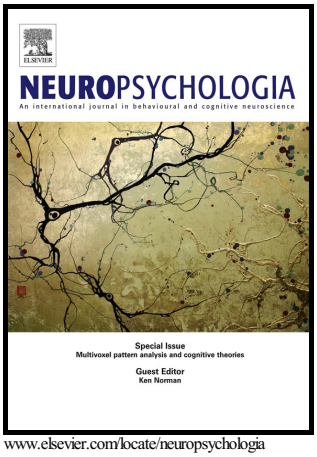
Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question

Sara Bögels, Marisa Casillas, Stephen C. Levinson



www.elsevier.com/locate/neuropsychologia

PII: S0028-3932(17)30502-X
DOI: https://doi.org/10.1016/j.neuropsychologia.2017.12.028
Reference: NSY6619

To appear in: *Neuropsychologia*

Received date: 29 August 2017
Revised date: 13 November 2017
Accepted date: 15 December 2017

Cite this article as: Sara Bögels, Marisa Casillas and Stephen C. Levinson, Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question, *Neuropsychologia,* https://doi.org/10.1016/j.neuropsychologia.2017.12.028

Planning versus comprehension in turn-taking: Fast responders show reduced anticipatory processing of the question

Sara Bögels[1,2]*, Marisa Casillas[1], Stephen C. Levinson[1,2]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[2]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

*Corresponding author, Email: s.bogels@donders.ru.nl. Donders Institute, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

Abstract

Rapid response latencies in conversation suggest that responders start planning before the ongoing turn is finished. Indeed, an earlier EEG study suggests that listeners start planning their responses to questions as soon as they can (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). The present study aimed to (1) replicate this early planning effect and (2) investigate whether such early response planning incurs a cost on participants' concurrent comprehension of the ongoing turn. During the experiment participants answered questions from a confederate partner. To address aim (1), the questions were designed such that response planning could start either early or late in the turn. Our results largely replicate Bögels et al. (2015) showing a large positive ERP effect and an oscillatory alpha/beta reduction right after participants could have first started planning their verbal response, again suggesting an early start of response planning. To address aim (2), the confederate's questions also contained either an expected word or an unexpected one to elicit a differential N400 effect, either before or after the start of response planning. We hypothesized an attenuated N400 effect after response planning had started. In contrast, the N400 effects before and after planning did not differ. There was, however, a positive correlation between participants' response time and their N400 effect size after planning had started; quick responders showed a smaller N400 effect, suggesting reduced attention to

comprehension and possibly reduced anticipatory processing. We conclude that early response planning can indeed impact comprehension processing.

Keywords: Turn-taking, EEG, N400, Language comprehension, Language production, Prediction

1. Introduction

Speakers in conversation appear to effortlessly achieve smooth, rapid transitions between turns at talk. At least two major psycholinguistic processes underlie these fluent transitions: comprehension of incoming speech and initial planning for producing the upcoming response. At first sight it may appear that these two processes occur sequentially; you may feel that you generally begin producing a response only after your interlocutor has finished speaking. However, estimates from psycholinguistic picture naming studies indicate that speakers need at least about 600 ms to plan a single word (Indefrey & Levelt, 2004) and much longer (about 1500 ms) to plan a simple sentence (Griffin & Bock, 2000). These figures are clearly *much* higher than the typical time between two turns at talk in conversation (e.g., Heldner & Edlund, 2010; Levinson & Torreira, 2015; Sacks et al., 1974). To respond with the quick timing typical of conversation (0-200 ms, e.g., Stivers et al., 2009), responders must therefore begin planning their responses while the previous turn is still unfolding. In other words, there must be some overlap between comprehension of the incoming turn and planning of the upcoming response for addressees during conversation.

In a recent EEG study, two novel neural correlates of speech planning suggested that listeners start planning their own turn as soon as they have enough information to do so (Bögels et al., 2015a). The timing of these neural correlates indicates that there is more overlap between comprehension and production than would be expected if listeners postponed response planning for as long as possible. We hypothesize that such extensive dual-tasking—in the form of simultaneous comprehension and production planning—might come at a cost. The present study's aim is two-fold: (1) to replicate the presence and timing of neural correlates for response planning found in Bögels and colleagues' (2015a) study, and (2) to investigate whether comprehension of the incoming turn suffers during simultaneous production planning of an upcoming turn.

1.1 Aim 1: The Timing of Response Planning

Roughly two different types of models on the timing of response planning in turn-taking can be distinguished, early planning and late planning models (for reviews, see Bögels & Levinson, 2017; Corps et al., 2017). Early planning models assume that listeners start planning as soon as they have enough information to do so, whereas late planning models state that listeners will wait as long as they can and only start planning just before the end of the current turn. A few studies have directly attempted to distinguish between these models. Of these, two have used a dual-task paradigm in which participants were asked to take turns at talk while simultaneously performing an additional task. One study asked participants to spontaneously converse with someone while also using a mouse to track a visual target on a computer monitor (Boiteau et al., 2014). Participants' ability to track the visual target was most impacted just before and during speaking, suggesting that listeners start planning only just before the end of the previous turn. Another study asked participants to continuously tap their fingers in a specific pattern while they labeled rows of pictures in alternation with a pre-recorded voice (Sjerps & Meyer, 2015). As in the mouse-tracking experiment, participants' finger-tapping performance decreased before and during speech, beginning about 500 ms before the offset of the pre-recorded speech. In a second experiment, Sjerps and Meyer (2015) replicated the first experiment while also recording participants' eye movements. Their eye-tracking results supported the finger-tapping findings: listeners began to look at the pictures that they were about to name only just before finger-tapping performance started to decrease. These two studies thus appear to suggest that listeners postpone their planning of the upcoming utterance for as long as possible and start planning only when the time to respond is imminent.

A number of factors in the two studies described above may have affected the apparent timing of participants' response planning. Sjerps and Meyer (2015) argue that their method measures the most cognitively demanding aspects of planning, but it may be that these kinds of motor tasks (finger tapping and mouse tracking) primarily interfere with the execution of another motor task—speaking—and not the entire arc of response planning. Additionally, while Boiteau and colleagues (2014) used an ecologically valid task (natural conversation), they did so at the cost of not being able to control the speech being comprehended and produced. For example, we do not know when during the interlocutor's turn the participant had enough information to start planning his or her response. On the other hand, Sjerps and Meyer (2015) very closely controlled the speech in their experiment by limiting it to formulaic descriptions of rows of images. However, their paradigm deviated

quite far from typical conversational interaction in that participants knew that they were talking to a pre-recorded voice and that there was no contingency between the spoken content of their and the computer's turns. In other words, participants were just alternating their speech with the computer's without the computer's speech having any bearing on their own speech plans and vice versa.

Another recent study used an interactive EEG paradigm to look into the processes of production planning during listening while participants were engaged in turn-taking (Bögels et al., 2015a). Using EEG precluded the need for an additional task (e.g., finger tapping or mouse tracking). Furthermore, the moment during each turn at which participants could first start planning their response was directly manipulated. The study employed quiz questions containing both crucial information for answering the question (e.g., *007* in examples 1 and 2 below) and more general information (e.g., *famous movies* in examples 1 and 2 below). The crucial information either appeared in the middle of the question (see example 1) or at the very end (see example 2), affecting when participants could begin to plan their answer.

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

The situation was also truly interactive for participants, who were asked to answer each quiz question posed by their interlocutor (the 'quiz master'), who was sitting in an adjacent room. In reality, the quiz master's questions had been pre-recorded for maximal control across participants, but the same quiz master did provide live feedback on their answers, leading participants to believe the entire interaction was live. For Bögels and colleagues (2015a), a challenge in using EEG was that the hypotheses were rather exploratory with respect to the neural correlates of production planning; few EEG studies have attempted live language production paradigms so far. For this reason, a second version of the experiment had no production component. This time participants listened to and remembered the questions without actually answering them (a "no-response" version). Two neural signatures were elicited in in the response-planning version of the experiment that were absent or much reduced in the no-response version of the experiment. The first was a large positivity in the ERPs starting around 500 ms after the onset of the crucial word (e.g., *007* in the examples above) that localized to language production areas in the brain. This was interpreted as reflecting production planning directly. The second neural correlate was a decrease in alpha power occurring with the same timing, which was localized to occipital and parietal areas.

This effect was interpreted as reflecting 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 production planning. Most importantly, the timing of these two components suggested that listeners started planning their response within 500 ms of the point when response planning became possible (e.g., when recognizing *007* in examples 1 and 2 above). Note that this point occurred an average of 2.4 seconds before the end of the question for the condition in which it was possible for participants to plan their responses early (see Bögels et al., 2015a, Table 1). These results then suggest that participants begin planning their responses as early as possible, contrasting with the conclusions from the other response-planning studies described above (Boiteau et al., 2014; Sjerps & Meyer, 2015).

One other recent study used eye-tracking and spoken response latencies to look at the same question (Barthel et al., 2016). Participants saw a set of objects and listened to a confederate who named a subset of those objects before they themselves subsequently named the remaining ones. The confederate's utterance could either end with an object label (that critically affected the participant's response plan) or an extra word (that did not critically affect the participant's response plan). Results showed that participants gazed at the first object they were going to name as soon as the last object label of the incoming turn could be recognized, supporting the idea that listeners start planning responses as soon as they are able. However, this point occurred relatively close to the end of the question in this study. Barthel and colleagues (2016) argue that the difference with the study by Sjerps et al. (2015) might lie in strategic effects; in some cases participants might opt for an early planning strategy whereas in other cases they choose a late planning strategy. Barthel and colleagues (2016) further argue that a late planning strategy might be more likely when social pressures are low, such as when no genuine interactive situation is present or when the responses are not relevant for a listener (as in Sjerps & Meyer, 2015).

Given these mixed results of previous studies on this topic, the first aim of the present experiment was to replicate the EEG results by Bögels et al. (2015a) with an adapted but still interactive paradigm (see section 1.3: The Present Study).

1.2 Aim 2: Planning versus Comprehension

In the remainder of this Introduction, we will assume that planning starts as soon as possible (in line with Bögels et al., 2015a) or that at least substantial overlap exists between comprehension of the incoming turn and production planning of the upcoming one, possibly followed by buffering of the pre-planned response in memory (see, e.g., Corps et al., 2017). If

so, it forces us to consider what possible consequences this overlap might have on the various sub-processes involved in taking turns in conversation. There are several indications in the literature that a production-comprehension overlap might result in less-than-optimal processing on either of the two tasks. Recent work shows that production planning requires sustained attention, especially in a dual-task situation (Jongman et al., 2015). Another study suggests that planning speech while hearing words affects later memory of those words, suggesting there is competition for attention both when comprehending words and when encoding them for memory (Gerakaki et al., unpublished data). One fMRI study accordingly found that the brain regions involved in semantic, lexical, and syntactic processing are mostly overlapping for language comprehension and production (Menenti et al., 2011). Another study used a repetition suppression paradigm to show that, not just the same brain regions, but the same populations of neurons appear to be involved in the production and comprehension of syntactic structures (Segaert et al., 2011). If indeed the same brain areas—even the same neural populations—are involved in these two processes, it is difficult to conceive of how comprehension of one turn and production planning of the next turn could proceed in parallel without some difficulty or loss of efficiency in one or both processes.

Prior work indeed already gives some indication that production planning is less efficient during comprehension (Barthel et al., 2016; Bögels et al., 2015a; Magyari et al., 2017). These studies all show faster response times when planning can start earlier. However, the gain in response time is generally much smaller than the extra time available, even taking into account an avoidance of vocal overlap between speakers. In other words, the ability to plan a response 600 ms earlier does not usually result in a response that begins 600 ms earlier. Moreover, eye-tracking data (Barthel et al., 2016) show that proportions of looks to objects that were to be named, increased more slowly when the turn was still ongoing than when there was no concurrent incoming material. Production planning therefore appears to proceed less efficiently while speakers simultaneously listen to incoming speech. But is the complement to this finding true as well? Is comprehension of incoming speech also affected when speakers simultaneously begin to plan a response? Answering this question is the second aim of the present study.

