Cognition 136 (2015) 304–324

Contents lists available at ScienceDirect

Cognition

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

Variation in dual-task performance reveals late initiation of speech planning in turn-taking

Matthias J. Sjerps [a,b,c,*], Antje S. Meyer [a,b]

[a] Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands
[b] Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University, Nijmegen, The Netherlands
[c] Department of Linguistics, University of California Berkeley, Berkeley, California, USA

ARTICLE INFO

Article history:
Received 21 August 2013
Revised 29 September 2014
Accepted 20 October 2014

Keywords:
Dual-task
Turn-taking
Speech production
Speech perception
Eye-tracking

ABSTRACT

The smooth transitions between turns in natural conversation suggest that speakers often begin to plan their utterances while listening to their interlocutor. The presented study investigates whether this is indeed the case and, if so, when utterance planning begins. Two hypotheses were contrasted: that speakers begin to plan their turn as soon as possible (in our experiments less than a second after the onset of the interlocutor's turn), or that they do so close to the end of the interlocutor's turn. Turn-taking was combined with a finger tapping task to measure variations in cognitive load. We assumed that the onset of speech planning in addition to listening would be accompanied by deterioration in tapping performance. Two picture description experiments were conducted. In both experiments there were three conditions: (1) Tapping and Speaking, where participants tapped a complex pattern while taking over turns from a pre-recorded speaker, (2) Tapping and Listening, where participants carried out the tapping task while overhearing two pre-recorded speakers, and (3) Speaking Only, where participants took over turns as in the Tapping and Speaking condition but without tapping. The experiments differed in the amount of tapping training the participants received at the beginning of the session. In Experiment 2, the participants' eye-movements were recorded in addition to their speech and tapping. Analyses of the participants' tapping performance and eye movements showed that they initiated the cognitively demanding aspects of speech planning only shortly before the end of the turn of the preceding speaker. We argue that this is a smart planning strategy, which may be the speakers' default in many everyday situations.

1. Introduction

A hallmark of natural conversation is turn-taking, with interlocutors alternating in adopting the roles of listener and speaker. Speakers normally manage to coordinate their contributions to a conversation in such a way that their utterances follow smoothly on from each other, rather than overlapping or being separated by long pauses (Sacks, Schegloff, & Jefferson, 1974). For instance, Stivers et al. (2009) analysed a corpus of polar (yes/no) question–answer sequences in ten languages and found that the average interval between a question and the answer was around 200 ms. Data from Dutch corpora containing a range of different utterance types has provided a similar estimate (Heldner & Edlund, 2010). Many authors have stressed that natural conversations are characterised by smooth transitions of turns (Sacks et al., 1974; Wilson & Wilson, 2005). Moreover, there is good evidence that inter-turn intervals can convey meaning; for instance, a

* Corresponding author at: Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University, Nijmegen, The Netherlands. Tel.: +31 243521207.

E-mail address: m.j.sjerps@gmail.com (M.J. Sjerps).

http://dx.doi.org/10.1016/j.cognition.2014.10.008
0010-0277/© 2014 Elsevier B.V. All rights reserved.

long pause before an answer to a request may indicate reluctance to comply (Jefferson, 1989; Roberts & Francis, 2013; Roberts, Margutti, & Takano, 2011). Thus, speakers know how to time their contributions appropriately.

However, very little is known about the way this timing is achieved. Short inter-turn intervals and occasional overlaps of turns indicate that speakers often begin to plan their utterance while still listening to the other person (De Ruiter, Mitterer, & Enfield, 2006; Levinson, 2013; Sacks et al., 1974). This is because planning a single content word (e.g., a name of a picture) may take close to a second (Indefrey & Levelt, 2004; Strijkers & Costa, 2011) and initiating a simple descriptive utterance, such as "The donkey kicked the man", may take about two seconds (Gleitman, January, Nappa, & Trueswell, 2007; Griffin & Bock, 2000).

But when do speakers begin to plan their turns? Do they typically begin to plan an utterance as soon as they have a rough idea what they might say, or do they only begin to plan when they can anticipate that the interlocutor's turn is about to end? If utterance planning and listening indeed co-occur in time, how do speakers distribute their processing resources across these tasks? And how do concurrent listening and speech planning affect each other? Addressing these and related questions is crucial for understanding how we speak and comprehend speech in everyday contexts. So far, however, most experimental psycholinguistic work has concerned monologues and little is known about the way listening and speech planning are coordinated in everyday conversations. In the present study, we developed a new paradigm to assess the coordination of speaking and listening in a simple turn-taking task.

