either before or after planning could start (e.g., the expected word 'fruit' versus the unexpected word 'healthy'). The size of the N400 effect (Kutas & Hillyard, 1980) elicited by comparing expected and unexpected words was then used as an indication of attention paid to that part of the question. Results showed that participants that responded quickly, presumably paying more attention to early production planning, showed an attenuated N400-effect after planning had started, thus apparently paying less attention to the rest of the question. This was not the case for participants who generally gave late responses. While this result is not directly relevant for the present study, it importantly shows that early planning can have consequences and may come with a cost. Another recent study (Corps, Crossley, et al., 2018) presented participants with yes/no-questions and asked them to either answer them or press a button when they thought the question would end. They manipulated the predictability of the questions and found that more predictable questions led to faster answers, but not to more

accurate button-presses. This suggests that participants started response planning once they could predict what the question was about, but that this information did not help to predict when the question would end (see also Casillas, De Vos, Crasborn, & Levinson, 2015 for a similar finding in a sign language).

Given these mixed findings, the literature is still inconclusive on distinguishing between the early and late planning hypotheses, while doing that is crucial to be able to understand the processes underlying the quick turn-taking in everyday conversation. Thus, more studies are necessary to get a clear answer. The studies so far differed in their ecological validity. The study by Sjerps and Meyer (2015), using a picture naming and finger-tapping task, asked participants to take turns with a recorded voice while the turns were non-contingent on each other. The four studies described above that found evidence for early planning all used interactive paradigms in which the participant's turn was contingent on the previous one. Moreover, in the studies by Bögels and colleagues (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015) and Barthel et al. (2016) the authors went to great lengths to make the participants believe that they were interacting with a live partner ('quiz master' or confederate), while in reality the turns of the partner were pre-recorded to enable greater control. While the paradigms were thus interactive, they were still very controlled and the items were carefully constructed to make sure participants could start planning their next turn at a specific position, sometimes very far from the end of the current turn. Moreover, the answers that participants had to give were mostly very short. Also, in a quiz-like situation, participants might feel time-pressured or evaluated. Combined, these factors create a rather different context for interaction than is the case in day-to-day conversation and this might affect the way participants plan their utterances. The one study described above that did use natural conversation (Boiteau et al., 2014), did not control or measure when participants could start planning their answers.

In the present study, participants' EEG is measured to look at on-line processing during listening to questions, at the same time taking a step towards more ecologically valid interaction within an interview-paradigm. The interaction was auditory only, given the restrictions on (eye) movements to properly measure the EEG. Furthermore, the largest part of the interaction consisted of participants answering questions from an experimenter, to be able to estimate when they might have started planning their response (see below). The experimenter was instructed to ask eight predetermined scripted questions in each interview, intermixed with spontaneous follow-up questions (23 per interview on average). Nevertheless, the interaction was highly natural and spontaneous, without any restrictions on the participant's responses. Within this setting, two main questions were asked. First, in such a conversation-like paradigm, when during the current turn are responders generally able to start planning their upcoming next turn? Second, do they indeed start planning close to that moment, or do they wait until the turn is almost finished? To answer the second question, the present study investigates whether the neural correlates found in the two earlier EEG studies described above (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015) can be replicated with a similar timing in this free interview paradigm.

To investigate these research questions, an experimenter performed short, informal interviews with participants about their personal occupations and interests (see example 1 and Table 1 for examples), while the participants' EEG was measured. Afterwards, all questions were coded for the word in the question which would enable participants (after recognizing it) to start planning an answer. This word will hereafter be referred to as the 'answer word', and the onset of that word as the 'answer point'. It was hypothesized that the position of the answer point would vary greatly from question to question but would occur quite far in time from the question end in a substantial amount of cases. Earlier behavioral findings in controlled paradigms have shown that gap lengths are shorter when participants can start planning earlier (Barthel et al., 2016; Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015; Magyari, De Ruiter, & Levinson, 2017). On the basis of these findings, a negative relation between planning time (the distance between the answer point and the question end) and response time was expected. That is, the earlier participants can start planning, the faster they are expected to respond. For EEG data analysis, the EEG was time-locked to the answer point and compared to data time-locked to the onset of control words at another random position in the sentence (either before or after the answer point), using cluster-analysis. EEG results were hypothesized to replicate the early positivity and alpha reduction found before (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015), for answer points relative to control points in the questions.

Given that this approach is rather new and exploratory, three additional control EEG analyses were performed with different selections of the data (see Section 2, Methods, for details). The first selection contained questions for which the answer and control words were matched on frequency, word type, and position in the sentence (this was not possible for the entire set, see Section 2.4, Coding below), to make sure that the EEG effects found could not be attributed to these variables. This selection is hereafter referred to as the 'matched selection' (or MS). The second subset contained only spontaneously asked questions by the experimenter, excluding the scripted questions ('non-scripted selection', NSS), to make sure the results were not mainly driven by specifically designed, manipulated questions. The third subset consisted only of questions for which both coders agreed on the answer point ('agreed selection', AS), see Section 2.4, Coding, below. In addition, given the free, uncontrolled nature of the data, the EEG cluster-analyses described above were supplemented by a linear mixed effects model, including several control variables.

2. Methods

2.1. Participants

Fifty-two right-handed participants from the Max Planck Institute database participated in the study. Their native language was Dutch and they did not have any hearing disorders. They gave informed consent and received 8–10 euros per hour for their participation in the entire EEG experiment. Data from 6 participants were excluded from the analysis (see Section 2.5, Data-analysis). The resulting 46 participants (10 males, 36 females) had a mean age of 22.1 years old (SD = 2.27, range: 18–28). Participants were recruited in the course of conducting two other EEG experiments (N = 24 from Bögels, Magyari, & Levinson, 2015; N = 22 from Bögels et al., 2019) but they performed the present study first.

2.2. Procedure

Before EEG preparation, participants filled out a short questionnaire with some personal information that was used to personalize some of the questions in the interview (see Section 2.3, Materials). After EEG preparation (but before starting the main experiment), participants sat down in a sound proof booth in front of a computer screen. They interacted with the experimenter outside of the booth via a microphones and loudspeakers. The experimenter of the main study for which participants were recruited always served as the interviewer (this person was thus different for the first 24 and last 22 participants). The experimenter told the participants she would have a short informal chat with them before the real experiment started. The interview started with a beep, which was both recorded in the audio and registered in the EEG signal. The participants saw a fixation cross on the screen during the interview and were instructed to try to look at the cross and not to move too much during the interview. Otherwise no restrictions were imposed on the participant. The experimenter was instructed to ask 8 pre-scripted polar questions during the interview (see Section 2.3, Materials), each followed by any follow-up questions that came to her

mind. When she could not think of further questions or the topic seemed finished, she moved on to the next pre-scripted question. The interviews lasted on average about 7 min (SD = 1.34, range: 4–11 min).

2.3. Materials

The experimenter asked eight pre-scripted polar questions that could each appear in two different versions, a long and a short version. Each participant heard four of the questions in the short version and four in the long version. These questions were designed for a different experiment, for which participants' responses were analyzed behaviorally (see Bögels & Torreira, 2015 for details). The eight questions were adjusted to each participant based on their personal details from a questionnaire (see Section 2.2, Procedure). An example of a short question is: "So, you play volleyball?" and of a long question is "So, you play volleyball on Thursday night?" (see also the first example in Table 1 for another scripted long question). Besides these pre-scripted polar questions, the experimenter freely asked a number of other open and polar questions which could differ per participant. These were spontaneous follow-up questions prompted by participants' answers to the scripted questions (e.g., asking with whom they played volleyball, whether they played competition etc.; see also Example 1 in the Introduction and examples 2–4 in Table 1). Speech from both the participant and the experimenter were recorded via the participant's microphone on one audio channel. After the experiment, the start and end of all questions and answers in each interview were measured in Praat (Boersma & Weenink, 2012) and the questions and answers were transcribed. On average, the experimenters asked 31 questions per interview (SD = 6.95, range: 17–54).

2.4. Coding

For each question in all interviews, two native Dutch-speaking raters independently selected the word in the question at which they thought the participant could start planning an answer (i.e., the answer word). They were instructed to estimate when the participant would have (or could anticipate) enough information such that he/she could have an idea what this question was about and thus could start thinking of an answer to the question. They were asked to only take into account the left part of the question (up to the word under consideration), as well as the preceding conversational context, but not what came after. It was thus possible that the speaker's intention at the end of the question turned out to be different than what appeared to be the case earlier in the question, at the answer word (this happened in 1.8% of questions in the full selection). However, since this information was not available to participants yet during the answer word, it is not likely to have affected their processing at that point (note that the later moment when it became clear what the real intention was, likely involving some form of reanalysis, was not included in the analysis). Raters were encouraged to use both the transcription and the audio recordings for their judgment (such that they could also take into account, e.g., intonation). They were also encouraged to take into account the preceding context of the question, since follow-up questions should be more predictable with context than in isolation. See Supplementary Materials (Section 1) for a complete list of instructions given to coders. The answer point, to which the EEG would be time-locked, was defined as the onset of the coded answer word, which would, after recognition of that word, lead to a possible start of production planning. This position was chosen since that would be equivalent to the earlier studies by Bögels and colleagues (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015) in which trials were also time-locked to the onset of the critical word (e.g., *007* in examples 2 and 3 in Section 1, Introduction).