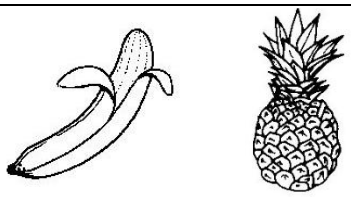
1.3 The Present Study

As said, the first aim of the present study is to replicate the earlier EEG study (Bögels et al., 2015a) with a slightly different paradigm (see below). We hypothesized that we would find similar neural correlates in response to the word that allows participants to start production

planning: a positivity in the ERPs and a reduction in alpha power in the time-frequency analyses, with a similar distribution, localization, and timing. However, given the small changes in the paradigm and the different stimuli in the present study, we might also find small differences in results. To identify robust effects, those not affected by these methodological differences, we analyzed the data using the same analyses as in the Bögels et al. (2015a) study. As described above, in that study the main experiment was compared with a no-response control experiment to establish whether the neural correlates were related to response planning per se. Given that the control experiment yielded absent or much reduced neural correlates, establishing them as relevant to response planning, we only included a response-planning experiment in the present study.

The paradigm was slightly adapted to address our second aim, which was to investigate whether comprehension of the incoming turn suffers when interlocutors are simultaneously planning a response. As in Bögels et al. (2015a), participants were led to believe that they were engaged in a live question-answer interaction with a partner. In reality, they only heard pre-recorded questions during the experiment. Different from the original study, participants' possible answers on each trial were limited to objects that they could see on a screen (e.g., a banana and a pineapple, see Table 1, top row) so that they would have expectations on each trial about the type of information that might be mentioned in their partner's question (e.g., "fruit"). The task was interactive: participants' responses were contingent on the information from the confederate on each trial and they believed their own responses contingently affected which object the confederate selected (see section 2: Methods for further details).

Table 1. Examples of the four conditions. An example of the visual display is given in the top row, followed by examples of the questions which were asked auditorily in the four conditions (A–D; with English gloss in italics). Critical informative words are printed in bold and critical uninformative words (expected or unexpected) are underlined.

		
early-planning, expected-word	A	Welk object is **krom** en wordt als <u>fruit</u> gezien? *Which object is **curved** and is considered to be a type of <u>fruit</u>?*
early-planning,	B	Welk object is **krom** en wordt als <u>gezond</u> gezien?

unexpected-word		*Which object is **curved** and is considered to be healthy?*
late-planning, expected-word	C	Welk object wordt als fruit gezien en is **krom**? *Which object is considered a type of fruit and is **curved**?*
late-planning, unexpected-word	D	Welk object wordt als gezond gezien en is **krom**? *Which object is considered to be healthy and is **curved**?*

As in Bögels et al. (2015a), we controlled when participants could begin to plan their response: either midway through the question or at the end. In the present study, each question contained two pieces of information for choosing between two pictures (e.g., a banana and a pineapple): one informative (e.g., *curved*) and the other uninformative (e.g., *fruit*). In the 'early planning' conditions (A and B in Table 1), the informative cue (*curved*) occurred in the middle of the question and in the 'late planning' conditions (C and D in Table 1) it occurred at the end of the question, similar to the prior study. We expected to replicate the neural correlates found in the earlier study for (early) response planning: a positivity in the ERPs and an alpha decrease in the time-frequency domain, both around 500 ms after onset of the informative word (e.g., *curved*).

New to the present study, we manipulated whether the uninformative information was expected or unexpected. For example, when viewing a banana and a pineapple (see Table 1), the word *fruit* (see Table 1, conditions A and C) is more expected than the word *healthy* (see Table 1, conditions B and D) even though both refer equally to the visible pair of objects (see section 2: Methods for details). These expected and unexpected word conditions were created to induce an N400 effect between them. The N400 is an index of the expectedness of a word in its context (Kutas & Hillyard, 1980).

Our experiment relies on the idea that the unexpected word will elicit a larger N400 than the expected word—the difference between them is called the *N400 effect*. Furthermore, we assume that the N400 effect is modulated by attention. In their review paper of the N400, Kutas and Federmeier (2011) argue that though a lack of attention to semantic processing of a stimulus might not eliminate the N400 effect, it clearly diminishes it. For example, a recent study showed that the size of the N400 effect was affected by task instructions, indicating it is not automatic and relies on attention at least to some extent (Brothers et al., 2017). Another study showed that unexpected versus expected words in focused position (as indicated by the context and by prosodic prominence) elicit a larger N400 effect than the same words in non-focused position, presumably because the former receive more attention (Wang, Bastiaansen, Yang, & Hagoort, 2011). One possible mechanism for this modulation by attention is through

anticipatory processes[1], as follows: unanticipated words elicit fairly large N400s but, as anticipation of a word increases (e.g., from context), words elicit smaller N400s. That is, N400 attenuation may rely on attention being allocated to anticipate upcoming words. If, for some reason, listeners' attention is held elsewhere, they may not be able to anticipate upcoming words as well, leading to a less attenuated (i.e., larger) N400. By extension, the N400 effect (the difference between expected and unexpected words) should be smaller in a situation of less attention.

Crucially, in the present study, then, the presumed N400 effect occurs in different positions in the early-planning questions (conditions A and B in Table 1) and the late-planning questions (conditions C and D in Table 1). In late-planning questions, the expected or unexpected word occurs in the middle of the turn, before the answer becomes known and before planning can begin, presumably still receiving full attention. By contrast, in early-planning questions, the answer becomes known first, in the middle of the turn, and the expected or unexpected word occurs afterwards, near the end of the question, and therefore possibly in overlap with response planning. So, if concurrent production planning indeed takes away attentional resources from comprehension, we hypothesize that the N400 effect (difference) between expected and unexpected words should be smaller after planning has started (i.e. in the early planning conditions; A and B in Table 1), because comprehension presumably overlaps with production planning.

2. Methods

2.1 Ethical Approval

All experiments were carried out in accordance with guidelines approved by the *Ethics Committee Faculty Social Sciences* of Radboud University in Nijmegen.

2.2 Participants

Thirty-three participants were recruited from the participant pool of the Max Planck Institute for Psycholinguistics. Data from one participant was excluded from analysis because of too many artifacts (see section 2.7: Data Analysis). The 32 remaining participants (7 male, 25 female) had a mean age of 21.1 years old (range 18 to 24). All participants were right-handed

[1] We use the term anticipation in a broad sense, referring to any way in which the current state of the system affects processing of the incoming information (see, e.g., Kuperberg & Jaeger, 2016). This is compatible with an interpretation of the N400 both in terms of 'prediction' or 'integration'.

native speakers of Dutch without hearing impairments. They gave informed consent before participating and received 8 euros per hour for their participation.

2.3 Materials

We selected pairs of object drawings from the Snodgrass picture set (Snodgrass & Vanderwart, 1980) in which each pair had at least one shared trait (e.g., *fruit* for the banana-pineapple pair) and at least one distinctive trait that could be used to single out one of the objects (e.g., *curved* for the banana). In a pre-test we then elicited shared traits for each pair (see Supplementary Materials) and constructed questions describing one of the two objects in each pair using the distinctive trait (e.g., *curved*) as the informative cue and a frequent (e.g., *fruit*) and infrequent shared trait (e.g., *healthy*) as the expected and unexpected uninformative cues, respectively. This lead to 120 items in 4 different conditions each (see Table 1 for an example and Table S1 for all items).

Our main experiment depends on the idea that participants do not know which object to respond with until they hear the informative cue (e.g., *curved*), so we verified in a second pre-test that the informative cues (e.g., *curved*)—but not the uninformative cues (e.g., *fruit*/*healthy*)—could be used to identify one of the objects in each item (see Supplementary Materials). Finally, our confederate participant pre-recorded each of the four questions (conditions A–D; Table 1) for each of the resulting 120 items (object-pairs), plus the filler questions (see section 2.4: Procedure), in the EEG chamber where the experiment would take place so that there would not be a noticeable shift in background noise when the confederate switched from live to prerecorded speech during the experiment.

2.4 Procedure

We measured participants' EEG as they took part in an interactive dialogue task with a confederate. We took care to ensure that participants conceived of the confederate as another participant until the study had ended. When participants were first invited to take part in the study, they were told that it required *pairs* of participants and that a second participant would therefore join them on the testing day. The same message went on to explain that only one person from each pair would be wearing an EEG cap and that this role was randomly assigned within each pair. In reality, all participants were assigned to the 'EEG participant' role. This was explained to them during post-experiment debriefing, when we revealed the

confederate's role. Once participants arrived and had gone through consent, EEG set-up began. Toward the end of this process, the confederate arrived and began her own consent paperwork.

Once both the participant and confederate were ready to proceed, the experimenter explained how the task worked, following the cover-story in which the confederate was treated as a participant. In a nutshell, the pair needed to work together to choose a target object from an array of objects on each trial. They sat at separate computer monitors, each of which displayed insufficient information to pick out the target object. Each trial unfolded over a series of steps (Figure 1): First, each monitor would show a fixation cross, then a set of object pictures. The participant's (P) screen always showed only two of the four objects they thought were shown on the confederate's (C) screen. Next, P's objects would disappear (replaced by a fixation cross) while C's monitor displayed text cueing her to ask specific information about P's objects (e.g., 'Welk object wordt als fruit gezien en is krom?'—*Which object is considered a type of fruit and is curved?*). Note again that C's monitor didn't actually change during this process; P was just made to believe that these steps were taking place in each trial. Then, given C's question, P would name the object that fulfilled these properties ('banaan'—*banana*) and would then hear C click on the object named by P. The click would initiate the next trial.

P was thus led to believe that C could not pick out the target object without P's response. This is a crucial aspect of the design because it enhances the illusion of interactivity and the relevance of P's response in the interaction. This design also ensured that P could not see any objects when listening to and answering C's question, since the objects had disappeared by then. We made this decision to reduce eye movement artifacts during the participant's response-planning phase. The pair completed three practice trials sitting side-by-side so that P could understand what C supposedly saw throughout each trial. Then P entered a sound-attenuated isolation chamber while the C stayed outside.

Once P entered the isolation chamber, he or she completed another ten practice trials followed by 136 test trials. While in the chamber, P only heard pre-recorded questions from C that were made to sound as if they were spoken live (Figure 1, rightmost panel). C stayed on stand-by throughout the experiment in case there was need for live interaction at any point (e.g., if there was a question between trials). The timing of the trials was as follows: P saw a fixation cross for 1000 ms, followed by a 3000 ms presentation of the two pictures, and then another fixation cross. The pre-recorded question started within a random interval between 500 and 1500 ms after the onset of the second fixation cross. When P's answer was complete

(e.g., 'banaan'), the experimenter monitoring the task pressed a button that initiated a randomly selected 500–2000 ms delay before playing the sound of C's button click. The button click sound was pre-recorded because C was not actually selecting images following P's response. Simultaneously with the click-sound, a blink signal appeared on P's screen for 2000 ms indicating that they could blink and rest their eyes. To increase the authenticity of the task, C's verbal cues on 16 of the test trials contained disfluencies. The total of 136 test trials therefore included 16 disfluent trials and the 120 fluent target trials, taken from the pre-tests (see section 2.3: Materials and Supplementary Materials). The 16 disfluent trials were considered to be filler trials and were excluded from all analyses.