The basic idea underlying the study was that initiating speech planning whilst listening to another person should increase the mental load for the speaker, and that this increase in mental load should lead to a performance decrement in a motor task carried out concurrently with the linguistic task. With this approach we built upon results of numerous dual-task experiments showing that performance in a cognitively demanding task deteriorates when it is carried out simultaneously with another cognitively demanding task rather than by itself (Baddeley, 1976; Becic et al., 2010; Bock, Dell, Garnsey, Kramer, & Kubose, 2007; Duncan, 1980; Kemper, Herman, & Lian, 2003; Kemper, Schmalzried, Herman, & Mohankumar, 2011; Lavie, 2005; Lavie, Hirst, de Fockert, & Viding, 2004; Meyer & Kieras, 1997; Pashler, 1984, 1994). There are various accounts of dual-task interference but most of them share the assumption that there is a limit to the overall amount of cognitive resources that can be attributed to concurrent cognitive tasks (Duncan, 1980; Kahneman, 1973; Marois & Ivanoff, 2005; Watanabe & Funahashi, 2014; Wickens, 1980). When capacity needs to be distributed across two tasks (rather than being exclusively dedicated to one task) performance in one or both tasks suffers (Somberg & Salthouse, 1982). Related accounts assume that dual-task interference arises because of limitations to central executive control or monitoring processes (Baddeley & Hitch, 1974; D'Esposito et al., 1995). In addition to these domain-general sources of interference, there may be interference in specific processing components, such as verbal working memory, visual processing, or motor planning, drawn upon by both tasks (Bergen, Medeiros-Ward, Wheeler, Drews, & Strayer, 2013; Pashler, 1994).

Most relevant to the current study are dual-task studies that have shown that speaking and listening are prone to dual-task interference. Much of this work concerned the way listening and speaking (for instance using a mobile phone) can be combined with driving and therefore has used braking, following, or lane-keeping tasks (Becic et al., 2010; Horrey & Wickens, 2006; Kubose et al., 2006; Kunar, Carter, Cohen, & Horowitz, 2008; Strayer, Drews, & Johnston, 2003; Strayer & Johnston, 2001). Other studies were carried out in the context of research on aging and combined linguistic tasks with motor tasks such as walking, finger tapping, or tracking a moving target on a computer screen (Kemper, Herman, & Nartowicz, 2005; Kemper et al., 2003). These lines of research have yielded abundant evidence for dual-task interference between speaking or listening and concurrent non-linguistic tasks. This demonstrates that non-negligible amounts of processing capacity are required for talking and listening (for corroborating evidence from studies using other paradigm see, for instance, (Caplan & Waters, 2013; Cleland, Tamminen, Quinlan, & Gaskell, 2012; Cook & Meyer, 2008; Ferreira & Pashler, 2002; Gordon, Eberhardt, & Rueckl, 1993; Mattys, Brooks, & Cooke, 2009; Papesh & Goldinger, 2012; Roelofs & Piai, 2011). Moreover, these studies have demonstrated that dual-task paradigms are suitable to measure differences in the capacity demands imposed by different linguistic tasks. A common (though not universal) finding is, for instance, that speaking interferes more with secondary task performance, and hence appear to require more capacity, than listening (Almor, 2008; Kubose et al., 2006; Kunar et al., 2008; Recarte & Nunes, 2003) but see (Kubose et al., 2006).

Recently Boiteau, Malone, Peters, and Almor (2014) used a dual-task paradigm to investigate the cognitive demands in turn-taking situations. In their study, participants' primary task was to engage in an unscripted 15-min conversation with a confederate (Experiment 1) or a friend (Experiment 2). The secondary task was a continuous visuomotor task, which consisted of tracking a moving target on a computer screen using the computer mouse. The tracking task was carried out by itself (control condition) and throughout the conversation. The authors recorded the participants' speech rate and fluency in the conversation and their performance in the tracking task, measured as the distance between the target and the cursor. Specifically, they examined the tracking performance in the tracking-only control condition and in 480-ms time windows at the beginning and at the end of utterances the participants heard or produced, and at the ends of pauses preceding or following the participants' utterance onsets. Boiteau and colleagues found that the participants' performance in the tracking task deteriorated in the conversation compared to the control condition. In addition they found that overall the participants' tracking performance was better during listening than during speaking or during the planning pauses preceding their utterances. Further analyses showed that the participants' tracking performance improved across the time window at the beginning of the interlocutor's utterances (i.e., when the participants were listening), but deteriorated across the time windows at the end of the interlocutor's utterances, in the planning pause and at the beginning of the participants' utterance.