The raters were trained to select the answer word reliably by coding all questions in six interviews (three with each experimenter as interviewer, randomly selected) and subsequently discussing the differences and reaching general agreements about them. This procedure was repeated for a second subset of six different interviews. Then both raters coded all interviews independently (including the trained ones). The coding proved to be difficult, as seen from relatively low inter-rater agreement of 57% in the first six interviews, but it improved to 73% for the second six interviews and was finally 78% over all questions. All disagreements were discussed among the two raters and a final agreement was reached about them. As an example of a disagreement, consider the last example in Table 1. The first part of the question in Dutch has the following word order: 'because it started annoying to get' (*omdat het vervelend begon te worden*). One of the raters initially marked 'annoying' as the answer word, while the other rater marked 'get'. While the exact grammatical structure of the question could only be known to the participant after recognizing the word 'get', the gist of the question could probably be predicted after recognizing 'annoying', which is indeed the final agreed-on answer word. However, one can imagine that discussion may exist about whether the participant will have predicted the gist at a certain point or not (this uncertainty is discussed further in Section 4.4.5 in the Discussion). All questions which led to initial disagreements were marked in the data such that they could be taken out in a later control analysis of the EEG data ('agreed selection', AS).

After these answer points were established for each question, the onset of another word in each question was chosen as a 'control point' to serve as a comparison to the answer point. These control points were assumed to provide a proper control for 'word processing' during the answer words, while not having any other specific type of processing in common, and crucially, where it was assumed that response planning would not be starting up (as was hypothesized to be the case after recognizing the answer words). Control points could occur either before (55.8% for the full selection, 49.5% for the matched selection) or after the answer word. Questions that consisted of only one word ($N = 5$) did not allow for selecting an unrelated control word so these were excluded from the EEG analysis. The initial hope was to be able to match the control words to the answer words on the following variables: word type, frequency, rank position of the word within the question, and timing (of word onset) within the question. In order to do so, word frequencies were taken from two different sources: the Subtlex corpus based on subtitles of Dutch television (Keuleers, Brysbaert, & New, 2010) and the CGN (Dutch spoken corpus, Oostdijk, 2000) since that appeared to be a better match to the spoken interaction measured here (despite the smaller number of words). However, matching answer and control words on all these variables proved difficult given that many questions were quite short and thus did not include many potential control words. Many of these short questions for example included only one content word (e.g., a noun) which was often the 'answer word'. Thus, the full selection was not fully matched on word type, in that the set of answer words contained a larger percentage of content words than the set of control words. Nor were they matched on frequency in that control words were on average more frequent than answer words (see Table 2). Therefore, a selection of questions was made for which answer and control words were overall matched on the above variables ('matched selection', MS, see Table 2). EEG analyses were performed both on this selection and on the whole set, hereafter referred to as the 'full selection' (FS, excluding some trials specific for the EEG or behavioral analysis, see Section 2.5, Data-analysis). All answer and control points were annotated by hand in Praat (Boersma & Weenink, 2012) in order to be able to time-lock the EEG signal to those exact positions.

2.5. Data-analysis

2.5.1. Behavioral analysis

The behavioral data-analysis included only the 46 participants that were also retained for the EEG analysis. Trials that had to be excluded for the EEG analysis (see Section 2.5.2, EEG Preprocessing and Cluster-analysis) were kept for the behavioral analysis, but outliers with a response time > 3 standard deviations away from the mean were

Table 2
Differences between answer and control words on several variables for all four selections. All measures are given for the EEG selection. Word type was coded by hand; nouns, verbs, adjectives, and proper names were regarded as content words and all other word types as function words. Both frequency measures (Subtlex and CGN; see Section 2.4, Coding) were log transformed. Position (words) refers to the order of the word in the sentence as counted from the beginning. Position from start (ms) refers to the time between question onset and word onset. Italics in the table indicate that the answer and control words are poorly matched on that variable.

	Full selection	Matched selection	Agreed selection	Non-scripted selection
Total # questions	1317	923	1029	973
Mean # questions	28.6 (16–47, *SD* = 7.05)	20.5 (10–34, *SD* = 6.19)	22.4 (14–36, *SD* = 5.60)	21.2 (8–40, *SD* = 7.10)
Word type	*Answer: 210 more content words* Answer vs. control	Control: 7 more content words Answer vs. control	*Answer: 240 more content words* Answer vs. control	*Answer: 123 more content words* Answer vs. control
Freq subtlex log	*3.17* vs. *4.03****	3.57 vs 3.68 (p = .03)	*2.95* vs. *4.11****	*3.41* vs *4.01****
Freq CGN log (+1)	*2.30* vs *3.14****	2.70 vs 2.77 (p = .09)	*2.11* vs *3.19****	*2.50* vs *3.13****
Position (words)	5.51 vs. 5.37 (n.s.)	5.60 vs. 5.65 (n.s.)	*5.38* vs. *5.12**	5.50 vs. 5.56 (n.s.)
Position from start (ms)	912 vs. 945 (n.s.)	921 vs. 976 (p = .05)	863 vs. 883 (n.s.)	917 vs. 977 (p = .04)

* $p < .05$.
*** $p < .001$.

removed (2.0% of the data). Note that the numbers of questions used in the behavioral and EEG analysis thus do not match exactly. Then, distributions of Response Time in milliseconds (relative to question offset), Planning Time in milliseconds (answer point to sentence end), and normalized position of answer point in the question (time from question onset to answer point, divided by question length) were inspected in a density plot created in R (Bates, Maechler, Bolker, & Walker, 2014), and mean, median, and mode were calculated.

A linear mixed-effects model was run using the lme4 package in R (Bates et al., 2014). For model coefficients, $|t| > 2$ was interpreted to correspond to significance at the 5% level (via the convergence of the t-distribution to the normal for large samples; cf. Baayen, Davidson, & Bates, 2008). Response Time (in seconds) was the dependent variable. This variable was not log transformed since removal of outliers already removed the long tails of the distribution and the negative values in the data introduce complexities into the log transform (see also Heldner & Edlund, 2010; Meyer, Alday, Decuyper, & Knudsen, 2018). Planning Time was the predictor of interest. If longer time to plan (in overlap with the question) is related to shorter response times, that would be consistent with the early planning hypothesis. Answer Length was also included as a predictor; longer answers can be expected to come later because they need more planning (Roberts, Torreira, & Levinson, 2015). Question length has sometimes been found to negatively affect response time (e.g., Bögels & Torreira, 2015; Magyari et al., 2017). Indeed, a small negative correlation between question length and response time was found in the present data ($r = -0.13$, $p < .001$). However, this variable is presumably highly correlated with planning time (in the present data the correlation is 0.78, $p < .001$) and it is likely that the negative correlation between question length and response time is due to the longer planning time for long questions. On the other hand, one study found later responses for longer questions (in individuals at clinical high risk for psychotic disorders, Sichlinger, Cibelli, Goldrick, & Mittal, 2019), which might be due to the complexity of the question (and Roberts et al., 2015 found a complex relationship between question length and response time). Thus, question length was not entered as a variable given its positive relation with Planning Time, but as a rough proxy for question complexity, Question Type (Open or Polar question) was added as a control variable. Polar questions are expected to be less complex because they ask for a choice between only two options versus open questions which ask for multiple options (e.g., Casillas, Bobb, & Clark, 2016). Thus, polar questions are expected to lead to shorter response times. As a manipulation check, time from question onset to answer point was also added to the model, since it was expected not to affect response times. In addition, four more control variables were added. Log Frequency (from the CGN) and Word Type (Content or Function word) of the answer word were hypothesized to potentially affect the time needed to process the answer word and consequently potentially affect response times. Agreed (Yes, No) and Scripted (Yes, No) were added to ensure they could not explain an effect of Planning Time. All continuous predictors were z-transformed and all categorical predictors were contrast coded (to 1,-1) before the analysis.

With respect to interactions, Planning Time was hypothesized to potentially interact with Answer Length, that is, its effect may be smaller for short than long answers. This is because short answers do not need much planning time, and therefore it might not matter (as much) whether planning time is long or short. Potentially a similar interaction would hold for Planning Time by Question Type, with a smaller effect of Planning Time for Polar questions. Therefore, Planning Time by Answer Length and Planning Time by Question Type interactions were also added to the model. Random effects were chosen to be maximal without overparameterization (e.g., Barr, Levy, Scheepers, & Tily, 2013). Thus, intercepts of Participant were included, as were random slopes for the main effects of Planning Time and Answer length by Participant (note that no random effects for questions could be added because the questions differed per participant). The addition of more random slopes to the model led to a singularity fit error. For this reason, main effects of the control variables (except for Answer Length) are not reported in the text or interpreted further, but the final model is reported in the Supplementary Materials (Section 6). All models were fit with maximum-likelihood estimation (i.e., REML = FALSE).