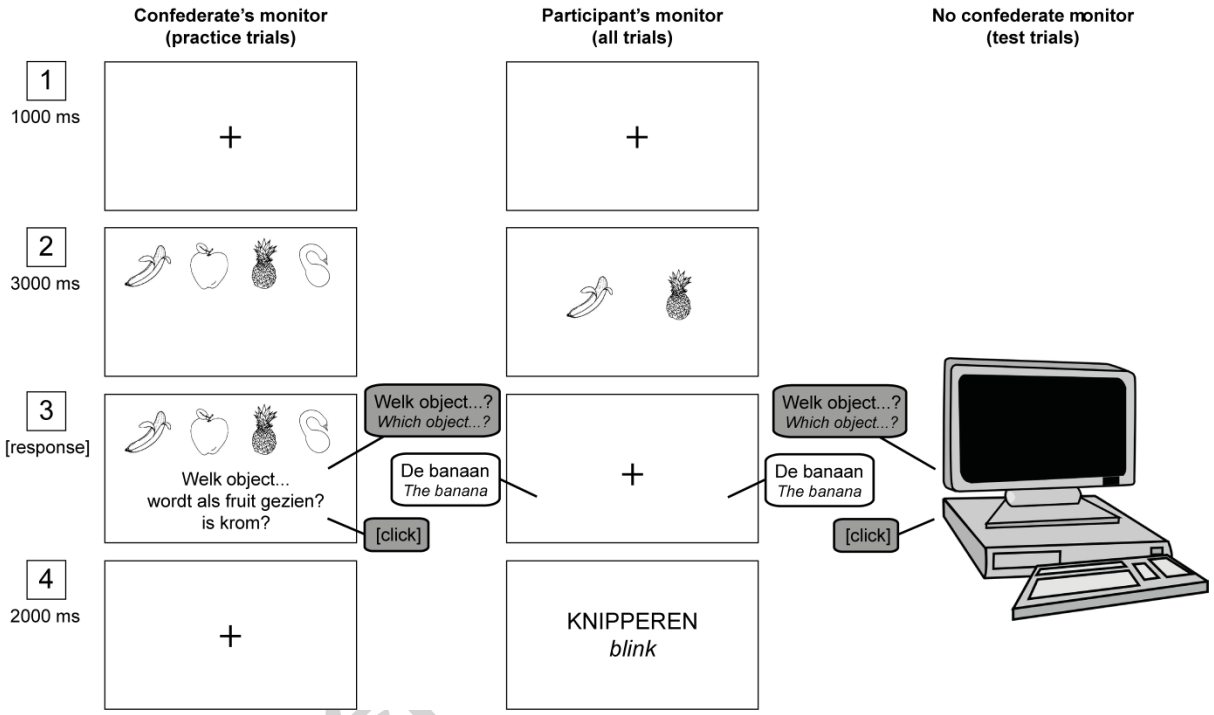


Figure 1. Example of trial structures. The three-part structure for practice trials (live confederate: left and middle panel) and test trials (pre-recorded confederate: middle and right panel).

The experiment lasted about one hour in total. When participants had completed all trials, they finished up with a short survey about the task, including whether or not they had suspected that their partner's speech was pre-recorded. Two participants considered this possibility at some point during the experiment (one suspected the confederate from the start). We performed additional analyses without these two participants which yielded qualitatively similar results. When asked directly to guess whether their partner was a confederate or not, 8 more guessed that it was (only two participants in total suspected

recorded speech). Most participants were surprised to learn that their partner had been a confederate.

2.5 Design

The experimental questions could differ on the factors Planning (early-planning, late-planning) and Expectedness (expected-word, unexpected-word). See Table 1 for an example of one item in the four conditions. Four lists were created, each of which were administered to a quarter of the participants. All lists contained all 120 items only once, in one of the four conditions, 30 items per condition. In addition, each list contained 16 filler items, 4 per condition. The conditions per item were rotated over the 4 lists in a Latin square design. All lists contained the items in the same order (except for changes in the conditions). The experiment was divided into 4 blocks of 34 questions each with pauses between blocks. Conditions and filler items were divided evenly over the blocks. The same condition appeared maximally two times in a row and filler trials were always separated by at least two other trials.

2.6 Apparatus

EEG was recorded from 61 active Ag/AgCI electrodes using an actiCap (e.g., Bögels et al., 2015a). Of these, 59 electrodes were mounted in the cap with equidistant electrode montage referenced to the left mastoid. Two separate electrodes were placed at the left and the right mastoid. Blinks were monitored through a separate electrode placed below the left eye and one of the 59 electrodes in the cap. Horizontal eye movements were monitored through two separate electrodes placed at each outer canthus. The ground electrode was placed on the forehead. Electrode impedance was kept below 10 kΩ. EEG and EOG recordings were amplified through BrainAmp DC amplifiers. EEG signals were filtered online with a band-pass filter between 0.016 and 100 Hz. The recording was digitized online with a sampling frequency of 500 Hz and stored for offline analysis.

2.7 Data Analysis

First, participants' answers were coded for errors, which were defined as non-responses, naming the wrong picture, or using a name that could not be attributed to one of the pictures. If it was clear from the name that the correct picture was meant, the response was coded as correct. Second, response times from the end of the question to the start of the answer were measured. Mixed-effects models were run on errors, hesitations, and response times to assess

the effect of Planning (early-planning, late-planning), Expectedness (expected-word, unexpected-word) and their interaction using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2014). In each of these models we used the maximal random slopes structure justified by the design (Barr et al., 2013). Only correct responses without hesitations were entered into the response time analysis.

Preprocessing and statistical analysis of EEG data was conducted using Fieldtrip (Oostenveld et al., 2011). The epochs in which participants were instructed not to blink were relatively long (longer than in Bögels et al., 2015a) and participants were speaking during the experiment, which led to noisy data in general. Therefore, we used an ICA approach to artifact removal to retain enough trials.

Trials with incorrect answers were discarded before EEG analysis. Each question contained two critical positions: one where the answer became known ('planning position') and one where either an expected or an unexpected word was presented ('N400 position'). Epochs were extracted from 500 ms before the start of a critical position until maximally 1500 ms after, but were always cut off at 100 ms before speech onset to avoid speech artifacts. Then, PCA was used to reduce data dimensionality for each participant to 40 components, which were then subjected to ICA (Gross et al., 2012; 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 3.6 (range: 2–8). 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 with a window of 200 ms immediately before the critical position. Epochs still containing eye artifacts or other artifacts that exceeded +/- 100 μV were discarded. As mentioned (section 2.2: Participants), one participant with less than 20 trials remaining in one condition was not analyzed further. This procedure resulted in an average of 26.7 remaining trials per condition per participant for the 'planning position' (range: 21–30 out of 30) and 26.3 remaining trials per condition per participant for the 'N400 position' (range: 20–30). The difference in number of remaining trials between to-be-compared conditions was maximally one trial.

These preprocessed data were then entered into event-related potential (ERP; used in all analyses) and time-frequency analyses (TF; used in only the replication analyses). For ERPs, epochs were filtered with a low-pass filter of 35 Hz and baselined with a window of 200 ms immediately before the critical position. Then, trials of the same condition were averaged per participant. 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 using a Hanning taper (Grandke, 1983) with a window of 500 ms for each frequency (the same as Bögels et al., 2015a). 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 Figure 5).

To test for statistically significant differences between conditions and reduce the multiple-comparison problem, we used the cluster-based approach (Maris & Oostenveld, 2007) implemented in the Fieldtrip toolbox for the ERP as well for the TF analysis. As in Bögels et al. (2015a), clusters were formed in time, space (neighboring electrodes), and frequency (for TF analyses) and 1000 randomizations were used for the permutation distribution. The critical alpha level was fixed to .05 (one-sided, given our hypotheses based on Bögels et al., 2015a). For significant clusters, we report *sum-t* statistics (the sum of all t-values in the cluster) and *p*-values. This robust cluster-based approach reduces the multiple-comparisons problem and controls family-wise error across participants in time and space (see Bögels et al., 2015a, for an elaborate description of this method). Analyses for all critical positions were performed within a time-range of 0–1500 ms for ERP analyses and 0–1200 ms for TF analyses.

To keep the replication analyses as close as possible to the original conditions in Bögels and colleagues' (2015a) study, the replication comparisons were only made between questions with expected words in them (conditions A and C in Tables 1 and 2). To evaluate the effects of the answer becoming known early in the turn, we compared the early-planning condition (A) at the planning position to the late-planning condition (C) at the N400 position (i.e., at the onset of the expected word in the first part of the question, see positions marked with a superscript '1' in Table 2). To evaluate the effect of the answer becoming known late, we compared the late-planning condition (C) at the planning position; to the early-planning condition (A) at the N400 position (i.e., at the onset of the expected word in the second part of the question, see positions marked with a superscript '2' in Table 2). Given that we only use the expected-word conditions in these analyses, we simply refer to the early-planning and the late-planning conditions for the replication-related findings in the Results.

Table 2. Time-locking positions. Positions used for time-locking in the two types of analyses are marked with vertical lines and a superscript number. Positions with the same

superscript numbers were directly compared in the analyses of conditions (A/C—replication analyses, A/B—N400 analysis 1, and C/D—N400 analysis 2).

	Replication Analyses	
early-planning, expected-word	A	Welk object is |[1] krom en wordt als |[2] fruit gezien? *Which object is curved and is considered to be a type of fruit?*
late-planning, expected-word	C	Welk object wordt als |[1] fruit gezien en is |[2] krom? *Which object is considered a type of fruit and is curved?*
	N400 Analyses	
early-planning, expected-word N400 after planning	A	Welk object is krom en wordt als |[3] fruit gezien? *Which object is curved and is considered to be a type of fruit?*
early-planning, unexpected-word N400 after planning	B	Welk object is krom en wordt als |[3] gezond gezien? *Which object is curved and is considered to be healthy?*
late-planning, expected-word N400 before planning	C	Welk object wordt als |[4] fruit gezien en is krom? *Which object is considered a type of fruit and is curved?*
late-planning, unexpected-word N400 before planning	D	Welk object wordt als |[4] gezond gezien en is krom? *Which object is considered to be healthy and is curved?*

For the N400 analyses, we first separately analyzed the N400 effect after planning (in early-planning questions; A/B) and before planning (in late-planning questions; C/D), comparing expected and unexpected words (at the same positions, see positions marked with superscripts '3' and '4' for the N400 after and before planning, respectively, in Table 2). We employed two-step analyses for emulating the interaction between Expectedness and Planning (see, e.g., Bögels et al., 2015b). We first calculated a *t*-statistic for the difference between the unexpected- and expected-word conditions, separately before and after planning. Then, we included the outcomes (*t*-values) of this first step statistic into a group statistic that compared the N400 effect before versus after planning. The comparison at the group level followed the cluster-based statistics approach described above. Next to this two-step analysis, we also report the more standard analysis based on direct differences between unexpected- and expected-word conditions in footnote 2.

Two different correlation analyses were performed over participants with each participant's average response time (over all four conditions) as one of the variables. First, we correlated participants' average response times with the average size of their N400 effects in the 300–500 ms window in one representative electrode (Cz, see small head in Figure 8; see Figure S1 for correlations in a larger set of electrodes). Second, we correlated their response

times with the average size of their positive effect (difference between conditions) in a 600–900 ms window in a representative electrode (see small head in Figure 9; see Figure S2 for correlations in a larger set of electrodes).

To identify sources underlying the electrode-level effects (only for replication analyses), a BEM (boundary element headmodel; Oostenveld et al., 2001) was used based on a template MRI aligned with the EEG electrode array. Parameters for the source analysis of ERP effects were chosen based on the earlier study (Bögels et al., 2015a) and significant electrode-level effects (600–1100 ms) and for frequency analyses based on significant electrode-level effects (alpha: 600–1200 ms and 8–14 Hz; beta: 500–800 ms and 16–20 Hz). ERP sources were identified using a Linearly Constrained Minimum Variance (LCMV) beamformer (Van Veen & Buckley, 1988) where we calculated a common LCMV filter for the two conditions together per participant. This common filter was then used to transform the participants' ERP signals into source (voxel) space for comparisons between conditions. For identifying generators of oscillations we employed Dynamic Imaging of Coherent Sources (DICS) beamformers (Gross et al., 2001) and also used common filters. Power values were calculated on an equidistant template 3D grid with a 1 cm resolution. Otherwise no anatomical constraints were imposed on the source localization. A regularization parameter (lambda) of 5% was used in both LCMV and DICS analyses. For statistical testing of the source-localizations underlying ERP and TF effects, we used the same cluster-based approach, in this case only clustering over voxels. For plotting purposes, the significant results were interpolated on a template brain based on the same anatomy from which the headmodel was created.