These results indicate that, compared to listening and articulating speech, speech planning is particularly high in capacity demands. This conclusion is consistent with earlier observations concerning the distributions of pauses and speech disfluencies in spontaneous speech, which tend to precede major planning units (Garrett, 1982; Grosjean, Grosjean, & Lane, 1979; Levelt, 1993). It is also consistent with experimental evidence from a variety of paradigms demonstrating that the early processes involved in utterance planning – identifying the concepts to be spoken about and selecting the corresponding lexical items – require central processing capacity (Belke, 2008; Cook & Meyer, 2008; Crowther & Martin, 2014; Roelofs & Piai, 2011; Shao, Roelofs, & Meyer, 2012; Wagner, Jescheniak, & Schriefers, 2010).

For the present purposes, the most important finding of Boiteau and colleagues is that the participants began to plan their utterance while they were still listening to the interlocutor. This is in line with the observation mentioned above that inter-turn-intervals in conversations are often so short that speakers must have begun to plan their utterance before the end of the preceding turn. Similar findings to those obtained by Boiteau and colleagues were reported by Ford and Holmes (1978), who asked participants to talk freely about various topics while categorising tones that were played at irregular intervals. Ford and Holmes found that the participants' responses to the tones were slower when the tones were played towards the end than at the beginning of clauses. They attributed the slow reactions to the processing load arising from planning the upcoming clause. Note, however, that Ford and Holmes's study concerned monologues and does not directly speak to the question of how speakers time their utterance planning in dialogue.

In the study by Boiteau and colleagues, the participants were engaged in unscripted 15-min dialogues about various topics. As the authors point out, studying spontaneous speech is invaluable for our understanding of how interlocutors allocate their processing resources in everyday conversations. However, researchers might sometimes wish to have tighter control of the content of the interlocutors' utterances. Though inter-turn intervals in natural conversations are typically short (around 200 ms), there is considerable variation around the average interval (Heldner & Edlund, 2010). This is not surprising. It is easy to think of many variables that may affect when speakers begin to plan their utterance and how long the planning processes leading to the initiation of the utterance take. For example, speakers first need to understand the interlocutor's speech act (whether it is, for instance, a question or request), they need to think of an appropriate reply, and, depending on the perceived time pressure, they might want to plan it fully or partially before beginning to speak. Many authors have stressed that speech planning is flexible (Swets, Jacovina, & Gerrig, 2013; van de Velde, Meyer, & Konopka, 2014), and this undoubtedly holds not only for the speakers' planning units and their choice of words and sentence structures, but also for the timing of their utterance planning.

Yet, the fact that speakers have some flexibility in their utterance planning does not mean that utterance planning is entirely unconstrained. A challenging task for psycholinguistic research of dialogue is to identify the variables that limit the speakers' allocation of processing resources and determine the possible planning strategies. Much of this research can be done by analysing large corpora of spontaneous utterances and extract the relevant variables in the way pioneered by Boiteau and colleagues. An alternative approach is to constrain the participants' utterances in certain ways and to observe which planning strategies they use. These approaches are complimentary to each other and should yield converging evidence about the way interlocutors manage dialogue. Here we describe a dual-task paradigm where the content of the interlocutors' utterances and, importantly, the presentation of the information that the speakers needed to prepare their utterance was tightly controlled by using a picture description task. As explained in more detail below, we used this paradigm to test two hypotheses about the onset of speech planning in this task.

To measure dual-task performance, we used a complex finger-tapping task (Kemper et al., 2003, see also Fraser, Li, & Penhune, 2010; Seth-Smith, Ashton, & McFarland, 1989). We used finger tapping because the use of pictorial materials in the linguistic tasks precluded the use of a visual tracking task as a secondary task. Tapping performance has been shown to decline under dual-task demands (Fraser et al., 2010; Hiscock, Cheesman, Inch, Chipuer, & Graff, 1989), indicating that tapping requires the allocation of cognitive resources. In addition, influences of a verbal task on tapping have been reported as an increase in the deviation from a fixed tapping rate when a story was retold aloud but also during passive listening (Seth-Smith et al., 1989). Continuous complex tapping does not require attention to exogenous stimuli (Theeuwes, 1991), but can be generated endogenously once the participants have learned the tapping pattern.