2.5.2. EEG preprocessing and cluster-analysis

Preprocessing and statistical analysis of EEG data was conducted using Fieldtrip (Oostenveld, Fries, Maris, & Schoffelen, 2011). Cluster-analyses were performed first, to keep the analysis as close as possible to the earlier, more controlled studies (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015). Subsequently, additional mixed-effects analyses were performed on the EEG data (see Section 2.5.3, EEG mixed-effects analyses). Since the free nature of the task led to relatively many artifacts (due to movements and eye blinks), an Independent Components Analysis (ICA) approach to artifact removal was used to retain enough trials. Epochs were extracted from question onset until speech onset, to avoid speech artifacts. Then, Principal Components Analysis (PCA) was used to reduce data dimensionality for each participant to 40 components, which were then subjected to ICA (also used by Bögels et al., 2018; see Gross et al., 2013; Oostenveld et al., 2011). These components were inspected visually and removed if they contained only noise and/or artifacts (e.g., caused by eye movements or very noisy electrodes). The average number of removed components was 7.1 (range: 2–16). The remainder of the components was used to recreate the EEG signal. Only for manual artifact rejection purposes, this signal was filtered with a low pass filter of 35 Hz, detrended, and baselined at the first 200 ms of the question. Epochs still containing eye artifacts or other artifacts that exceeded ± 100 μV were discarded. Six participants with too many artifacts after ICA were not analyzed further, resulting in 46 participants that entered the analysis. To create the 'full selection' for the EEG analysis, one-word questions, which did not allow selection of an independent control word, were

also removed, as well as questions with overlapping speech from the participant during the answer or control word, since that would lead to speech artifacts in the EEG signal. This lead to 28.8 questions remaining per participant on average (range: 16–47). The matched selection (see Section 2.4, Coding) included 45 participants with an average of 20.7 questions per participant (range: 10–34), see also Table 2 for values on the other two selections.

The EEG signal was time-locked to the answer point and to the control point in each question. Event-related potential (ERP) and time-frequency analyses (TF) were then performed. For ERPs, epochs were filtered with a low-pass filter of 35 Hz and baselined with a window of 200 ms immediately before the time-locking point. Then, questions were averaged per participant, separately for each time-locking point. For time-frequency representations, no filtering or baselining was performed, but a linear trend was removed from the data before the analysis. The power of each frequency between 4 and 30 Hz (with steps of 1 Hz) was calculated on the extracted epochs of individual trials between maximally −1 and 1.5 s using a Hanning taper (Grandke, 1983) with a window of 500 ms for each frequency, with incremental time steps of 50 ms. For illustration purposes, relative differences were calculated between conditions, dividing the absolute power difference between conditions by the sum of the power in both conditions (see Fig. 3).

To test for statistically significant differences between different time-locking points and reduce the multiple-comparison problem, the cluster-based approach (Maris & Oostenveld, 2007), implemented in the Fieldtrip toolbox, was used for the ERP as well as for the TF analysis. Clusters were formed in time, space (neighbouring electrodes), and frequency (for TF analyses) and 1000 randomizations were used for the permutation distribution. The critical alpha level was fixed to 0.05 (one-sided, given our hypotheses based on Bögels, Magyari, & Levinson, 2015). For significant clusters, sum-t statistics (the sum of all t-values in the cluster) and *p*-values are reported. This robust cluster-based approach reduces the multiple-comparisons problem and controls family-wise error across participants in time and space (see, e.g., Bögels, Magyari, & Levinson, 2015, for an elaborate description of this method). Analyses for all critical positions were performed within a time-range of 0–1500 ms for ERPs and 0–1250 ms for TFRs (the time-interval analyzed for TFRs was smaller because values at each time-point are calculated from a 500 ms window around that time point; time windows were comparable to Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015). Note that trial lengths differ, leading to a different number of trials entering the analyses for each time point, that is, the number of trials and thus the power to detect an effect diminishes with time. All cluster-analyses were based on participant averages per condition, while plots represent grand averages of participant averages.

Analyses were done for all four selections. All found effects are reported in the Results section, but with a focus on effects that are present in both the full and the matched selection (MS) and/or appear to be replications of effects found in earlier studies (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015) in the interpretation and discussion.

2.5.3. EEG mixed-effects analyses

Given the observational nature of the present data, and the presence of many potential confounding variables, additional linear mixed-effects analyses were performed on the ERP and TFR data, confined to a pre-defined time(−frequency) window chosen on the basis of earlier findings (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015) and visual inspection of the grand averages (Figs. 2 and 3). Nine regions of interest (ROIs) were defined, consisting of 6 or 7 electrodes each, by crossing the factors Anterior-Posterior (Anterior, Mid, Posterior) and Left-Right (Left, Mid, Right; see Fig. S1 in Supplementary Materials, Section 2, for the exact division of all electrodes over the nine ROIs). For the ERP analysis, the dependent variable was the average voltage (per trial) in a 400 to 800 ms window after answer/control word onset over all electrodes in each ROI. For the TFR analysis, the dependent variable was the average power per trial) in an 800–1000 ms window after answer/control word onset, between 9 and 13 Hz over all electrodes in each ROI. Outliers with an average (voltage or power) > 3 standard deviations away from the mean were removed. For ERPs (voltage), no further transformation was applied because it includes both negative and positive values. For TFRs (power), however, values are only positive and thus the distribution was clearly right-skewed. Power was thus log-transformed, rendering the residuals more normal. The analyses for ERP and TFR were otherwise identical. The predictor of interest was Condition (Answer, Control word). The two ROI variables Anterior-Posterior (reference: Posterior) and Left-Right (reference: Right) were added to the model, together with their interactions with Condition and the three-way interaction between the two ROI variables and Condition. Furthermore, the following control variables were added: Log Frequency (from the CGN) and Word Type (Content, Function) of answer/control word, Position of answer/control word (from question onset), Agreed (Yes, No), and Scripted (Yes, No), to make sure that these variables could not explain a potential effect of Condition. All continuous predictors were z-transformed and all categorical predictors were contrast coded (to 1, −1) before the analysis. In accordance with maximal random effects, intercepts of Participant and random slopes for Condition by Participant were added. The addition of more random slopes to the model led to a singularity fit error. For this reason, main effects of the control variables are not reported in the text or interpreted further, but the final models are reported in the Supplementary Materials (Sections 8 and 9). In case of interactions between Condition and the ROI variables, follow-up analyses were done for smaller ROIs separately. All models were fit with maximum-likelihood estimation (i.e., REML = FALSE).

3. Results

3.1. Behavioral results

The behavioral data used for the analyses is publicly available (Bögels, 2020). Fig. 1 displays density plots of the distributions of Response Time (relative to question offset, panel A; see Fig. S2 in Supplementary Materials, Section 3, for an indication of individual variability between participants in this measure), Planning Time (position of answer point relative to question offset, panel B), and normalized position of answer point in the question, relative to question onset (divided by question length, panel C). Panel A shows a wide distribution that is somewhat right-skewed as is typical of responding in conversational contexts (Stivers et al., 2009), including gaps as well as overlaps of varying lengths. The mode of the distribution can be seen to lie around 200 ms (M = 380, MED = 336 ms), which is a value typically reported in the literature (e.g., Heldner & Edlund, 2010; Stivers et al., 2009), showing that the response latencies in this context (EEG, non-face-to-face) are very natural. As can be seen in Panel B, the Planning Time (until question offset) is about 700 ms in the most frequent case (M = 1374, MED = 1083 ms). However, given that the distribution is very right-skewed, in many cases the Planning time is much longer. Since the possible Planning Time is constrained by question length, the presence of many short questions in the data (see Fig. S3 in Supplementary Materials, Section 4) might render the mode of Planning Time relatively small. Therefore, to determine how early in a question listeners can generally start planning, it might be more informative to look at the relative position of the answer point in the question, which is displayed in panel C. This panel shows that, most frequently, the answer point appears after less than one third of the question has passed (M = 37.8%, MED = 36.2%; but note that this is the *start* of the word that enables listeners to start planning their answer). Thus, it appears that listeners often at least have the option of planning their answer in overlap with a large part (about 2/3) of the question. These general observations were qualitatively similar when only including

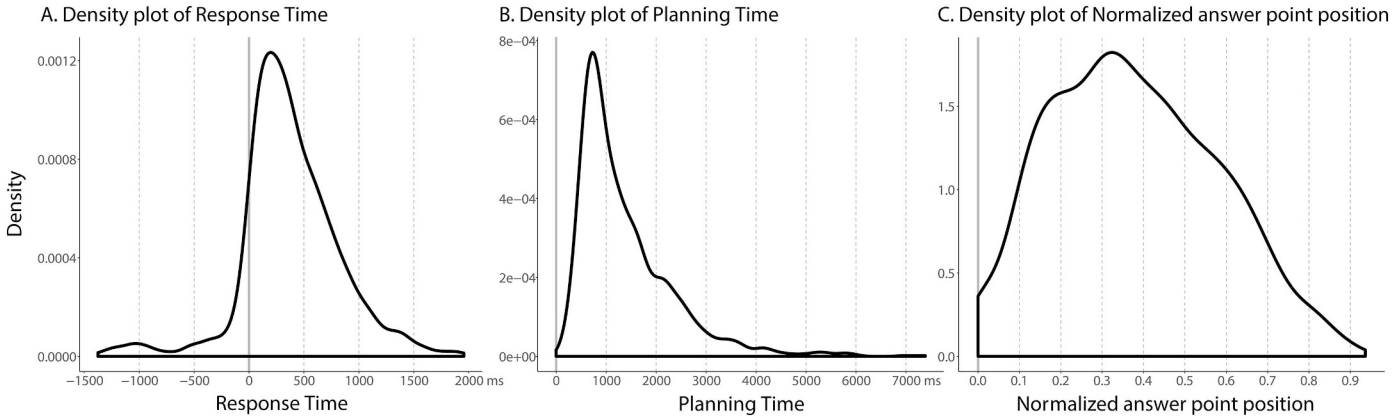

Fig. 1. Density plots for several data parameters. Panel A displays the distribution of Response Time relative to question end (mode around 200 ms). Negative values indicate overlaps. Panel B displays the distribution of Planning Time (time between the answer point and question offset; mode around 700 ms). Panel C displays the distribution of the normalized position of the answer point within the question (measured from question onset; mode around 0.32).

questions that were spontaneous (non-scripted selection) or for which both coders agreed (agreed selection; see Table S1 in Supplementary Materials, Section 5).