3 Results

3.1 Behavioral results

The percentage of errors in naming the right object was low: 3.9% over all conditions. A logistic mixed-effects model with errors as the dependent variable, Planning (early-planning, late-planning), Expectedness (expected-word, unexpected-word) and their interaction as the main predictors, and random intercepts for participant and item, showed that the likelihood of errors did not differ across conditions. After removing errors, 5.8% of the remaining data had hesitations (filled pauses or partial repeats) in them. A logistic mixed-effects model with hesitations as the dependent variable, Planning (early-planning, late-planning), Expectedness (expected-word, unexpected-word), and their interaction as the main predictors, random intercepts for participant and item, and random slopes of Planning, Expectedness, and their

interaction for participants, showed that the likelihood of hesitations did not differ across conditions. Hesitations were removed for the analysis of response times. Response latencies were right-skewed, as is typical of responding in conversational contexts (Stivers et al., 2009). To better meet the assumptions of our statistical model, we removed latencies longer than three times the standard deviation (5.3% of the responses). We report model results based on these original response latencies below; log-transformed latencies made no substantial further improvements to the normality of the model's resulting residuals and the original values can be interpreted more directly. See Figure 2 for a density plot of response times (without outliers) for the four conditions.

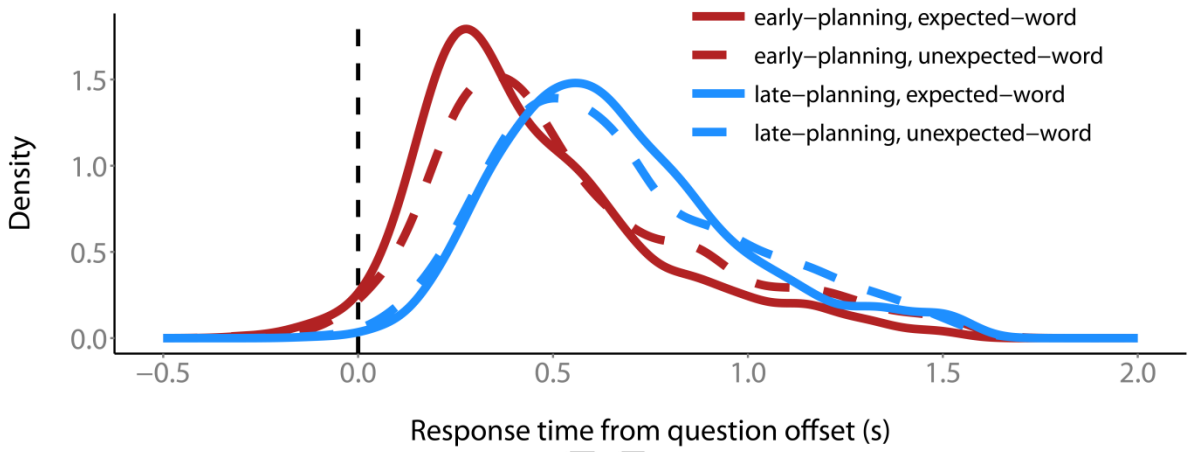


Figure 2. Density plot of response times from question offset for the four different conditions.

We built a linear mixed-effects model of response time relative to question offset, with Planning (early-planning, late-planning), Expectedness (expected-word, unexpected-word), and their interaction as the main predictors. The model also included random intercepts for participant and item, with random slopes of Planning and Expectedness for both participants and items. This model showed a main effect of Planning: responses were faster in the early-planning condition (M = 498 ms) than in the late-planning condition (M = 664 ms; t = 7.84), consistent with the behavioral results of the earlier study (Bögels et al., 2015a). This suggests that, consistent with our design and the prior findings, planning for the early-planning conditions began before the question had ended. There was also a main effect of Expectedness: responses were faster for questions with expected (M = 556) than with unexpected words (M = 606 ms; t = 4.34). Finally, the model showed an interaction between Planning and Expectedness (t = -3.28) such that there was an effect of Expectedness in the

early-planning condition (expected-word *M* = 453, red solid line; unexpected-word *M* = 545 ms, red dashed line; Figure 2), but not in the late-planning condition (expected-word *M* = 661, blue solid line; unexpected-word *M* = 667 ms, blue dashed line; Figure 2). That is, when the unexpected word came in the middle of a question (i.e., the late-planning condition) its effect may have already been resolved by the time participants gave their answer, whereas when it came at the end of a question (i.e., the early-planning condition), the unexpected word prolonged the response latencies.

3.2 Aim 1: Replication results

3.2.1 ERPs. The left part of Figure 3, Panel A, illustrates the ERPs on one representative electrode for the early-planning condition at the onset of the critical word (enabling retrieval of the answer) relative to the onset of the expected uninformative word in the late-planning condition (not enabling retrieval of the answer). The right part of Figure 3, Panel A, illustrates the ERPs for the late-planning condition at the onset of the critical word (enabling retrieval of the answer) relative to the onset of the expected uninformative word in the early-planning condition (not enabling retrieval of the answer). Figures of equivalent conditions from Bögels and colleagues' (2015a, Figure 1) study are reproduced in Figure 3, Panel B, for side-by-side comparison. In general, we see a large positive effect starting around 500 ms for the early-planning condition and a bit earlier for the late-planning condition, consistent with Bögels et al. (2015a). For the early-planning condition, this positivity is preceded by an N400 effect. Cluster-analyses indeed show a negative effect (224–476 ms, *sum-t* = -8673, *p* = .01) for the early-planning condition only. Thus, the N400 effects do not appear to be entirely consistent with Bögels et al. (2015a), with only an N400 effect in the early planning condition in the present study and only an N400 effect in the late planning condition in Bögels et al. (2015a, see present Figure 3, Panel B). We will come back to this in the Discussion. More importantly, the positive effects were highly reliable at critical word onset for both the early-planning condition, relative to the expected word occurring early (588–1500 ms, *sum-t* = 41668, $p < .001$) and the late-planning condition, relative to the expected word occurring late (260–1324 ms, *sum-t* = 127320, $p < .001$; note that 1324 ms is the end of the analyzed window for this condition given the start of articulation). Given their similarity in shape, timing, and distribution to Bögels et al. (2015a; see present Figure 3), we relate these positivities to response planning.

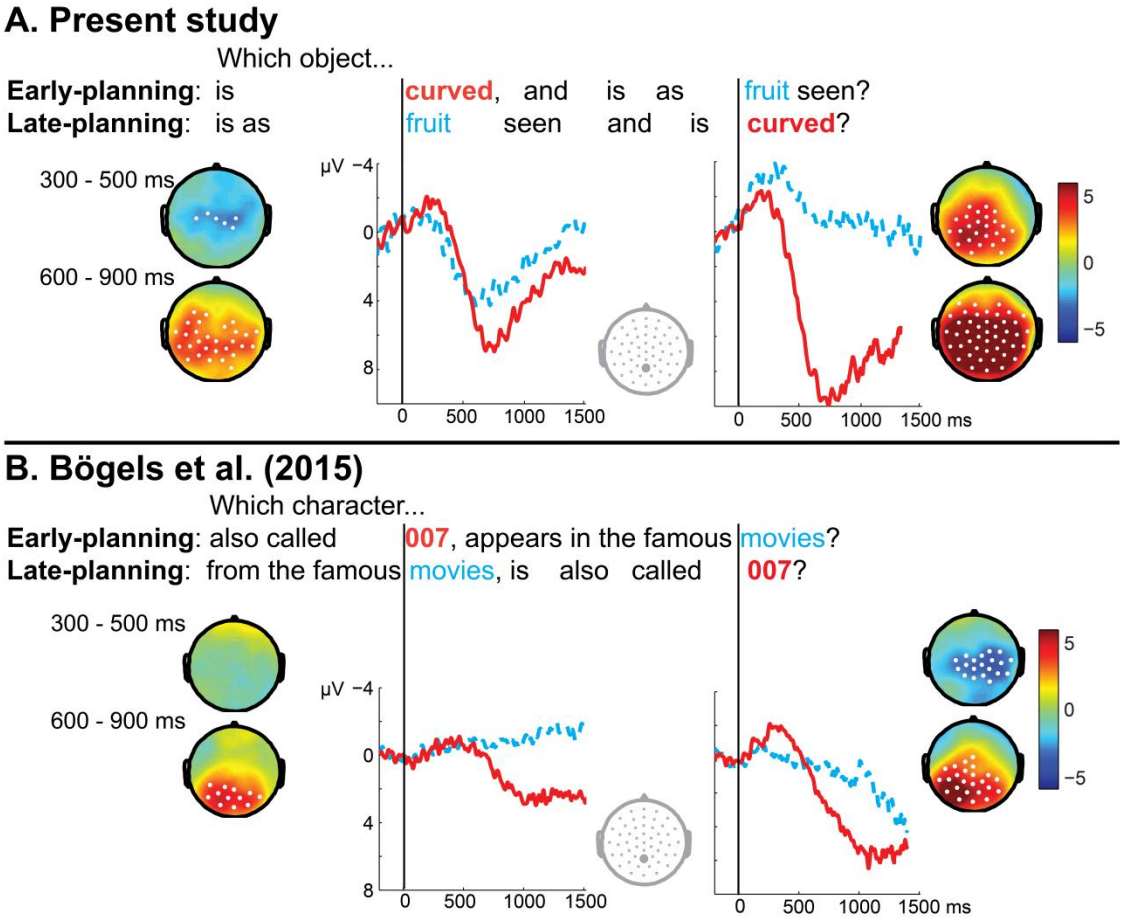


Figure 3. ERP comparisons with Bögels et al. (2015a) for early- versus late-planning conditions. Grand average ERPs for a representative electrode (Pz) from the present study are given in Panel A and results from Bögels et al. (2015a, part of Figure 1) in Panel B for comparison. The effects elicited by the early-planning condition (relative to the expected word occurring early) are given on the left and the effects elicited by the late-planning condition (relative to the expected word occurring late) are given on the right. Critical (informative) words are always indicated by red solid lines and expected words at equivalent positions by blue dashed lines. Topographical plots are given for the N400 time window (300–500 ms) and a time-window for the positivity (600–900 ms). Colors indicate t-values. Electrodes that show a significant effect in more than 70% of the time window are highlighted in white.

A localization of the positivities was performed between 600 and 1100 ms (see Bögels et al., 2015a). The localization for the positivity elicited by early-planning questions only led to a marginally significant cluster (*sum-t* = 292, $p = .086$) confined to left motor areas and extending to an area near the right temporo-parietal junction (top part of Figure 4, Panel A, figures from Bögels et al., 2015a are reproduced in Panel B for comparison). The positivity elicited by late-planning questions showed one cluster (*sum-t* = 4181, $p < .001$) localized at distributed sources in the brain (see Figure 4, Panel A, bottom). Local maxima for this cluster were found in similar language-production related areas (Indefrey & Levelt, 2004) as found

in Bögels et al. (2015a), including areas in the temporal lobe (especially the posterior part and temporal pole) and the inferior frontal gyrus. In addition, the cluster comprised motor areas not found by Bögels et al. (2015a) at the end of the question. The cluster was again most pronounced at the left hemisphere with some extensions to the right hemisphere as well, especially the temporal lobe.