We used the following procedure: On each trial of the experiment, participants saw a display featuring two rows of pictures, with each row containing two pairs of pictures separated by an arrow (see Fig. 1). The arrow between the two objects of a pair indicated whether the array should be described as "put the X above the Y" or "put the X below the Y". At trial onset, one row, selected at random, was described by a pre-recorded speaker (speaker 1). Because all pictures on the screen were different, participants knew

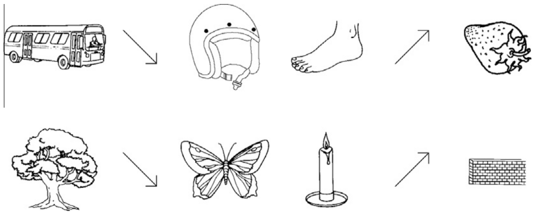

Fig. 1. Example display.

as soon as they had comprehended the first picture name which row was described first and which row would be left to be described second. There were three conditions differing in the participants' task: In the Tapping and Speaking task, participants listened to speaker 1 and then described the remaining pairs in the same fashion, saying for instance, *Put the tree under the butterfly and put the candle above the wall.* Four seconds after the offset of the utterance of speaker 1 a stop signal (a tone and a red dot) was presented and the pictures disappeared. Participants were instructed to complete their utterance before the stop signal. Thus, they were encouraged to respond relatively fast, as one might do in normal dialogue. Subsequently, they had to indicate whether they thought their utterance had been correct and whether they had completed it before the tone. They did this by reading aloud the word "goed" (correct) or "fout" (wrong), which appeared on the screen after the stop signal. Throughout the entire experiment, the participants continuously tapped a complex pattern (index finger, ring finger, middle finger, little finger). We measured their speech onset latencies, the durations of the spoken words, and their performance in the tapping task (the number of correct taps per second).

The second task, Tapping and Listening, involved a similar sequence of events. Again, participants first heard an utterance referring to the objects in the top or bottom row of the display, chosen at random. However, instead of describing the remaining pictures themselves, they then listened to a second pre-recorded speaker (speaker 2) doing so. Subsequently, they indicated by reading aloud the word "goed" or "fout" whether the second speaker's utterance was correct and completed before the stop signal. In other words, the participants overheard a pair of turns and evaluated the second speaker. Finally, in the Speaking Only task, participants performed the same tasks as in the Tapping and Speaking task but without tapping. The judgement task at the end of the trial was introduced to make sure that the participants in the Tapping and Listening task paid attention to the second speaker. In order to render the trial structure in all tasks as similar as possible, the self-monitoring task was included in the speaking conditions.

We expected that the participants' tapping performance might decline as soon as the utterance of speaker 1 began. More importantly, in the Tapping and Speaking task, we expected a further decline in tapping performance relative to the Tapping and Listening task as soon as the participants began to plan their utterance. Concerning the timing of this decline we contrasted two hypotheses: That speakers begin to plan their utterance as soon as they have sufficient information to do so (the Early Hypothesis), or that they wait until the end of the interlocutor's turn (the Late Hypothesis). Both planning strategies could easily be applied in our task: As soon as speaker 1 had named the first object, the participants knew which row of objects they should describe themselves and could initiate their own speech planning (as per the Early Hypothesis); however, since there was only moderate time pressure to respond, they could also wait until speaker 1 had named the last object and then begin to plan their utterance (as per the Late Hypothesis). Based on the earlier evidence concerning short inter-turn intervals one might expect that speakers would initiate their speech planning early. By contrast, the results obtained by Boiteau and colleagues suggest that speech planning might begin late, around the offset of the preceding utterance.

In addition to the tapping performance, we compared the participants' speech onset latencies and noun durations in the Tapping and Speaking task and in the Speaking Only task. This was done to assess to what extent the tapping task altered the way the participants planned and produced their utterances. One might expect some interference between tapping and speaking to arise at the level of motor execution (Bodwell, Mahurin, Waddle, Price, & Cramer, 2003). This would complicate the interpretation of any dual-task interference effects. Mutual interference at a level of motor execution would be expected to influence articulation as well as tapping. By measuring the noun durations we could assess the extent of motor interference in this task.