A linear mixed-effects model was run as described in Section 2.5, Data-analysis. Planning Time had a negative effect on Response time ($\beta = -0.138$, $SE = 0.019$, $t = -7.465$), showing that the more time participants had to plan before question end, the shorter their response time, while controlling for several control variables. Furthermore, there was a positive effect of Answer Length ($\beta = 0.158$, $SE = 0.022$, $t = 7.170$), with longer answers showing longer response times. The interactions included in the model were not statistically significant (Planning Time*Answer Length: $\beta = 0.014$, $SE = 0.010$, $t = 1.401$; Planning Time*Question Type: $\beta = 0.006$, $SE = 0.013$, $t = 0.508$; see Table S2 in Supplementary Materials, Section 6, for the full model).

A control analysis was performed to rule out that the effect of Planning Time on Response Time was due to only very short planning times (< 400 ms) that would not enable interlocutors to come in at a 'normal' response time of 200 ms, because there was less than 600 ms left to plan. If that would be the case, the results would be compatible with both the early and the late planning model. Therefore, the analysis was repeated using only trials where the planning time was longer than 400 ms (excluding a further 3.1% of the data and leaving 1364 trials for analysis). This yielded very similar results as the first analysis: shorter response latencies with longer planning time ($\beta = -0.134$, $SE = 0.019$, $t = -7.089$), longer response latencies with longer answers ($\beta = 0.153$, $SE = 0.022$, $t = 7.035$), and no interactions (all $|t| < 1.5$; see Table S3 in Supplementary Materials, Section 6, for the full model).

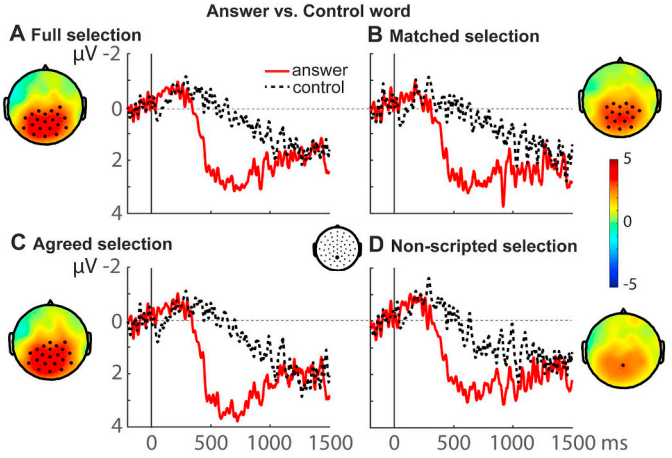

Fig. 2. ERP results for answer versus control time-locking points. Grand average ERPs for a representative electrode (location in the small head in the middle), showing a large positivity starting around 400 ms after the onset of answer words (red solid line) relative to control words (black dashed line). Panels A-D present results for the full, matched, agreed, and non-scripted selection, respectively. Topographical plots are displayed for the answer versus control point averaged over a 400–800 ms time window (where the positivity is largest) and show a predominantly posterior distribution of the effect. Colors indicate t-values. Electrodes that show a significant effect in > 70% of the time window are highlighted in black. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Event-related potentials

The EEG data used for the analyses reported here are publicly available (Bögels, 2020). Fig. 2, panels A–D display grand average waveforms time-locked to the answer point (red solid line) and the control point (black indented line) for all four selections in a representative electrode (see Fig. S4 in Supplementary Materials, Section 7, for an indication of the individual variability in the size of the positivity). In all panels, a large positivity is visible for answer relative to control words, starting around 400 ms and lasting until about 1000 m s, which seems largest over more posterior electrodes. A cluster-analysis comparing answer to control words between 0 and 1000 ms showed a significant positive cluster for all four selections (FS: 354–864 ms, sum-t = 16,969, p = .002; MS: 398–830 ms, sum-t = 12,919, p = .008; AS: 350–962 ms, sum-t = 20,512, p = .002; NSS: 446–758 ms, sum-t = 7395.8, p = .022). Note that the fact that the clusters did not reach all the way up to 1000 ms (as visual inspection would suggest) might be due to an increase of noise because the trial numbers gradually drop as the trial gets longer (because trials were cut off when the participant started to speak their response). The topographical plots in Fig. 2 show the distribution of the positivities over the scalp between 400 and 800 ms, confirming a mostly posterior distribution.

A linear mixed-effects model was run on the average voltage between 400 and 800 ms in 9 ROIs as described in Section 2.5.3, EEG mixed-effects analysis. The average voltage over all ROIs was more positive for Answer relative to Control points ($\beta = 0.404$, $SE = 0.164$, $t = 2.471$), when taking into account all control variables. Moreover, an interaction between Condition and Anterior-Posterior was found for the Anterior versus Posterior ROI ($\beta = 0.315$, $SE = 0.094$, $t = 3.354$). Follow-up linear mixed-effects models were performed for the Anterior, Mid, and Posterior ROIs separately with the same variables, except for the variable Anterior-Posterior and interactions between Condition and Left-Right, since no three-way interactions between Condition and the two ROI variables were present in the main analysis. These analyses showed more positive voltages for Answer relative to Control points in the Mid ($\beta = 0.447$, $SE = 0.188$, $t = 2.371$) and Posterior ROIs

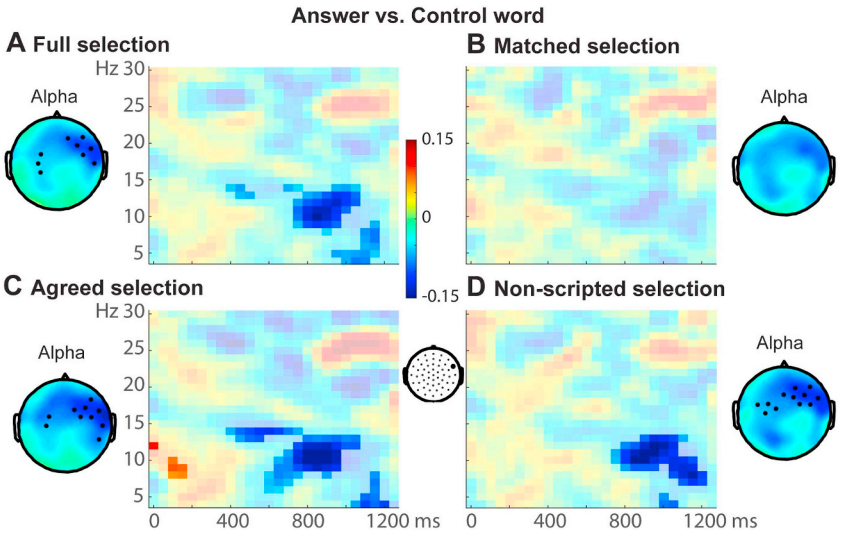

Fig. 3. Time-frequency results for answer versus control words. Time-frequency results are presented for a representative electrode (see head in the middle), generally showing decreased alpha power around 800–1000 ms after the onset of answer words relative to control words. Panels A–D present results for the full, matched, agreed, and non-scripted selections, respectively. Colors in all plots indicate the relative difference between raw power in the relevant conditions (the difference divided by the sum of both conditions). In time-frequency plots, the relative difference is given in transparent colors with the statistically significant cluster overlaid in opaque colors. Topographical plots are given for alpha effects (average over 9–13 Hz, 800–1000 ms), showing a predominantly left anterior distribution of the alpha decrease. Electrodes that are significant in at least 30% of the time/frequency window are highlighted in black. (For interpretation of colour in this figure, the reader is referred to the web version of this article.)

($\beta = 0.707$, $SE = 0.169$, $t = 4.174$), but not in the Anterior ROI ($\beta = 0.038$, $SE = 0.194$, $t = 0.169$). See Tables S4-S7 in Supplementary Materials, Section 8, for the full models including results for all variables.