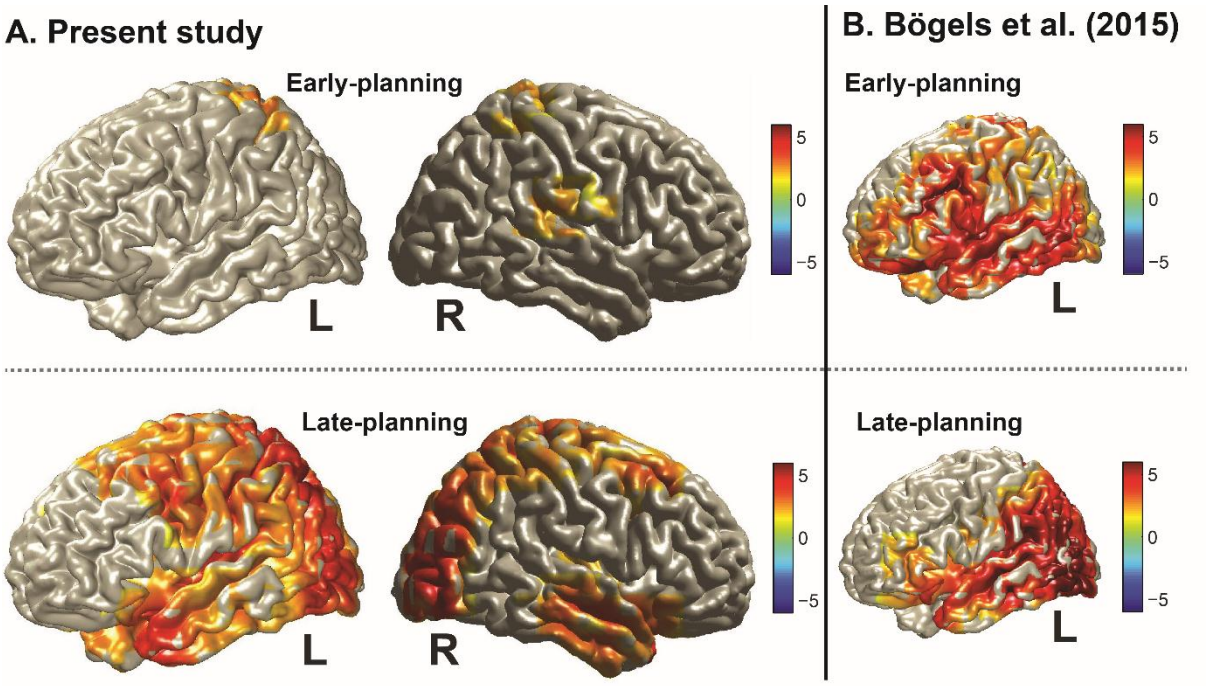


Figure 4. Localizations of ERPs; comparisons with Bögels et al. (2015a) for early- versus late-planning conditions. In Panel A, localizations of the positivities in the ERPs (600–1100 ms) elicited by the early-planning condition (marginally significant, $p = .086$) are shown at the top and by the late-planning condition are shown at the bottom for both hemispheres. Results from Bögels et al. (2015a, part of Figure 2) are reproduced in Panel B for comparison. Colors indicate *t*-values.

3.2.2 Frequencies. The left part of Figure 5, Panel A, shows time-frequency distributions for the early-planning condition at the onset of the critical word relative to the onset of the expected word in the late-planning condition. We found a modulation of power around the alpha band (around 8–14 Hz) which was similar to Bögels and colleagues' (2015a) findings (see Figure 5, Panel B, reproduced from Figure 3 of Bögels et al., 2015a). Alpha power was reduced (*sum-t* = -2724, $p = .014$) between about 600 ms until the end of the analyzed window (1250 ms after the start of the time-locking point). The second negative cluster was marginally significant (*sum-t* = -1551, $p = .085$), indicating reduced low beta power (about 15–20 Hz) between about 500 and 900 ms. The right part of Figure 5, Panel A, shows time-frequency distributions for the late-planning condition at the onset of the critical word

relative to the onset of the expected word in the early-planning condition. Here a stronger reduction in beta power (about 15–20 Hz) was found between about 500 and 800 ms (*sum-t* = -1990, p = .045) and a marginally significant reduction in alpha power (*sum-t* = -1683, p = .069), in some electrodes extended between 300 and 1200 ms. Furthermore, an increase in theta power (4–6 Hz) is found in this condition (*sum-t* = 6293, p = .002). The latter effect is probably related to the strong ERP effects in this condition (cf. Figure 3, Panel A, right graph) and we will therefore not discuss this result further.

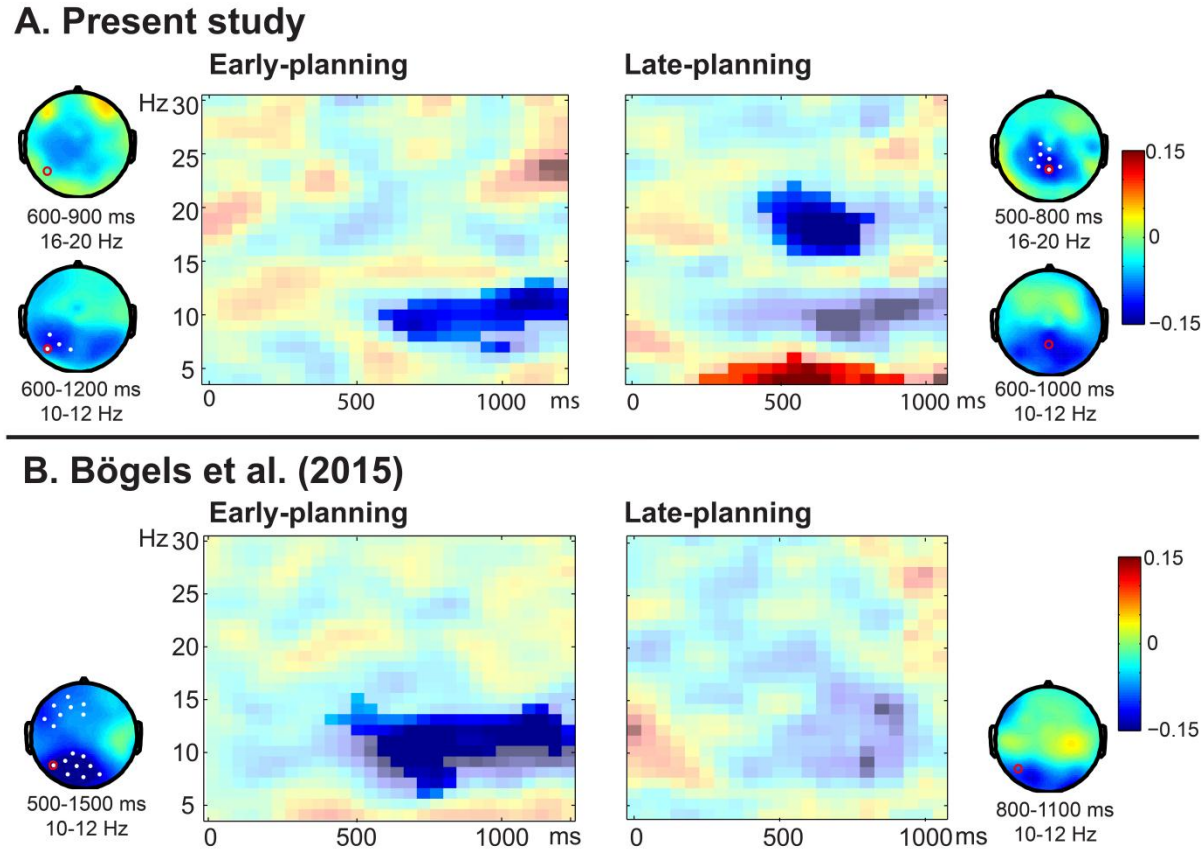


Figure 5. Time-frequency comparisons with Bögels et al. (2015a) for early- versus late-planning conditions. Time-frequency results are given for a representative electrode per comparison (see red circle in each topographical plot). Results from the present study are given in Panel A and results from Bögels et al. (2015a; part of Figure 3) are reproduced in Panel B for comparison. The effects elicited by the early-planning condition (relative to the expected word occurring early) are given on the left and the effects elicited by the late-planning condition (relative to the expected word occurring late) are given on the right. Colors in all plots indicate the relative difference between raw power in the relevant conditions. In the 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 appropriate time windows and for the 10–12 Hz range for alpha and the 16–20 Hz range for beta effects. Electrodes that are significant in more than 70% of the time window are highlighted in white.

To summarize the relevant effects, exactly as in the experiment by Bögels et al. (2015a), we found reduced alpha power in the early-planning condition and a trend towards the same effect in the late-planning condition starting within 500 ms after the critical (informative) information was presented. Additionally, in the present experiment we found a reduction in beta power starting around the same time (but lasting for a shorter period) in the late-planning condition and a trend towards the same effect in the early-planning condition. Given the similarity in alpha effects to Bögels et al. (2015a) and the similar functional significance that has been associated with alpha and beta reduction effects in the literature, we interpret both the alpha and beta effects as related to the production preparation as well. Specifically, we think these effects reflect a switch in attention from predominantly listening to the questions to more actively beginning to focus on production planning.

We performed a localization of both the alpha (8–14 Hz, 600–1200 ms) and the beta effects (16–20 Hz, 500–800 ms). Neither of these analyses yielded significant clusters at the early critical word, despite the significant alpha effect. At the late critical word, the beta analysis yielded one negative cluster in the analysis without regularization (lambda 0; *sum-t* = -615, $p = .05$). Figure 6, Panel A shows the effect to be confined to posterior brain areas, reminiscent of the localization of alpha effects in our earlier study (see Figure 6, Panel B, reproducing part of Figure 4 from Bögels et al., 2015a for comparison). The alpha analysis at this late critical word yielded only a marginally significant negative cluster (*sum-t* = -576, $p = .07$) encompassing similar brain areas (see Figure 6).

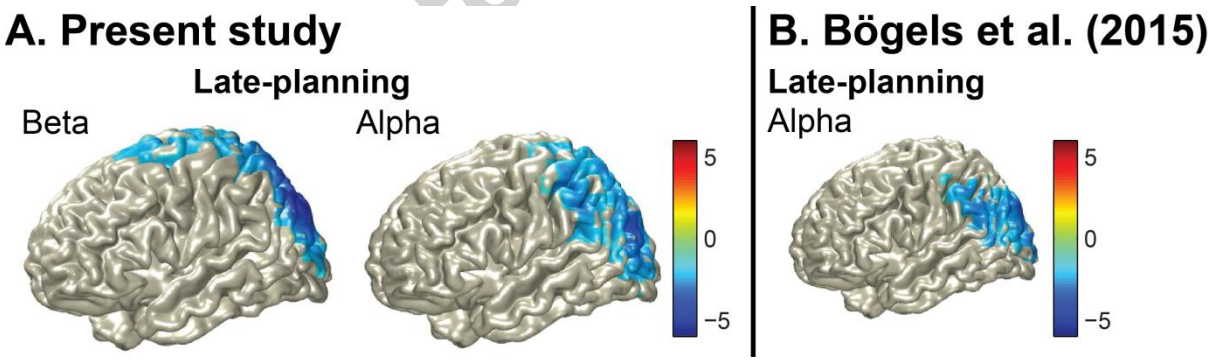


Figure 6. Localizations of time-frequency results; comparisons with Bögels et al. (2015a) for late-planning conditions. Panel A shows Localizations of reduced beta (left) and alpha (middle) power for the late critical position. Results from Bögels et al. (2015a, part of Figure 4) are reproduced in Panel B for comparison. Colors indicate *t*-values.