In sum, the current project investigated dual-task interference in speaking and listening using a continuous secondary task, complex tapping. Performance in the tapping task at different moments in time should indicate when participants engage in the central-capacity demanding components of speech planning. We created a situation that allowed participants to plan their utterances early. If they start to prepare their turn as soon as they know what to say, interference should begin to arise early during their interlocutor's turn, around the offset of the first noun of the first speaker, following the Early Hypothesis. Alternatively, if they wait until the end of their interlocutor's speech, interference should only arise around the offset of their interlocutor's speech, following the Late Hypothesis. In addition, follow-up analyses, involving a moving analysis window, allowed us to get a more precise estimate of the onset of speech planning.

Two experiments were conducted using the paradigm described here. They differed in the amount of tapping training the participants received before the experiment. In Experiment 1, the participants' speech and tapping performance were recorded. In Experiment 2, their eye movements were registered as well. As will be explained below, this allowed us to track the participants' allocation of visual attention and provided additional information about the initiation of speech planning and about the usefulness of complex tapping as a continuous measure of mental load.

2. Experiment 1

2.1. Method

2.1.1. Participants

Thirty-eight participants (six men) from the Max Planck Institute participant pool were tested. They were university students with a mean age of 22 years. The data obtained from four participants were discarded because they either failed to tap during the initial 3 s of a trial, or failed to name the pictures correctly on more than 75% of the speaking trials. All participants indicated that they were right-handed. They received a small financial reward for their participation. The instructions stressed that the participants should feel free to ask any questions they might have and that they could leave the experiment at any time. Ethical approval for the study had been given by the Ethics Board of the Faculty of Social Sciences of the Radboud University Nijmegen.

2.1.2. Apparatus

The experiment was controlled by a desktop computer with Presentation software (Neurobehavioral systems). Finger tapping was recorded by means of a purpose-built four-button box attached to the computer. The buttons consisted of microphones that required very little pressure to record a response. The participants' speech was recorded using a Sennheiser ME 64 microphone capsule. Auditory stimuli were presented using Sennheiser HD 280-13 headphones. Praat software (Boersma & Weenink, 2009) was used to generate the auditory stimuli and measure the duration of the participants' utterances. Stimuli were recorded at a sample rate of 44,100 Hz.

2.1.3. Materials

2.1.3.1. Visual stimuli. Forty pictures were selected from the database of 590 single-object line-drawings generated by Severens, Van Lommel, Ratinckx, and Hartsuiker (2005). The data base provides norms for these pictures in Belgian Dutch. Pictures were selected according to the following criteria applied in the order listed here: The picture name was monosyllabic or disyllabic; the average log picture naming latency was within 1.5 Standard deviations (SD) of the mean latency for the pictures in the data base; the frequency of the picture name was within 1.5 SD of the mean log frequency; the picture name was monomorphemic.

Using these pictures, five practice items and 90 experimental items were created. Each item featured eight semantically unrelated objects, randomly arranged in four pairs on two rows (see Fig. 1 for an example and Appendices A and B for a listing of all items, along with average length and frequency of the items per position as items were not counterbalanced on position). Each picture appeared 18 times on experimental trials (six times per condition) and three times on practice trials (once per condition).

In each row, arrows pointing up or down appeared between the first and second object and between the third and fourth object. Each display featured two arrows of each type, assigned randomly to the four positions. The line drawings were sized to have a maximal length and width of 7 cm, corresponding to a visual angle of approximately 6.5 deg for the participant. The length of the arrow was 4.5 cm, corresponding to approximately 4 deg.

2.1.3.2. Auditory stimuli. The auditory stimuli were created in two steps. First, a female speaker of Dutch (speaker 1) described both rows of each of the 95 displays, saying, for instance, *Zet de boom onder de vlinder en zet de kaars boven de muur* (*Put the tree under the butterfly and put the candle above the wall*). First, she described all items in the top rows of the displays and then all items in the bottom rows. Across all 95 picture arrays, a random selection of 48 top and 47 bottom descriptions was used as speaker 1 utterances for the experiment. The descriptions lasted between 2.67 s and 3.72 s, with an average of 3.02 s.