Thus, for all selections a robust positivity was found quickly after participants presumably could start planning their response. The positivity was still reliable on the middle and posterior parts of the skull when controlling for several control variables in a mixed-effects model. This replicates previous findings that were obtained under more controlled circumstances (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015; see e.g., Bögels et al., 2018, Fig. 2 for comparison). The effect appears similar in timing (although maybe starting and ending somewhat earlier) and distribution to the earlier found effects, but the size appears somewhat smaller. These observations are further discussed in Section 4, Discussion.

3.3. Time-frequency

Fig. 3 displays the relative difference in power (the difference divided by the sum) between answer and control words in a representative electrode, for all four selections, in panels A-D. Cluster-analyses between 0 and 1250 ms and 4–30 Hz showed one positive cluster for the agreed selection only (sum-$t = 1637.2$, $p = .047$, but note that this does not survive a 2-tailed test), mostly spanning theta to alpha frequencies at the start of the time window. However, given that this effect was not replicated in any of the other selections and might therefore be due to differences in frequency or other non-matched variables between the answer and control words, it will not be interpreted further. Analyses for three of the four selections showed a negative cluster (FS: sum-$t = -3368.8$, $p = .002$; AS: sum-$t = -3781.0$ $p < .001$; NSS: sum-$t = -2115.5$, $p = .018$), but no effects were present in the matched selection ($p > .17$). As is visible in Fig. 3, the negative effects appear somewhat scattered over different frequencies but mostly cluster around 800–1000 ms in time. One part of the clusters appears to occur around the theta frequency (5 Hz) and around 1200 ms. This theta effect will not be interpreted further, because it was not expected based on the previous results (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015). Another part of the negative cluster lies around the alpha frequency (around 10 Hz and slightly above) and is reminiscent of the effect that was found in the more controlled studies. However, comparing the distribution of the alpha effects (see topographical plots in Fig. 3) to that of the previously found effects (e.g., Fig. 5 of Bögels et al., 2018), the current distribution appears a bit more anterior than distributions in previous studies, which were mostly posterior.

A linear mixed-effects model was run on the average power between 800 and 1000 ms and between 9 and 13 Hz in 9 ROIs as described in Section 2.5.3, EEG mixed-effects analysis. The average power did not differ significantly between Answer and Control points over all ROIs ($b = -0.025$, $SE = 0.019$, $t = -1.32$), when taking into account all control variables, nor were there any interactions between Condition and the ROI variables (all $|t| < 1.1$; see Table S8 in Supplementary Materials, Section 9, for the full model).

In sum, although a negative alpha effect was found for Answer points in three of the selections, it was not significant in the matched selection, nor in the linear mixed-effects model taking into account all control variables. Therefore, it cannot be ruled out that this effect is due to low-level differences between answer and control words (such as frequency). The negative alpha effect will be discussed further in the Discussion.

4. Discussion

The background of the present study are two competing hypotheses on when interlocutors in conversation start planning their next turns (see also Bögels & Levinson, 2017), namely as late as possible to minimize overlap (late planning hypothesis), or as early as possible because of the importance of coming in on time (early planning hypothesis). The first aim of the present study was to see whether this is actually a valid distinction, by investigating when participants in a more natural interaction can generally start planning their response to questions. The second aim was to see whether neural correlates of response planning found earlier in controlled studies could be replicated in a more natural situation. To investigate these questions, a spontaneous, interactive situation was created that still enabled EEG measurements and an estimation of the possible start of response planning. That is, participants were interviewed informally by an experimenter, who asked some scripted but mostly spontaneous questions. According to naïve coders, participants could start planning their response relatively early in the questions, that is, only about one third into the question. Furthermore, the more time responders had to plan their answers before question offset, the sooner they responded. The EEG results showed a robust positivity quite soon after the answer point, relative to a control point in the sentence, replicating earlier studies in which the answer points were manipulated. Furthermore, an alpha suppression was found for three of the selections, which is reminiscent of the alpha suppression found in earlier studies (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015). However, it was not replicated in the matched selection, nor in a mixed-effects model taking into account several control variables. Moreover, it had a slightly different timing

and distribution as in earlier studies. These results are discussed in turn below and subsequently implications of these findings for theories of turn-taking are outlined.

4.1. Listeners can start planning early

For the distinction 'as early' and 'as late as possible' to be valid, one has to know when within the current turn the position occurs at which prospective speakers would actually have enough information to start planning their response. The two raters in the present study did not always agree on when the answer point occurred, which will be discussed further in Section 4.4.5 below. When taking their joint decision, the results showed that, on average, this position occurred around 1400 ms before question offset, and its most frequent position was around 700 ms before question offset. This might appear as quite short, but assuming that planning takes 600 ms (minimally) and assuming a gap of 200 ms, this would mean that planning can in most cases start at least a few hundred milliseconds before the 'as late as possible' position (which would be around 400 ms before question offset). Moreover, this perhaps short mode of 700 ms planning time is likely due to the relatively large number of short questions in the present study (see Fig. S3 in Supplementary Materials, Section 4). Since short questions impose a lower bound on the possible amount of planning time, this leads to a situation in which the bulk of the planning times is short as well. Many more questions exist where participants had more time to plan than 700 ms, than questions where they had less time. This situation might be representative of conversations in general. Thus, it may be more informative to look at the relative position of the answer point in the sentence, which occurs after only around one third of the question in most cases and after 38% of the question on average. These results show that listeners can actually often start planning after having heard only a small part of the question.

One might question whether the off-line coders were able to correctly identify the answer point in the same way that listeners would do on-line, especially since they did not agree in about 22% of the questions. However, including whether coders agreed on the answer point as a control predictor did not change the behavioral results. Moreover, the position of the coded answer point clearly appears related to the participants' behaviour and neural responses (see below).

This data also gave an opportunity to find behavioral evidence that listeners start planning earlier than is strictly necessary to start 'on time'. Several earlier studies (Barthel et al., 2016; Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015; Magyari et al., 2017) have shown, using manipulated first turns, that interlocutors started their response earlier when they could start planning 'early' relative to 'late' within the previous turn. The present study is the first that uses, in part, spontaneous questions to show a similar effect, this time using a graded measure of planning time (i.e., answer point relative to question offset). There was a negative relationship between time available for planning and response time, which could not be explained by very short planning times only. This effect thus shows that listeners generally make at least some use of the time available for planning and do not wait until the end of the question to start planning.

4.2. Replicating neural correlates for production planning

The second aim of the present study was to see whether the neural correlates for production planning during turn-taking, found and replicated using controlled stimuli (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015), would extend to more natural stimuli. In order to show this, the answer point for each question first had to be determined post-hoc, rather than manipulating it up front. Second, a control word needed to be defined in the same stimuli to compare the ERPs and oscillations to. Given that it was difficult to match answer and control words, because of the shortage of control words to choose from within (especially short) questions, the main results were also assessed for a matched subset of stimuli (on frequency, position in the question, and word type) and linear mixed-effects models were performed including several control variables. Note that predictability of answer/control words from the preceding context was not measured directly, although this is also an important factor that can affect EEG results. However, it appears reasonable to assume that part of the variance explained by predictability would be captured by the other variables that were controlled for; more predictable words presumably occur later in utterances, are more frequent, and may be more likely to be function words. Though it cannot be excluded that the answer and control words were not exactly matched on predictability, even in the matched selection, the fact that no N400 effect (Kutas & Hillyard, 1980) is found in any of the selections, makes it highly unlikely that answer and control words differed strongly on this variable. If there would be a difference in predictability between the two conditions, one would expect answer words to be less predictable. Answer words (as defined in the current study, as well as by Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015) cannot be very predictable because they first enable listeners to start planning their response. If they would be predictable from the context up to that word, listeners would already be able to start planning their response based on that context, which would render the last word (or an earlier word) of the context the actual answer word. This reasoning does not hold for control words. Thus, it is highly unlikely that the found positivity is actually an N400 for control words. If, in contrast, answer words would have been more predictable than control words, this would very likely have led to an N400 effect for answer words (as found in Bögels et al., 2018 for example), given that the N400 is very highly correlated with the cloze probability of a word (i.e., $r = 0.9$; Kutas & Federmeier, 2011, page 623). Note that some studies comparing expected and unexpected words do find positivities (mostly in addition to N400 effects), but this mostly concerns high cloze contexts followed by an unexpected word (e.g., Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007) or incongruous words that are really impossible in their context (e.g., Van Petten & Luka, 2012). These situations are unlikely to happen in the present study, given the natural, unmanipulated stimuli used.

4.2.1. Positivity in ERPs

The ERPs showed a robust positivity for answer relative to control words in all four selections and the effect also held up in the linear mixed-effects model for the mid and posterior ROIs, including several control variables. The found cluster lasted roughly between 400 and 800 ms after answer point onset and had a posterior distribution. Comparing this effect descriptively to the positivity found in earlier studies (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015), the polarity, peak timing, and distribution clearly appear to be similar. A first possible difference concerns the size of the effect. Based on visual inspection of Fig. 2 of the present study and Fig. 3 of Bögels et al. (2018, also including a panel displaying results by Bögels, Magyari, & Levinson, 2015), the largest effects in the earlier studies descriptively appear at least 3 to 4 microvolts as compared to roughly 2 microvolts in the present study. Although one should be careful interpreting this difference between studies and participants, the size of the difference is such that it may be meaningful. Indeed, more controlled studies are likely to find larger effects, given that the stimuli were carefully constructed to make sure that planning could only start after hearing that one critical word. In the present study, in contrast, the questions were mostly spontaneous and it there may not always be one specific word that suddenly enables participants to start planning their response. That means that there might be more variability within the questions as to whether participants indeed started planning at exactly this moment. This idea is corroborated by the non-perfect agreement of coders on the answer point (78%). Indeed, when only the agreed answer points are taken into account (see Fig. 2, panel C), the effect appears to be a bit larger than in the full selection, suggesting less variability around the coded answer point. See also Section 4.4.5 below for further discussion

on this point.