3.3 Aim 2: Comprehension vs. Production Results

Figure 7 shows ERPs for 15 distributed electrodes, time-locked to the onset of the expected and unexpected words coming either in the middle of the question (before planning could have started, in the late-planning condition, top graph; hereafter referred as effects *before* planning) or near the end of the questions (after planning had presumably begun, in the early-planning condition, bottom graph; hereafter referred to as effects *after* planning). Both graphs show an N400 effect, that is, a larger N400 for the unexpected relative to the expected word. The N400 effect after planning appears to start somewhat later and the distribution appears to be a bit more posterior. Analyses yielded an N400 effect before planning (180–660 ms, *sum-t* = -18006, p = .002) and after planning (300–656 ms; *sum-t* = -12468, p = .009). In addition, a positive effect for the unexpected word relative to the expected word was found before planning as well (1214–1500 ms, *sum-t* = 4608, p = .03). Such an effect was absent after planning (see Figure 7, Panel B). Since we were interested in the potential differences between the two N400 effects, we performed a 2-step interaction analysis (see section 2.7: Data Analysis for details). This analysis yielded no differences in or near the N400-window but only yielded a late significant cluster (964–1500 ms, *sum-t* = -5551, p = .023) reflecting the late positivity present before planning, but not after planning.[2]

In summary, we found N400 effects for unexpected relative to expected words both before and after planning could have started that were not reliably different from each other. Furthermore, a very late positivity for unexpected versus expected words was present before planning but was absent after planning. We must, however, be careful about interpreting this late effect because the positions before and after planning differ considerably around the laye time window. Specifically, for the position after planning (where no late positivity was found), participants might already have heard silence at that point or may have even started speaking (leading to loss of power because these trials were removed).

[2] A cluster-analysis on the differences between expected and unexpected words yielded no differences in N400 effects nor any other effects between before and after planning positions (*ps* > .23)

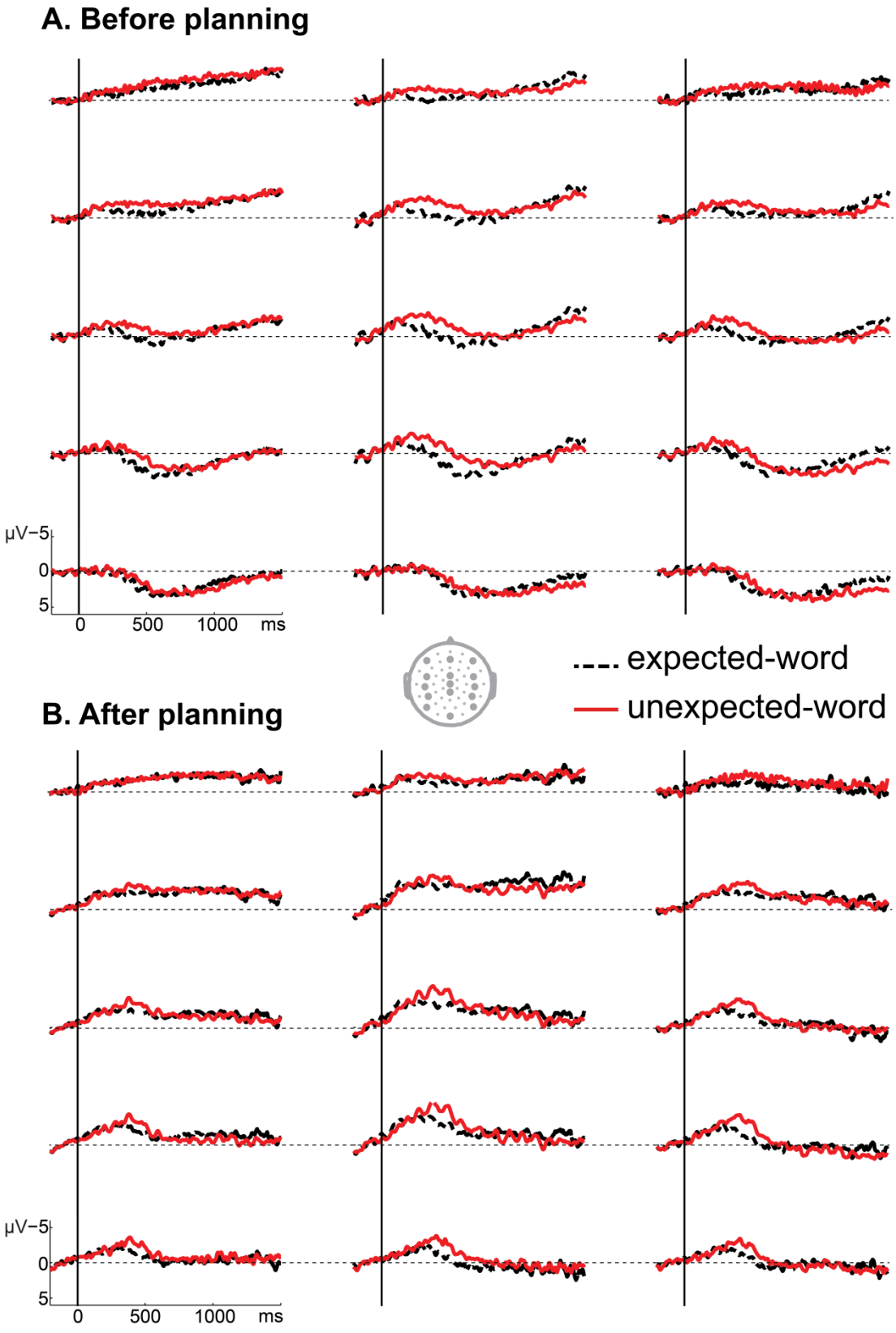


Figure 7. N400 effects to unexpected versus expected words before and after planning has started. Grand average waveforms time-locked to onset of the expected (e.g., *fruit*; black dashed line) and the unexpected word (e.g., *healthy*; red solid line) before planning (Panel A) and after planning (Panel B). A representative subset of 15 electrodes is shown, the locations of which are indicated on the head in the middle.

3.4 Individual differences analyses

Given that our hypothesis of a smaller N400 effect after than before planning was not borne out, we wanted to see whether participants might have followed different response strategies; if some prioritized quick responding more than others, the stimuli would have elicited a smaller N400 effect for fast-responding participants, but not slow responders (see also Barthel et al., 2016 for a similar idea). To look into this, we calculated the average response time (from the end of the question) over all four conditions for each participant and correlated this value with the average size of their N400 effect (in a representative electrode, Cz, see Figure 8; see Figure S1 for correlations at multiple electrodes). The Pearson correlation between average response time and the N400 effect before production planning was not significant (r = -.131, p = .475), whereas the Pearson correlation between average response time and the N400 effect after planning could have started was significant and negative (r = -.456, p = .009). Removal of three possible outliers (see Figure 8, Panel B) still led to a significant negative correlation (r = -.536, p = .003). Figure 8, Panels A and B, show scatterplots illustrating these correlations. Thus, the N400 effect before planning was not dependent on overall response time, whereas the N400 effect after planning was smaller for participants with shorter response times. This suggests a potential trade-off between a fast production planning strategy (reflected by a short response time) and a focus on comprehension processes (reflected by a large N400 effect). Figure 8, Panel D, shows the N400 at Cz after planning could have started for two groups of participants, shown here with a median split in overall reaction time between participants (for completion we display the same results before planning could have started in Panel C). In Panel D, the N400 effect (i.e., the difference between unexpected and expected words) at Cz is clearly present for 'slow responders' (dashed lines) whereas it is much smaller for 'fast responders' (solid lines). This figure suggests that the difference in N400 effect is caused predominantly by expected words. That is, unexpected words (the two red lines) yield a similar N400 in fast and slow responders. In contrast, expected words (the two black lines) diverge: slow responders show a typical reduction of the N400 for expected words, suggesting that they anticipated the word based on the context, whereas fast responders show much less N400 reduction for expected words, possibly because they did not anticipate this word as much.

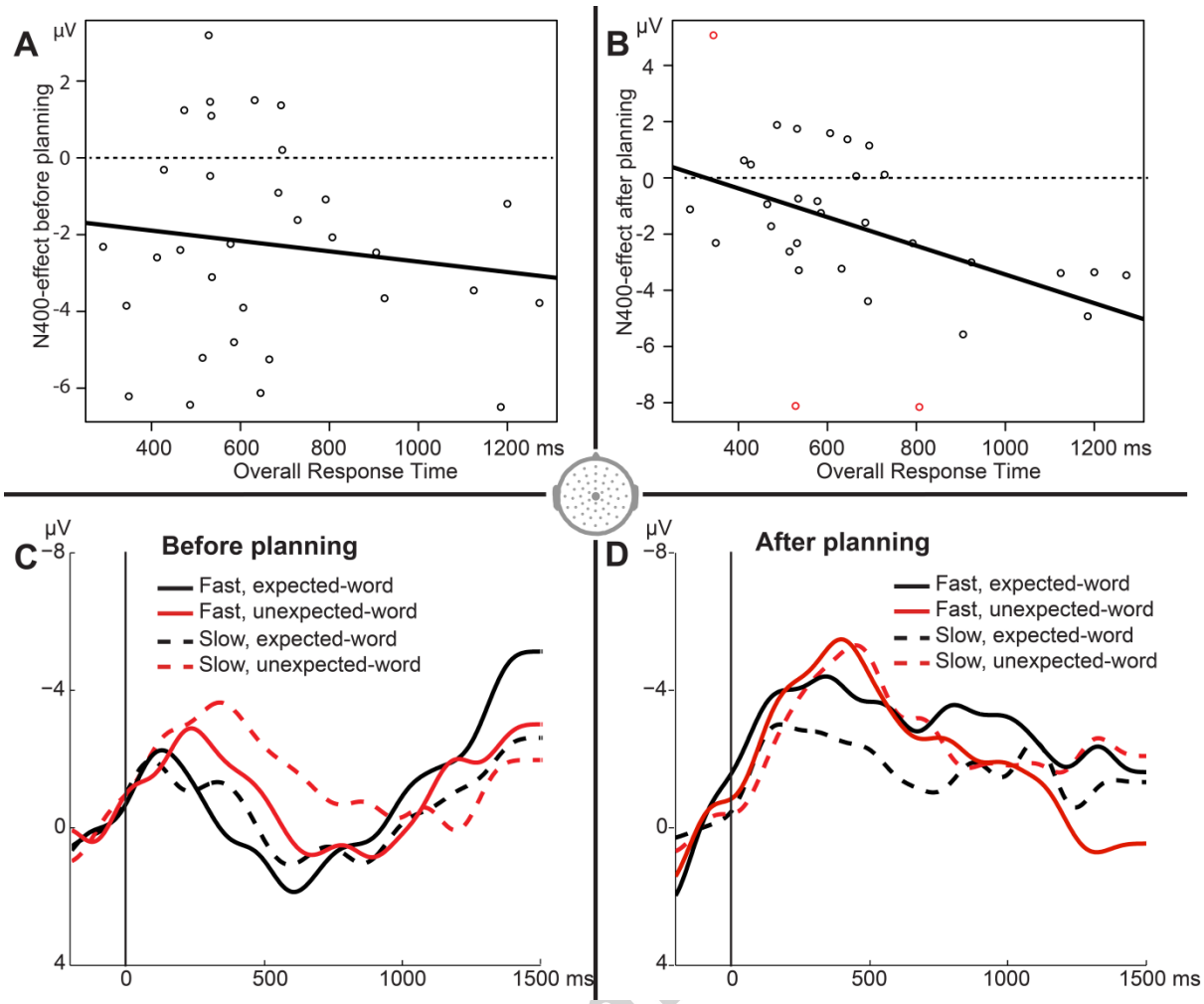


Figure 8. N400 individual differences analyses. Panels A and B show scatterplots for the correlation between participants' overall response time and the average N400 effect (between 300 and 500 ms in a representative electrode, see small head in the middle) before planning (A) and after planning (B). The data points indicated in red in Panel B reflect possible outliers, without which the negative correlation was still significant. Panels C and D shows grand average waveforms at Cz (see head at the bottom) for expected and unexpected words before (C) and after (D) planning for 'fast' and 'slow' responders separately (median split). For visualization only, the waveforms were filtered with a low-pass filter of 5 Hz.