To create the recordings of speaker 2 (to be used in the Tapping and Listening task), a male speaker of Dutch was recorded while he was performing the Tapping and Speaking Task in the same way as the participants in the experiments. He first practiced the tapping task and was familiarised with the pictures. Then he saw each of the displays, heard the description of speaker 1 and described the remaining object pairs. Four seconds after speaker 1 offset a tone was presented and a dot appeared on the screen. The speaker aimed to complete his description before this audio-visual stop signal appeared. Subsequently, he indicated whether or not his description was correct by saying either "goed" (correct) or "fout" (incorrect). He carried out the tapping task during the entire recording session. Most of the 90 experimental utterances were correct and fluent. On five trials the speaker used a wrong noun to describe a picture (e.g., "vlinder" (butterfly) instead of "fles" (bottle)) and on 15 trials he completed the utterance too late (initiating the fourth object name later than 150 ms before the onset of the stop signal). This amounts to an error rate of 22%. The utterances of speaker 2 were spliced out from onset to offset. Their average duration was 3.7 s (with a minimum of 2.7 s and a maximum of 5.5 s).

In the Tapping and Speaking task and in the Speaking Only task the participants only heard the utterances by speaker 1. In the Tapping and Listening task, they heard speaker 1 followed after 250 ms by speaker 2. This inter-turn interval was our estimate of a likely inter-turn interval in this task. It was slightly longer than the 200 ms obtained by Stivers et al. (2009) because the utterances produced by the participants were more complex than simple yes/no answers. Speaker 2 sentences were truncated at 4 s after speaker 1 offset, that is, the stop signal was played instead of the end of the sentence. 45% of the utterances of speaker 2 were truncated, but typically only by one phoneme.

2.1.4. Design

There were three experimental conditions tested within participants: Tapping and Speaking, Tapping and Listening, and Speaking Only. Each task was tested in a separate block. The order of the blocks was counterbalanced across participants. The 90 experimental items were randomly assigned to three sets of 30 items each, and the assignment of sets to conditions was counterbalanced across participants. The experimental items within blocks were presented in a random order. Each block began with five practice items. In total, eighteen experimental lists were created (six presentation orders of conditions crossed with three assignments of conditions to sets of pictures). 16 lists were seen by two participants and two lists were only seen by one participant each.

2.1.5. Procedure

In the experimental session, the participants were first asked to study a booklet showing the line drawings that were used in the experiment along with the object names. They were instructed to use these names in the description task.

Then they performed a tap-training to familiarise themselves with the button box and the tapping sequence. Declines in tapping performance under load have been reported to be greater for the dominant than the non-dominant hand (Simon & Sussman, 1987). Furthermore, effects of concurrent talking have been shown to be stronger for complex finger tapping sequences than for simple tapping (Kemper et al., 2003). Therefore, we asked the right-handed participants to tap a complex pattern of (1 (index finger), 3 (ring finger), 2 (middle finger), 4 (little finger)) with their dominant hand. The training was terminated as soon as the participant had performed 50 consecutive correct taps. Training times varied across participants between 2 and 10 min.

Next, participants were asked to complete an extended version of the Edinburgh Handedness Inventory (Oldfield, 1971) to confirm their right-handedness. Then they received another booklet showing the 40 pictures, now without their names, and were asked to name them. Naming errors were immediately corrected by the experimenter.

Then the participants were instructed for the first block of the main experiment. The instructions depended on the task. Additional instructions were given prior to each block. For the Tapping and Listening task participants were instructed to listen to the descriptions by speaker 1 and 2 and then to indicate whether speaker 2 had described the displays correctly and had completed the utterance on time (before the stop signal). They did so by saying aloud “goed” (correct) or “fout” (incorrect). The participants were encouraged to tap the sequence they had practiced as fast and accurately as possible throughout the test block, but they were told that the linguistic task was more important than the tapping task.

In the Tapping and Speaking task, the participants were instructed to listen to speaker 1, and then to describe the two remaining object pairs. They were asked to try to complete the utterance before the stop signal was presented. They should then indicate whether they thought they had described the display correctly and completed the description on time by saying “goed” or “fout”. Again, they were asked to tap as fast and accurately as possible throughout the test block, but to give priority to the linguistic task. Finally, in the Speaking Only task participants were instructed to listen to speaker 1, describe the remaining object pairs and again judge the correctness and timing of their own utterance.