A second possible difference is the timing of the effect. Although cluster-based analyses do not allow for precise estimations of the timing of effects (Sassenhagen & Draschkow, 2019), descriptively the present positivity appears to start somewhat earlier than most of the earlier found positivities, which generally did not appear to start until after 500 ms (except for the one comparing answer and control words at the end of the sentence in Bögels et al., 2018). One possible reason for this in the present study might be the absence of a negativity (N400) preceding the positivity; a pattern sometimes found in the previous, more controlled studies (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015). A second possible reason for an earlier effect in the present study might be the nature of the answers. In the two controlled studies, participants' responses were always single nouns in the context of a quiz-like question. Especially in Bögels, Magyari, and Levinson (2015) the answers needed to be retrieved from long-term memory, which might take some time. In contrast, in the present study the responses were much more variable in complexity. On the one hand, they could be much longer than one word, but on the other hand, they could also be much simpler (like a 'yes'-response) or they could be built up incrementally, which might arguably be started up faster. This explanation is corroborated by comparing the modes of response times in the present study (around 200 ms) with those in the more controlled studies (around 300–800 ms, cf. Fig. 2 in Bögels et al., 2018). However, this explanation could not be confirmed in an exploratory analysis comparing questions with short versus long answers (median split per participant) at the answer point, which showed no significant differences in the positivity (latency or size) between long and short answers. Future research going both into the direction of careful experimentation (e.g., carefully controlling for answer length and other factors) and into the direction of collecting EEG data on larger corpora of more spontaneous conversation is thus needed to determine more precisely which factors affect the latency and size of the positivity. Next to the earlier start in the present study, it also seems to end earlier than in previous studies, but this is likely due to less power in later time windows due to varying trials lengths.

Overall, despite small variability in size and latency of the positivity, it closely resembles previous results in the context of turn-taking. Previously (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015) this positivity was interpreted as a neural correlate of response planning, based on a localization of the positivity in areas related to language production and comparisons with a control study without speech (Bögels, Magyari, & Levinson, 2015). The present results are interpreted in a similar way, showing that upcoming responders start planning as soon as they have enough information to do so, consistent with the early planning hypothesis. Crucially, this is the first time that this is shown in a naturalistic context.

4.2.2. Alpha decrease

The results in the time-frequency domain, for three of the four selections, showed a lower alpha power around 800 ms after the answer point as compared to the control point. However, the matched selection did not show the same effect, nor was it replicated in a linear mixed-effects model including several control variables. Thus, it cannot be excluded that the effect in the other selections was due to low-level differences between answer and control words, such as frequency. Even if the effect found in the other three selections was due to the start of response planning, it is unclear whether the present effect has the same underlying source as the alpha effects found in previous, more controlled studies (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015), given differences in timing and distribution. Bögels, Magyari, and Levinson (2015) speculatively interpreted this effect as an attention switch from attending exclusively to the spoken input (leading to an increase in alpha over visual areas; see, e.g., Jensen et al., 2002) to spreading attention towards other processes, such as production planning. The distribution of the alpha effect in the present study is more anterior/temporal in the present study, rather than posterior, as in earlier studies. This may be due to a less visual nature of the answers required in the present study, as compared to general knowledge quiz questions (Bögels, Magyari, & Levinson, 2015) or remembering the name of a just presented picture (Bögels et al., 2018). Furthermore, the alpha decrease descriptively appears quite a bit later in the present study (around 800 ms) relative to earlier studies (around 500 ms) and is less extended, which may be due to noisier data with more jitter. Further turn-taking EEG studies are needed both to determine exactly what such alpha decreases reflect and, if they are related to turn-taking, to shed more light on the variables that affect their timing and strength.

4.3. Generalizability

The main aim of the present study was to see if the results of earlier studies in terms of timing and neural correlates of planning in turn-taking, based on controlled experimental stimuli, could be extended to a more natural situation. To this end, an informal interview setting was created, with a mix of some pre-scripted and mostly spontaneous questions. Similar settings occur in participants' real-life experiences using language, for example when giving background information at the doctor's office. Still, the exact situation can be argued to be no fully ecologically valid situation since it is characterized by an asymmetry between interviewer and participant, where the former asks all the questions and the latter answers them, pre-scripted questions are included, and the interaction is not face-to-face, in order to prevent too much movement for the EEG recordings. These worries may be mitigated by showing that the mode of the response times relative to question offset was around 200 ms, just as has been generally reported for spontaneous conversations (e.g., Heldner & Edlund, 2010; note that response times in Bögels, Kendrick, & Levinson, 2015 and Bögels, Magyari, & Levinson, 2015 were generally longer with modes between 300 and 800 ms, depending on the condition). Moreover, the same pattern of results was found when including only spontaneous (non-scripted) questions.

Still, the present results might be easiest to generalize to telephone conversations, including service interactions (e.g., with a plumber or cable provider). However, it is to be expected that planning also starts early in face-to-face conversations, where perhaps even more cues (like gestures and facial expressions) can contribute to an early 'answer point' at which the listener has enough information to start planning. Still, future EEG studies should confirm this, in which movements in face-to-face conversation might be minimally controlled or filtered out in analysis.

In the present study, as in the previous, more controlled ones, the focus was on question-answer pairs: when during a question will listeners start planning their response? However, turn-taking in actual conversation does not only occur between questions and answers (or more broadly, adjaceny pairs, Schegloff, 2007). After a question has been answered, the questioner could ask another question. After one person tells an anecdote, someone else could tell one of their own. Much of the corpus research that looked at response times in conversation (e.g., Heldner & Edlund, 2010) has taken all floor transfers together and still finds a mode or response time around 200 ms, as in the present study on questions only. The similarity in distribution suggests that different types of floor transfers will behave similarly and that planning thus might start early, also when the planned utterance is not an answer to a question. Still, it would be interesting to look at planning of other turns that are not answers to a question more closely in future research. However, this might not be possible in a similar way as has been done in the present study. Given that questions 'require' an answer (Schegloff, 2007), it is possible (although not very easy as seen in the present study) to code at what position in the question that answer could be given by a responder. This appears a good proxy for when the responder can first start planning their next turn. For turns that are not answers to a question, this may be quite different. If the upcoming

turn is not contingent on the previous one (as in an out-of-the-blue question, or a story that one wanted to tell for some time but did not get the chance to), it could in principle be prepared at any time. It appears that the start of planning in those cases might be initiated speaker internally, rather than being elicited by something within the previous turn. In that case it would be hard to estimate at which exact moment planning would start. On the other hand, also turns other than answers might often be contingent on the previous turn (as in follow-up questions). Future research could investigate to what extent coders would agree on 'answer points' in a spontaneous symmetrical conversation containing all kinds of turns. As a next step, neural correlates of planning different kinds of turns could be investigated.

4.4. Theoretical implications

The present results thus appear to favor the 'early planning' model as outlined in the Introduction. The found neural signatures suggest that planning starts early, also in a natural turn-taking situation and do not appear to be compatible with the 'late planning model'. In the present section, potential implications of adopting the early planning model are discussed, as well as some open questions and speculations. The different sections below will discuss (1) how the early planning model assumes responders come in on time; (2) why the neural signatures are unlikely to be related to monitoring for turn-ends; (3) whether the early and late planning model are mutually exclusive; (4) how conversationalists might manage the overlap between comprehension and production planning; (5) the potential individual variability in the possible start of planning between conversationalists; and (6) how responders manage the short gaps found in conversation.

4.4.1. Articulation based on turn-final cues

According to the early planning model as described in the Introduction, next to early planning, listeners try to determine turn-ends on the basis of cues occurring at the ends of turns and launch their articulation of the (presumably pre-planned) turn on the basis of these cues. Bögels and Torreira (2015) showed that turn-final prosodic information (e.g., intonation and lengthening) is one of the types of information used by listeners to determine turn-ends. However, prosodic information is likely not enough (and may not be 100% reliable, see, e.g., Gravano & Hirschberg, 2011). In that same paper by Bögels and Torreira (2015), it was argued that general 'linguistic' completion, consisting of syntactic, semantic, pragmatic, *and* prosodic completion, probably cues the turn end. However, even at (manipulated) points of full linguistic completion followed by more speech, still only a part of listeners identified this as the turn end on-line (by pressing a button). This result suggests that in some cases, listeners require linguistic completion followed by a small amount of silence before they are sure enough to launch their turn. Thus, identifying linguistic completion plus in some cases a small amount of silence may allow for a large part of the turn-taking distribution. Still, turn-taking is not always perfect. Listeners may sometimes come in too early and end up in overlap with the current speaker. However, this may often be solved quickly because one of the two speakers may notice the overlap and stop speaking immediately, leading to only a very short overlap (Levinson & Torreira, 2015, page 17). In other cases, listeners may not have planned their response in time, or may not have detected the turn-final cues, leading to relatively late responses.