Given our interpretation of the positive ERP effects reported above (section 3.2: Replication Analyses) as reflections of the start of production planning, we also investigated whether average response time was correlated with the average size of these positive effects (in a representative electrode just below Cz, see Figure 9; see also Figure S2 for correlations in a larger set of electrodes). We found that the positivity in response to the early-planning condition (relative to the expected word occurring early as a control) was not correlated with response time (r = -.046, p = .804). The positivity in response to the late-planning condition (relative to the expected word occurring late as a control) was negatively correlated with this measure (r = -.384, p = .030). However, this correlation was no longer significant (r = -.252,

p = .171) after removing a possible outlier (see Figure 9, Panel B). Figure 9, Panels A and B show scatterplots illustrating these correlations and Panels C and D show the ERP results for fast and slow responders based on a median split in response times (cf. Figure 3, Panel A which shows these results for the whole group in an adjacent electrode). Panel D shows a larger positivity that appears to start a bit earlier for fast than slow responders, providing some support for the interpretation that the positivity is related to production planning.

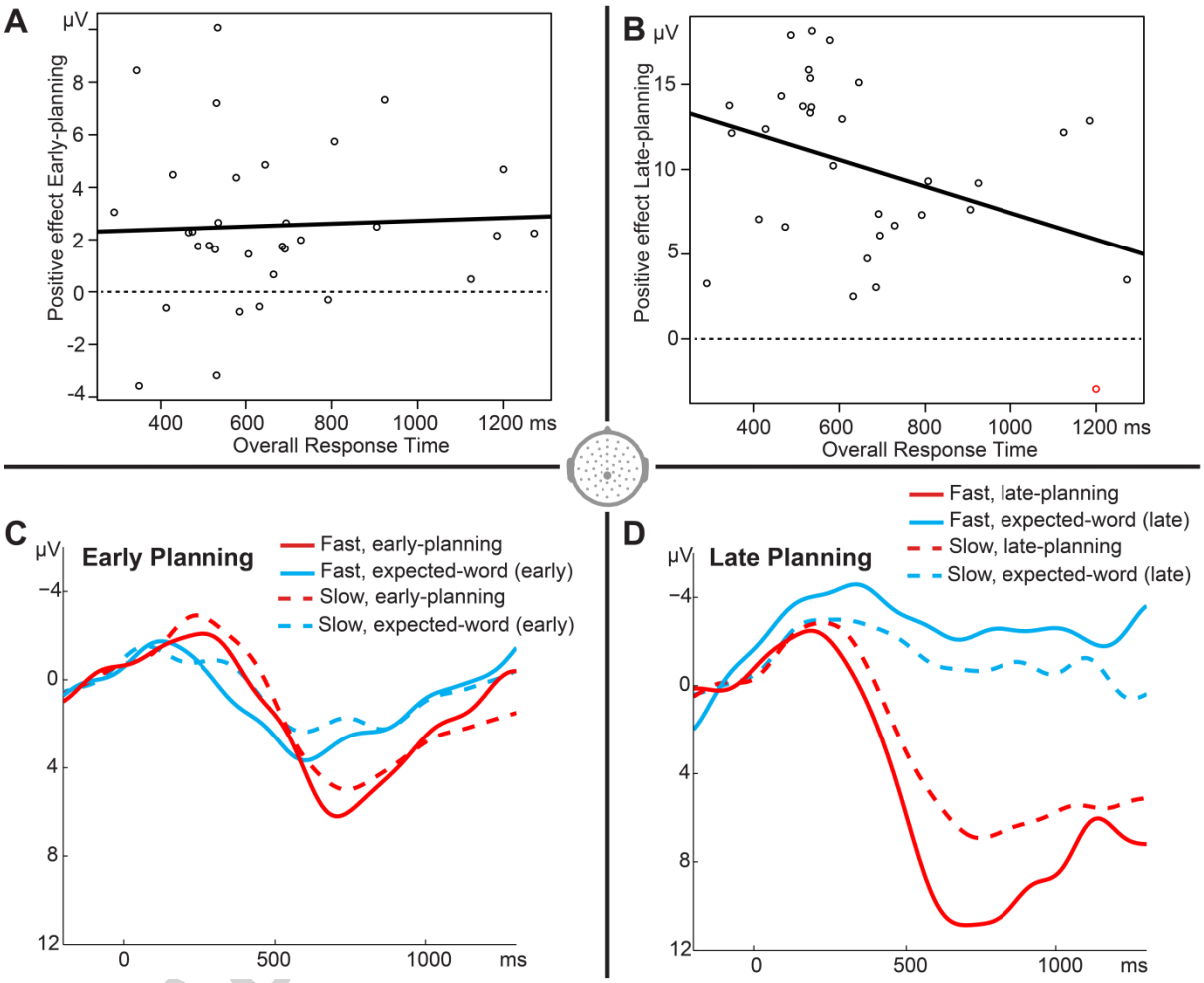


Figure 9. Positivity individual differences analyses. Panels A and B show scatterplots for the correlation between participants' overall response time and the average positivity (between 600 and 900 ms in a representative electrode, see small head in the middle) at the early position (A) and the late position (B). The data point indicated in red in Panel B reflects a possible outlier, without which the negative correlation was no longer significant. Panels C and D show grand average waveforms at a representative electrode (see head) for early (C) and late (D) critical and expected words for fast and slow responders separately. For visualization only, the waveforms were filtered with a low-pass filter of 5 Hz.

4 Discussion

4.1 Aim 1: Replication of Bögels et al. (2015a)

The first aim of the present study was to replicate the neural correlates of production planning during turn-taking found by Bögels and colleagues (2015a), but in a different experimental paradigm. As part of this, we replicated the behavioral finding that, when answers can be retrieved earlier in the question, participants' response times are faster (cf. Barthel et al., 2016; Bögels et al., 2015a; Magyari et al., 2017). As before, the present study also found that the facilitative effect of starting production planning early was not as large as it could have maximally been, suggesting less efficient production planning during simultaneous comprehension. Regarding the neural correlates of production planning during turn-taking, we replicated the central effects found by Bögels and colleagues (2015a), namely the positive effect in the ERPs and the alpha reduction in the time-frequency analysis. Given the similarity of the findings, we interpret them in the same way. That is, we take the positivity in the ERPs as a neural correlate of production planning per se, whereas the alpha reduction is interpreted as a switch in attention from comprehension to production planning. The timing of these effects then corroborates Bögels and colleagues' (2015a) conclusion that production planning appears to start as soon as it can, even if this might be in the middle of the ongoing question. That said, we also found some subtle differences in the results of the two studies, which we discuss in more detail below.

On average, participants in the present study answered a bit earlier than those in Bögels et al. (2015a), especially for late questions (about 200 ms earlier). This general difference is likely to stem from the fact that making a binary choice between two objects takes less time than answering open trivia questions, as was done in Bögels et al. (2015a). More importantly, as in Bögels et al. (2015a), participants answered faster when they could start planning earlier.

Regarding the ERP results, Figure 3 shows that the large positivity was clearly replicated with a very similar posterior scalp distribution and a similar timing, starting at or before 500 ms after the moment at which participants could start retrieving the answer. Looking at the slope and start of the positive component at the positions where answer retrieval could start (red lines), the positivity appears to start a bit earlier and to be a bit steeper in the present study than in the study by Bögels et al. (2015a). This might be due to the fact that the position of the informative word was more predictable in the present study, given the more predictable sentence structures. Moreover, in the present study the two

options for the answer were given at the start of the trial—before the question began—which might have made it easier to launch planning early (consistent with the slightly faster average response times). Interestingly, looking at the top left graph of Figure 3, the control condition also appears to show a positive component, albeit not as strongly as the critical condition. This might again be due to the predictability of the position of any potentially critical word in the questions. Participants might immediately try to retrieve the correct answer when hearing such a word, but since they are not able to do so in the control condition, this process has to stop prematurely and the positivity remains smaller.

The size of the positivity appears to be larger in the present study than in the study by Bögels and colleagues (2015a), especially at the end of the question (Figure 3, right graphs). This might be due to slight differences in the control conditions used at the end of the question (blue dashed line in Figure 3). In Bögels and colleagues' (2015a) study the control condition was the last word of the question, so it is likely that language production processes were activated, especially those related to articulation. However, in the present study the control condition at this position was always the expected uninformative word (e.g., *fruit*) followed by one or more words (typically one, included to avoid wrap-up effects). One would expect final language production processes (e.g., articulation) leading to a small positivity in the control condition to start at the very end of the question, which occurs a bit later relative to the time-locking point in the present study compared to the study by Bögels and colleagues (2015a). Indeed, Figure 3 suggests a positivity for the control condition (blue dashed line) that starts somewhat later than in the Bögels and colleagues (2015a) study.

In the present study, a small N400 effect was found preceding the large positivity when the answer was known early in the question relative to the expected uninformative word occurring early in the sentence. Such an effect is not surprising given that the informative word was unexpected by design and the expected uninformative word was expected by design, leading to a standard N400 effect. Such an N400 effect was not found by Bögels and colleagues (2015a), probably due to the fact that the control word in the middle of the question in that study was not specifically designed to be expected. Conversely, in contrast to Bögels et al. (2015a), in the present study no N400 effect was found when the answer became known at the end of the sentence relative to the expected uninformative word as the control condition (Figure 3, Panel A, right graph). This might be due to issues alluded to above; the critical condition leads to a very rapid positivity, whereas the positivity in the control condition is delayed until the end of the sentence (one or more words later). For these

reasons, the strong difference in the positivity seems to overlap with and override any potential N400 effects.

With respect to the frequency results, we clearly replicated the reduction in alpha power for the early questions, with a similar (slightly left) parietal distribution and a similar onset around 500 ms after the onset of the informative word. A similar effect was only marginally significant in the late questions, as was the case in Bögels and colleagues' (2015a) study. Additional to the reduction in alpha power, in the present study we also see a beta power reduction in the late condition, which is only marginally significant in the early condition. The distribution of this beta reduction appears a bit more central, but the timing is similar to the alpha reduction effect, starting around 500 ms after critical word onset. We can only speculate about the functional significance of this effect. One option is that it serves a similar function to the alpha reduction, given that beta reduction has also been claimed to reflect a stronger engagement of task-relevant brain areas (e.g., Wang et al., 2012). This option is corroborated by the similarity in localization of the two effects (see Figure 6 and below). Both effects could then be interpreted as a boost of the visual system, relative to the auditory system, reflecting a focus of attention on other processes than auditory processing and language comprehension (e.g., involving visual imagery of the correct answer). Such a visual process might be even more likely in the present study than in Bögels and colleagues' (2015a) since the two alternative responses were originally presented as pictures and participants might use visual imagery to bring one of them back. Alternatively, the beta reduction might be due to some specific characteristic of the present procedure. For example, in contrast to Bögels et al. (2015a) a relatively high memory load was imposed from the beginning of the question by having participants keep two pictures in memory. Such rehearsal in visual short-term memory has been related to increased beta power (Tallon-Baudry et al., 1999). At the moment the answer can be selected, the memory load decreases to only one picture, possibly leading to a decrease in beta power.