Fig. 2 shows the trial structure. In all tasks, each trial began with a blank interval of two seconds, followed by a one-second preview of the eight-object display. Then, the pre-recorded description of speaker 1 (with an average duration of 3.02 s) was played while the display remained in view. In the Tapping and Listening task, the pre-recorded utterance of speaker 2 followed 250 ms after the offset of speaker 1. Four seconds after the offset of the utterance of speaker 1 a tone (with a duration of 1 s and a frequency of 400 Hz) was played and the pictures disappeared and were replaced by a red dot in the centre of the screen. After the offset of the sound and dot, the words “goed” (correct) and “fout” (incorrect) appeared next to each other on the screen. The participant read aloud one of them to indicate whether or not the utterance by speaker 2 was correct. The words disappeared after 2.5 s, and the next trial began.

The trial structure in the Tapping and Speaking and Speaking Only tasks was the same, except that the participant heard the description by speaker 1, then described the remaining pictures themselves, and finally evaluated the correctness of their own utterance.

As explained in the Introduction, the stop signal was introduced to make sure that the participants in the Tapping and Speaking and in the Speaking Only tasks tried to respond promptly. The judgement task was included to make sure that participants paid attention to the spoken materials in the Tapping and Listening task. To keep the procedure as similar as possible across conditions, the self-monitoring task was included in the speaking conditions.

2.1.6. Analysis approach for tapping performance

For the tapping analyses, all data were used, including those with naming errors. This ensured that an equal number of trials contributed to the three tasks. Because of the high sensitivity of the microphone-buttons, on some occasions button presses as well as button releases activated the button-response trigger. Therefore, all instances where the same button-trigger was activated twice within 400 ms were discarded (26% of the taps). Based on the remaining data, a button-press was labelled as correct if it followed the correct predecessor. For instance, as the pattern was 1-3-2-4, the predecessor for button 3 had to be 1, and for button 1 it had to be 4. The first tap in a block was always coded as correct. 11% of the valid taps were discarded because they were incorrect. Tapping rate was calculated as the number of correct taps placed per second in a particular time window. Thus, tapping rate was an aggregated score, combining tapping speed and correctness.

Tapping rates were modelled spanning from 5 s before the end of the turn of speaker 1 until 4 s after the turn of speaker 1 (i.e., the end of the speaker 2 response window). Analyses were carried out by comparing the quality of fit

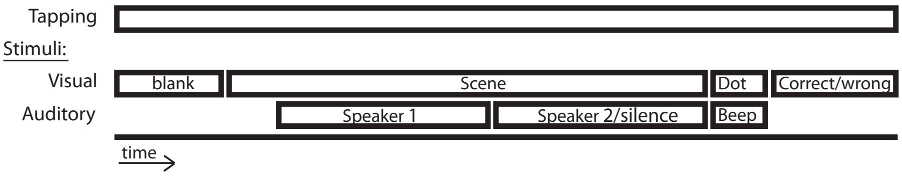

Fig. 2. Schematic representation of the sequence of events in a trial.

between models built with the lmer() function in the lme4 package (Bates, Maechler, & Bolker, 2013) in R (R_Core_Team, 2013). To assess whether planning was initiated early or late in the Tapping and Speaking task four predictor variables were constructed. These variables either had the value 1 or 0, depending on the assumed time course of interference between speech planning and tapping (see bottom panel of Fig. 3).

The main hypotheses were the Early versus Late Planning hypotheses. For the Early Planning predictor the value was initially 0 and set to 1 from the offset of the first noun of speaker 1 (i.e., this variable was tailored to the speech times of each individual trial). For the Late Planning predictor the value was initially also 0 but was only set to 1 starting from the offset of the fourth noun of speaker 1. We did not know whether interference would decrease again after the initial planning phase. Conceivably, not only planning object names, but also maintaining an utterance plan in working memory and monitoring the correctness of the utterance may be capacity-demanding. Therefore, two predictors each were created for the Early and the Late Hypotheses. For the Temporary predictors, we assumed a planning time of 600 ms per word, following averages reported in the literature (Indefrey & Levelt, 2004; Strijkers & Costa, 2011). This resulted in a 2.4 s (4 * 600) window where the predictor value was 1 instead of 0. After this period the predictor was set back to 0. The Continuous predictors had the same onset rise times as their Temporary counterparts but their value did not return to 0 for the rest of the analysis window. A final model was included that did not include any Planning predictor (i.e., a base model) to assess whether the inclusion of the Planning models was warranted at all.