4.4.2. Monitoring for completion

Given that the early planning model thus proposes a combination of early planning and late articulation (triggered by linguistic completion), this implies that listeners are also monitoring the turn for such completion. One may then wonder whether the neural signatures presumably related to response planning (i.e., the positivity and/or alpha power decrease) could also be a reflection of this monitoring for the turn end (cf. Piai et al., 2015 who interpreted an alpha power decrease as monitoring for the go-signal to speak). However, it appears unlikely that the onset of such monitoring would be time-locked to the moment when planning can start. The two processes of planning and turn-end detection appear to be independent (Corps, Crossley, et al., 2018). If that is the case, monitoring for the turn-end should be going on during the entire turn and not start exactly at the moment that planning can start. It would be especially difficult under this proposal to explain why these same neural signatures were also found in one specific condition within the earlier controlled studies (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015). Time-locking to words occurring towards the ends of questions, these signatures were found for words that enabled the start of planning at that late moment, in comparison to questions for which planning had already started much earlier. It is unlikely that monitoring for the turn-end was NOT happening at final words of questions where planning could have started already much earlier.

4.4.3. Intermediate model?

Given the strict distinction between the early and late planning models as described in the Introduction, one may wonder whether these two models are the only options or whether elements of either model would be compatible with the other. The reason they were described in this way is partly historical. One of the first on-line studies on turn-taking (De Ruiter, Mitterer, & Enfield, 2006) suggested that late prosodic cues were not relevant for turn end identification and that they in fact probably occurred too late to be useful since that would not allow for the short gaps seen in conversation. This view thus suggested that planning starts as late as possible, and for this to work, listeners need to be able to estimate turn ends quite precisely from minimally 400 ms before the end (in order to realize gaps of about 200 ms). Otherwise, it is unclear how listeners would determine when they should start planning. This prompted an opposing view, which separated response planning into two processes, one involving all processes up to articulation (starting early) and another involving only articulation (starting late; e.g., Levinson & Torreira, 2015). In the meantime, final cues have been shown to be used by listeners (Bögels & Torreira, 2015) and most of the current research appears to point to an early start of planning (including the present study). However, this is not to say that lexical prediction, as deemed crucial in the late planning model, is not relevant for turn-taking at all. On the contrary, it is very likely that listeners try to predict the content of the current utterance in order to be able to start planning their utterance as early as possible (e.g., Corps, Crossley, et al., 2018). Whether lexical prediction is also used to better estimate turn ends is more debated, with some studies showing that predictability helps (e.g., Magyari & De Ruiter, 2012) and others that it does not (Corps, Crossley, et al., 2018). However, there is no principled reason why the early planning model would exclude a role for lexical prediction in turn-end estimation. For example, if listeners estimate, based on lexical prediction, that the end of the turn is drawing near, they may start paying more attention to turn-final cues.

4.4.4. Managing comprehension-production overlap

If upcoming responders in conversation indeed start planning their response as soon as they can, and this moment often occurs relatively early on in the turn, as suggested by the present study, how then do conversationalists manage the large amount of overlap between comprehension and production planning? This is an open question, which requires future research, but a few speculations can be brought forward. First, a recent study (Fairs et al., 2018) shows that performing two (low-level) linguistic tasks at the same time, leads to quite some interference, which appears to be problematic for overlap in conversation. However, first, the overlap may be problematic for certain stages in comprehension and production (such as lexical selection) but not others. These other stages may be performed in parallel with one another, rendering the overlap more feasible. As for the more demanding stages, these may be, for example, performed intermittently, with conversationalists

rapidly switching between comprehension of the ongoing turn and production planning. How many resources are allocated to either of the two processes may then be subject to individual differences (see, e.g., Bögels et al., 2018). Furthermore, the difficulty of performing the two tasks in overlap may also be alleviated because they are heavily related. Responses are generally contingent on the questions that precede them, which may allow for some form of priming perhaps leading to facilitation between the two tasks as well.

On the other hand, a recent study on turn-taking (Barthel & Sauppe, 2019) showed that production planning led to higher processing load (as measured by pupillary responses) when it started in overlap with the current turn than when it could only start when that turn had finished. Thus, it is very well possible that comprehension of the ongoing turn, production planning, or both, are of lesser quality (or speed) when done in overlap. Still, apparently, conversationalists are willing to trade the higher processing load and possible suffering of one or both processes for shorter turn transitions. By minimizing turn timings in this way, they can be used as a source of information in conversation (see also the discussion by Barthel & Sauppe, 2019, p. 10–11), for example indicating whether a response will be preferred or not.

4.4.5. Variability in planning onset

The present study shows that upcoming responders can often start planning their response relatively early in a question, after hearing only about one third of it. However, the agreement between the two raters was not very high. Initially their agreement was only 57%, which climbed up to 78% after some training. Apparently, even knowing the conversational context and having access to lexical and prosodic information did not always lead to agreement about the moment at which the responder could have started planning. One piece of information that may not have been available to the coders was the common ground between the interviewer and the participant. However, since they knew each other only since the start of EEG preparation, it is unlikely that this would have a large effect on top of the discourse context here. It is more likely that the exact start of planning is actually not always clear-cut (although it may be very clear in other cases). Even in the hypothetical situation that two people would receive the same question in the same context, they might not both feel they have enough information to start planning their response at the same time. What that point is, may depend, for example, on how well responders can predict the upcoming words, and/or how willing they are to rely on this prediction for starting response planning. As an example, if you meet someone for dinner and they start saying 'How…?' some people may predict 'How are you?' and immediately start planning a response to that, whereas others may wait to hear if the speaker will not instead ask 'How is your brother?', in which case an entirely different response has to be planned. Some people (in some situations) may be willing to take the risk that they will have to abandon and revise their production plan (or even give an inappropriate response), whereas others may not be. Thus, the extent to which coders are able to determine the 'correct' answer word (that is, the one used by the responder) may differ per speaker and situation. This variability may in part account for the high inter-individual variability in the size of the positivity (displayed in Fig. S4 in Supplementary Materials, Section 7).

In any case, the early planning model would predict that listeners start planning as soon as they are confident enough that they know what the question is about. An alternative explanation, put forward by an anonymous reviewer, would say that cases of low agreement between raters point to high uncertainty in conversationalists as well, which would lead them to forego these potential answer points and start planning only (much) later. However, this would predict much earlier response times for the agreed than the non-agreed questions in the present dataset, which does not appear to be corroborated by Table S1 (in Supplementary Materials, Section 5), which shows comparable values for the full and agreed selections.

This observation does complicate the measurement of the answer point in any given question. The present study used two raters and already showed that the positivity appeared more robust for questions on which the raters agreed. More fine-grained questions may be answered in future research with multiple raters to create a probability profile of the answer point over the question. Other than simply asking raters to indicate the answer point, one could also use a gating paradigm and ask raters to guess what the question is about. The probability of the answer point would then be proportional to the consistency between raters. Although this method would not make sure that the answer point is identified accurately for this specific speaker, it may give more confidence. A next challenge would then be to time-lock and analyze the EEG signal relative to these graded answer points. An alternative may be to ask the interviewed participants themselves after the interview to indicate for each question when they had started planning, although it is not clear whether such introspection would be reliable.

4.4.6. Responding on time

The present study showed that, in the most frequent case, listeners have about 700 ms to plan before the end of the question. Assuming that responders most frequently achieve a gap of about 200 ms, this means that they have 900 ms planning time in total. Looking at picture naming studies, planning a single word takes about 600 ms (Indefrey & Levelt, 2004). However, longer utterances like simple sentences take longer (about 1500 ms, Griffin & Bock, 2000) which would already lead to a challenge to achieve 200 ms gaps. Moreover, it can be expected that planning might proceed somewhat less efficiently when upcoming responders are at the same time still (partly) listening to the ongoing turn (e.g., Barthel et al., 2016; Bögels et al., 2018). So how, then, can listeners generally achieve gaps of 200 to 300 ms? First, if planning a longer turn, they may (in some cases) decide to plan this incrementally, that is, they may first plan and utter one word and then go on to plan the rest of the utterance (e.g., Ferreira & Swets, 2002). Relatedly, responders may decide to utter filler words, such as 'uh' once they realize that it will take them too long to utter a content-full response. In these ways, listeners may try to preserve the short response latencies even if the planning time available is not exactly sufficient. Furthermore, although the most frequent turn transitions may lie around 200 ms, the distribution is generally much wider and also includes much longer gaps. The present study showed that such longer gaps are more likely to occur when planning time is shorter and when responses are longer (see e.g., examples 2 and 3 in Table 1). Future research could qualitatively examine responses for which planning times are relatively short and/or answers are relatively long to see how the corresponding gap lengths can be explained. One potential strategy responders may use to preserve short gaps is to keep (initial) responses short when planning time is limited (cf. Ferreira & Swets, 2002, who show that incremental production is under strategic control).