As in the study by Bögels and colleagues (2015a) we again attempted to localize the effects found in the ERPs and time-frequency analyses. When it was possible to identify a significant localization, the effects were similar to those found by Bögels et al. (2015a). That is, localizations of the positivity in the late-planning condition again comprised areas that have been related to language production such as the left temporal lobe and the left inferior frontal gyrus. In this case even motor areas were involved, in contrast to the localizations of the positivity in the late condition by Bögels and colleagues (2015a). This makes sense, given the fact alluded to above, that the expected word in the control condition appeared one or

more words before the end of the question, such that motor areas related to articulation would be more active in the critical condition than in the control condition. Localizations of the alpha and beta effects at the late questions were again similar to Bögels et al. (2015a), comprising mainly posterior and parietal areas. No significant localization effects were present at the early condition, neither for the ERPs, nor for time-frequency results. The reason for this is unclear but we must keep in mind that localization of EEG data can be difficult (see, e.g., Leahy et al., 1998), especially if a standard head model is used, as in the present study. Future research, for example using MEG combined with structural MRI, might be better able to shed light on the exact neural underpinnings of the effects found here.

In sum, the present study largely replicates the prior findings of two neural correlates of production planning during turn-taking (Bögels et al., 2015a). Most of the subtle differences in results are likely related to differences in the procedure of the two experiments. The present study then lends support to those initially exploratory findings, showing that they are robust to irrelevant changes in the set-up and circumstances of the turn-taking situation. Moreover, these neural correlates were again found very soon after participants could first start planning their responses, lending further support to the idea that response planning in turn taking starts as early as it can.

4.2 Aim 2: Comprehension vs. Production

The second aim of the present study was to see whether comprehension of the incoming turn would suffer from simultaneous production planning of the response. Our behavioral results already shed some light on this question since we found longer response latencies for questions with an unexpected word compared to an expected word, especially when it occurred at the end of the question. This suggests that listeners generally still processed these words, even if they were already planning, but it was unclear whether they processed them to the same extent as before they could have started planning. In addition, we elicited an N400 effect by including unexpected and expected uninformative words in different positions in the questions. We had hypothesized that the differential N400 effect (unexpected vs. expected words) might be smaller when these words occurred late in the question, after production planning had already started, than when they occurred early in the question, before production planning could have begun. In the overall dataset, we indeed found a differential N400 effect at both positions, but these two effects were statistically equivalent. From this finding alone, combined with the conclusion that production planning indeed starts early

when it can (see previous section) one might conclude that comprehension of the incoming question does *not* appear to suffer from simultaneous production planning.

However, participants' overall response latencies from question offset varied drastically (see Figure 2), with some participants' average response times around 300 ms and others larger than a second. Moreover, we found a positive correlation between their response latencies and the size of their N400 effect after planning could have started: 'slow responders' showed a larger N400 effect between unexpected and expected words than 'fast responders' after planning could have started. Crucially, this cannot be just a difference between fast and slow responders in their general anticipatory processing, since no such correlation was found between response latency and N400 size when the unexpected word was heard *before* planning could have started. So, one might speculatively infer that when fast responders encounter informative words early in the question, they allocate a large part of their attention to production planning, which lessens the attentional resources for comprehension of the ongoing turn, but leads to a faster response. Conversely, when slow responders encounter the informative words early in the question, they may not allocate as much of their attentional resources to production planning yet, which leaves more attentional resources available for comprehension of the ongoing question, but comes at the cost of a relatively late answer. This would explain why we see a positive correlation between average response time and size of the N400 effect after response planning could have started.

If we assume that slow responders indeed allocate more resources to comprehension, what kind of processes would be involved? Perhaps slow responders use these resources for anticipation of upcoming words. From Figure 8, Panel D, the N400 component in response to unexpected words appears virtually identical for fast and slow responder, whereas their response to expected words appears to differ. That is, slow participants show a (typical) attenuated N400 in response to expected words, suggesting that they anticipate or even predict these words on the basis of the context. On the other hand, for fast responders the N400 component in response to expected words is almost as large as to unexpected words, suggesting that fast responders are less engaged in anticipatory or predictive processing. These results thus suggest that one way in which production planning might affect comprehension is by reducing anticipatory processing for incoming speech.

How can we integrate both kinds of results reported above, then? The first part, together with the earlier results by Bögels et al. (2015a) suggests that conversationalists start planning their response as soon as they can. On the other hand, the second set of results suggests that there are individual differences in the amount of attention allocated to

production versus comprehension (or anticipatory processing). We found exploratory indications that faster responders may show a larger positivity at the end of the question. However, since this correlation appeared to depend on one outlier, we have to be very careful in interpreting this result. If it would be replicated in later studies, it would corroborate the idea that the positivity is related to production planning processes and suggest that a larger positivity may reflect more resources put into planning. However, we did not find any correlation between response time and the positivity early in the question. This is especially puzzling because fast responders showed a smaller N400 effect after planning could have started (suggesting a decline in comprehension) and one would then expect them to also show a larger positivity reflecting a stronger investment in early production processes. In general, though the positivity found in these two studies clearly appears related to production planning, its exact functional relevance remains to be uncovered. For example, the correlation at the end of the question (described above) might suggest that the timing or size of the positivity relates to the amount of resources put into production planning. But it could also correspond to the amount of material that has to be planned or the speed with which planning is started. Future research designed specifically to shed more light on these relationships is necessary before any strong conclusions can be drawn. For example, such studies could manipulate answer length, vary the amount that has to be planned, or the motivation for responding quickly (e.g., speeded vs. non-speeded responses).

Another open question is which mechanism(s) conversationalists use to divide their attention between production and comprehension. For example, these processes might be carried out in parallel with differing amounts of resources allocated to either task. Alternatively, some sort of rapid switching between comprehension and production processes might be going on. In the latter case, it is still possible that all or most people generally start planning as soon as they can, but some of them keep focusing on the production process throughout the ongoing turn, whereas others switch back to comprehension often.

If conversationalists indeed employ different strategies in allocating cognitive resources to comprehension versus production-processes during turn-taking, what factors do these strategies depend on? One potential answer comes from stable individual differences. For example, an earlier study found individual differences in language processing between males and females (Wang et al., 2011); only male participants showed a differential N400 effect when the critical words were and were not in focus. Another potential answer comes from a link between working memory and dual task performance; individuals with higher working-memory capacity have been found to name pictures faster and show less interference

from a secondary task in picture naming (Piai & Roelofs, 2013; but see Miyake et al., 2000). On the other hand, capacity itself is not the only thing at issue here; it might just be more efficient to give one task priority over the other, even if an individual is able to carry them out in parallel.

In addition to possible stable individual differences, within-individual variation in response strategy may also come into play: what are the circumstances under which conversationalists are more or less likely to try and respond quickly? Barthel and colleagues (2016) argued that early production planning might be more likely when responses are contingent on earlier turns and when another person is present. On the basis of the present results, one could speculate that interactants always have to balance the amount of effort or attention they put into early production planning, which affects both their response time and their remaining resources for comprehending incoming speech. The precise balance between these processes for any individual at any point in time might depend on a great number of factors, such as qualities of the ongoing interaction (e.g., competitive vs. friendly, equal vs. hierarchical etc.) and more stable personal characteristics of the individual (e.g., introvert vs. extravert). Investigating which of these factors affect early production planning would be a fruitful avenue for future research.

4.3 Conclusions

The present study is among the first experimental studies to look at language production and comprehension in combination, and in an ecologically valid interactive situation. The results strengthen two previous findings. First, that response planning in turn-taking starts early—soon after critical information for responding becomes available—even if this point is midway through an ongoing turn. Second, neural correlates that were previously found to be related to (early) production planning in turn-taking were largely replicated in the present study, despite some significant changes in the interactional task, suggesting a robust neural signature for response planning during live interaction. The present study also showed, for the first time, that comprehension—more specifically, anticipatory processing of the incoming turn—can be disrupted by early planning strategies, but that this allows participants to respond quickly to the question at hand. Such a trade-off between responding quickly and comprehending efficiently may be an important insight into the inner workings of conversational turn-taking. Characterizing the conditions under which participants shift their

attention in future research may be key to understanding how these psycholinguistic processes work together in the course of everyday interaction.

5 Acknowledgements

We thank Ruben van den Bosch and Annick Bosch for their assistance during the experiment. We are grateful to the members of the IFL and Dialogue projects at the Max Planck Institute for Psycholinguistics for extensive discussion of this work.

6 Funding sources

This work was supported by an ERC Advanced Grant (269484 INTERACT) to SCL and NWO Veni Innovational Research Scheme (275-89-033) to MC.

7 References

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., 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*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bögels, S., Barr, D. J., Garrod, S., & Kessler, K. (2015b). Conversational Interaction in the Scanner: Mentalizing during Language Processing as Revealed by MEG. *Cerebral Cortex, 25*(9), 3219-3234.

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.

Bögels, S., Magyari, L., & Levinson, S. C. (2015a). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports, 5*, 12881.

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.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of memory and language, 93*, 203-216.

Corps, R. E., Gambi, C., & Pickering, M. J. (2017). Coordinating Utterances During Turn-Taking: The Role of Prediction, Response Preparation, and Articulation. *Discourse Processes*, 1-11.

Gerakaki, S., Sjerps, M., & Meyer, A. (submitted). Concurrent speech planning affects memory for heard words.

Grandke, T. (1983). Interpolation algorithms for discrete Fourier transforms of weighted signals. *Instrumentation and Measurement, IEEE Transactions on, 32*(2), 350-355.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274-279.

Gross, J., Baillet, S., Barnes, G. R., Henson, R. N., Hillebrand, A., Jensen, O., et al. (2012). Good-practice for conducting and reporting MEG research. *NeuroImage*.

Gross, J., Kujala, J., Hämäläinen, M., Timmermann, L., Schnitzler, A., & Salmelin, R. (2001). Dynamic imaging of coherent sources: studying neural interactions in the human brain. *Proceedings of the National Academy of Sciences, 98*(2), 694-699.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555-568.

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*(1), 101-144.

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.

Jongman, S. R., Roelofs, A., & Meyer, A. S. (2015). Sustained attention in language production: An individual differences investigation. *The Quarterly Journal of Experimental Psychology, 68*(4), 710-730.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32-59.

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.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

Leahy, R., Mosher, J., Spencer, M., Huang, M., & Lewine, J. (1998). A study of dipole localization accuracy for MEG and EEG using a human skull phantom. *Electroencephalography and clinical neurophysiology, 107*(2), 159-173.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6*, 731.

Magyari, L., de Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in psychology, 8*, 211.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190.

Menenti, L., Gierhan, S. M., Segaert, K., & Hagoort, P. (2011). Shared language overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological Science, 22*(9), 1173-1182.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*, 156869.

Oostenveld, R., Praamstra, P., Stegeman, D., & Van Oosterom, A. (2001). Overlap of attention and movement-related activity in lateralized event-related brain potentials. *Clinical Neurophysiology, 112*(3), 477-484.

Piai, V., & Roelofs, A. (2013). Working memory capacity and dual-task interference in picture naming. *Acta psychologica, 142*(3), 332-342.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2011). Shared syntax in language production and language comprehension—an fMRI study. *Cerebral cortex*, bhr249.

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.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory, 6*(2), 174.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences, 106*(26), 10587-10592.

Tallon-Baudry, C., Kreiter, A., & Bertrand, O. (1999). Sustained and transient oscillatory responses in the gamma and beta bands in a visual short-term memory task in humans. *Visual neuroscience, 16*(03), 449-459.

Van Veen, B. D., & Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *ASSP Magazine, IEEE, 5*(2), 4-24.

Wang, L., Bastiaansen, M., Yang, Y., & Hagoort, P. (2011). The influence of information structure on the depth of semantic processing: How focus and pitch accent determine the size of the N400 effect. *Neuropsychologia, 49*(5), 813-820.

Wang, L., Jensen, O., Van den Brink, D., Weder, N., Schoffelen, J. M., Magyari, L., et al. (2012). Beta oscillations relate to the N400m during language comprehension. *Human brain mapping, 33*(12), 2898-2912.

Highlights

- Interlocutors start planning their response as early as possible in turn-taking
- Positivity and alpha suppression replicate as neural correlates of planning onset
- Fast responders show diminished anticipatory processing of the ongoing turn
- Comprehension can suffer from simultaneous response planning