Note, furthermore, that individual differences are substantial here as well. Fig. S2 (in Supplementary Materials, Section 3) displays the overlaid density plots of all participants, showing that the most frequent response time differs per participant (at least between 100 and 600 ms or so). Thus, interlocutors may differ in how motivated they are to keep the gap as short as possible, and even in what as short as possible means for them. This may result in different strategies such as differing amounts of incremental planning (Ferreira & Swets, 2002).

5. Conclusion

Looking at questions from spontaneous interviews, as one form of conversation, the present study showed that coders can agree, to some extent, on the position in the question that allows responders to start planning their response. This position turned out to occur relatively early in the question, after only about one third of the question has passed. This means that responders have, in theory, quite some time to plan their response in overlap with the question, depending on the

actual question length. Moreover, the more time responders have to plan (in overlap with the question), the earlier they respond, which suggests that they indeed use this time to plan, and do not wait until the very end of the question to start planning.

In addition, in this ecologically valid corpus, a positivity was found in the ERPs quickly after responders could start planning, which was interpreted as a neural correlate of the start of response planning in more controlled studies (Bögels et al., 2018; Bögels, Magyari, & Levinson, 2015). This finding boosts the conclusion that production planning of the upcoming turn starts as soon as it can, even in natural circumstances. This suggests that overlap between comprehension and production, although perhaps costly, is not necessarily minimized. Instead, the gist of the current turn is continuously anticipated, with planning starting as soon as enough information has been gathered, likely because it is so important to respond on time in conversation.

CRediT authorship contribution statement

Sara Bögels: Conceptualization, Methodology, Formal analysis, Investigation, Writing - review & editing.

Acknowledgements

Thanks to Floor Arts, Dorine van Belzen, Annet Veenstra, Amy Abelmann, and Cielke Hendriks for their assistance during the experiment and with coding of the questions. Thanks to Marisa Casillas and the other members of the Interactional Foundations of Language and Dialogue projects at the Max Planck Institute for Psycholinguistics for extensive discussion of this work. Thanks to the members of the M3 meeting at the Radboud University for advice on linear mixed-effects modelling.

Funding

This work was supported by the European Research Council under an advanced grant (269484 INTERACT) to Stephen C. Levinson.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104347.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Barthel, M., & Sauppe, S. (2019). Speech planning at turn transitions in dialog is associated with increased processing load. *Cognitive Science, 43*(7), e12768. https://doi.org/10.1111/cogs.12768.

Barthel, M., Sauppe, S., Levinson, S. C., & Meyer, A. S. (2016). The timing of utterance planning in task-oriented dialogue: Evidence from a novel list-completion paradigm. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01858.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. Retrieved from https://github.com/lme4/lme4/http://lme4.r-forge.r-project.org/.

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer. Retrieved from http://www.praat.org.

Bögels, S. (2020). Neural correlates of turn-taking in free interviews/Raw Data. *The language archive (MPI Nijmegen)*https://hdl.handle.net/1839/72888153-5074-42c5-aa42-a63fe9259473.

Bögels, S., Casillas, M., & Levinson, S. C. (2018). Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question. *Neuropsychologia, 109*, 295–310. https://doi.org/10.1016/j.neuropsychologia.2017.12.028.

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. https://doi.org/10.1371/journal.pone.0145474.

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2019). Conversational expectations get revised as response latencies unfold. *Language, Cognition and Neuroscience, 0*(0), 1–14. https://doi.org/10.1080/23273798.2019.1590609.

Bögels, S., & Levinson, S. C. (2017). The brain behind the response: Insights into turn-taking in conversation from neuroimaging. *Research on Language and Social Interaction, 50*(1), 71–89. https://doi.org/10.1080/08351813.2017.1262118.

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5*, 12881. https://doi.org/10.1038/srep12881.

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics, 52*, 46–57. https://doi.org/10.1016/j.wocn.2015.04.004.

Boiteau, T. W., Malone, P. S., Peters, S. A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology. General, 143*(1), 295–311. https://doi.org/10.1037/a0031858.

Casillas, M., Bobb, S. C., & Clark, E. V. (2016). Turn-taking, timing, and planning in early language acquisition*. *Journal of Child Language, 43*(6), 1310–1337. https://doi.org/10.1017/S0305000915000689.

Casillas, M., De Vos, C., Crasborn, O., & Levinson, S. C. (2015, July 23). The perception of stroke-to-stroke turn boundaries in signed conversation. 315–320. https://pure.mpg.de/pubman/item/item_2161478_10/component/file_2161477/CasillasDeVos-CogSci2015-Final.pdf.

Corps, R. E., Crossley, A., Gambi, C., & Pickering, M. J. (2018). Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition, 175*, 77–95. https://doi.org/10.1016/j.cognition.2018.01.015.

Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. *Discourse Processes, 55*(2), 230–240. https://doi.org/10.1080/0163853X.2017.1330031.

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language, 82*(3), 515–535. https://doi.org/10.1353/lan.2006.0130.

Fairs, A., Bögels, S., & Meyer, A. S. (2018). Dual-tasking with simple linguistic tasks: Evidence for serial processing. *Acta Psychologica, 191*, 131–148. https://doi.org/10.1016/j.actpsy.2018.09.006.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84. https://doi.org/10.1016/j.brainres.2006.06.101.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language, 46*(1), 57–84. https://doi.org/10.1006/jmla.2001.2797.

Grandke, T. (1983). Interpolation algorithms for discrete Fourier transforms of weighted signals. *IEEE Transactions on Instrumentation and Measurement, 32*(2), 350–355. https://doi.org/10.1109/TIM.1983.4315077.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language, 25*(3), 601–634. https://doi.org/10.1016/j.csl.2010.10.003.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274–279. https://doi.org/10.1111/1467-9280.00255.

Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., … Schoffelen, J. (2013). Good practice for conducting and reporting MEG research. *NeuroImage, 65*, 349–363. https://doi.org/10.1016/j.neuroimage.2012.10.001.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555–568. https://doi.org/10.1016/j.wocn.2010.08.002.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00255.

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001.

Jensen, O., Gelfand, J., Kounios, J., & Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral Cortex, 12*(8), 877–882. https://doi.org/10.1093/cercor/12.8.877.

Kendrick, K. H., & Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes, 52*(4), 255–289. https://doi.org/10.1080/0163853X.2014.955997.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643.

Kubose, T. T., Bock, K., Dell, G. S., Garnsey, S. M., Kramer, A. F., & Mayhugh, J. (2006). The effects of speech production and speech comprehension on simulated driving performance. *Applied Cognitive Psychology, 20*(1), 43–63. https://doi.org/10.1002/acp.1164.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205. https://doi.org/10.1126/science.7350657.

Levinson, S. C. (1983). *Pragmatics.* Cambridge: Cambridge 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. https://doi.org/10.1016/j.tics.2015.10.010.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00731.

Magyari, L., Bastiaansen, M. C. M., de Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive*

Neuroscience, 26(11), 2530–2539. https://doi.org/10.1162/jocn_a_00673.
Magyari, L., & De Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology, 3*(376), 1–9.
Magyari, L., De Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.00211.
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024.
Meyer, A. S., Alday, P. M., Decuyper, C., & Knudsen, B. (2018). Working together: Contributions of corpus analyses and experimental psycholinguistics to understanding conversation. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00525.
Oostdijk, N. (2000). *Het Corpus Gesproken Nederlands. 5*, 280–284.
Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Intelligence and Neuroscience, 1*(1–1), 9. https://doi.org/10.1155/2011/156869 2011.
Piai, V., Roelofs, A., Rommers, J., Dahlslätt, K., & Maris, E. (2015). Withholding planned speech is reflected in synchronized beta-band oscillations. *Frontiers in Human Neuroscience,* 9. https://doi.org/10.3389/fnhum.2015.00549.
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.
Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology, 56*(6), e13335. https://doi.org/10.1111/psyp.13335.
Schegloff, E. A. (2007). *Sequence organization in interaction: Volume 1: A primer in conversation analysis.* Cambridge University Press.
Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension—An fMRI study. *Cerebral Cortex, 22*(7), 1662–1670. https://doi.org/10.1093/cercor/bhr249.
Shitova, N., Roelofs, A., Coughler, C., & Schriefers, H. (2017). P3 event-related brain potential reflects allocation and use of central processing capacity in language production. *Neuropsychologia, 106*, 138–145. https://doi.org/10.1016/j.neuropsychologia.2017.09.024.
Sichlinger, L., Cibelli, E., Goldrick, M., & Mittal, V. A. (2019). Clinical correlates of aberrant conversational turn-taking in youth at clinical high-risk for psychosis. *Schizophrenia Research, 204*, 419–420. https://doi.org/10.1016/j.schres.2018.08.009.
Sjerps, M. J., & Meyer, A. S. (2015). Variation in dual-task performance reveals late initiation of speech planning in turn-taking. *Cognition, 136*, 304–324. https://doi.org/10.1016/j.cognition.2014.10.008.
Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … others (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences, 106 (26), 10587–10592.
Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015.