Linear mixed-effects regression (Lmer) models of the tapping data with the four different predictors were created and compared to each other based on AICc (Aikaike Information Criterion corrected) measures of model fit. We report the fitted values for all models that received Delta AICc values below 2. All models were fit with the Maximum Likelihood method for fitting (i.e., with REML set to FALSE in lmer modelling). The models were subsequently compared by assessing their AICc values with the aictab() function from the AICcmodavg package (Mazerolle, 2013) implemented in R. As these models are nonnested, a comparison of χ^2 values (using, for example, the anova() function) is not valid. This approach applies to all analyses reported below.

2.2. Results

2.2.1. Tapping data

The top panel of Fig. 3 displays the tapping rates, aggregated into 19 500-ms-bins, aligned to the offset of speaker 1 (i.e., focussing on the turn-taking point, indicated by the vertical line in the figure). To exemplify, the bin at time 1 s encompasses observations between 0.75 and 1.25 s after the offset of speaker 1. The figure provides an indication of the development of dual-task interference on tapping rates across the trial. Each time point represents the average tapping rate in the Tapping and Speaking (filled squares) or the Tapping and Listening (open diamonds) task. At the top part, the vertical lines are indications of the maximal and minimal durations of the speaker 1 part (which varied in duration due to differences in word lengths). It can be observed that tapping rates were quite similar until shortly before speaker 1 offset, when a sudden decrease in tapping rates was observed in the Tapping and Speaking task. Tapping rates increased again around 2 s into the speaker 2 phase, which was, on average, around the time of the initiation of the third noun. During the judgement phase, tapping rates in the Tapping and Speaking task recovered, and were in fact better than in the Tapping and Listening task.[1]

A first analysis included the factors Task (with the levels Tapping and Speaking and Tapping and Listening, with the latter modelled on the intercept); Block (with the values −1, 0, and 1, for the 1st, 2nd and 3rd block, respectively); and the Planning predictors, varying between 0 and 1 depending on the assumed alignment and duration of interference. An average value of the predictor was calculated for each 0.5 s window as for the tapping rates. The models included the main fixed effect for Block, the main fixed effects for Task, and the interaction term between Task and the Planning predictors. In addition, the models contained random effects for those interactions for subjects and items thereby assuming a maximal random structure for the critical interaction test (Barr, 2013). This approach takes into account that participants may differ in baseline tapping rates and in the strength of the effect of speech planning on their tapping performance.

The top part of Table 1 presents the results. It can be observed that the model assuming continuous interference from the offset of the fourth (last) noun of speaker 1 fitted the data best. The optimal model revealed a significant effect on the Intercept ($B = 2.891$, $SE = 0.094$, $p < 0.001$), reflecting an average tapping rate of about 2.9 taps per second for the Tapping and Listening task during the speaker 1 part. Non-significant values were observed for the main effect of Task ($B = -0.028$, $SE = 0.071$, $p = 0.700$), suggesting that before the offset of noun 4 of speaker 1 there was no significant difference in tapping rate between the two tasks. A significant main effect was observed for Block ($B = 0.196$, $SE = 0.045$, $p < 0.001$) as participants' overall tapping rates increased as the experiment progressed. A significant effect was observed for the Planning predictor ($B = -0.151$, $SE = 0.053$, $p = 0.004$), indicating that a decrease in tapping performance occurred for the Tapping and Listening task after the offset of the first speaker. This may be related to the fact that the participants were asked to monitor the correctness of the utterances of speaker 2, but not of speaker 1. More importantly, a significant interaction was observed between Task and the Planning predictor ($B = -0.399$, $SE = 0.077$, $p < 0.001$). This shows that in the Tapping and Speaking task there was a much larger

[1] Readers may be concerned that our decision to display the tapping scores aligned to Speaker 1 offset does not allow for a fair visual inspection of the Early Planning Hypothesis. Note, however, that alignment to noun 4 offset could not obscure substantive interference aligned to the offset of noun 1 because up to noun 4 the lines representing the tapping scores in the two conditions are virtually on top of each other. In the bottom panel of Fig. 3, where predictors are also aligned to speaker 1 offset, the hypothesised increase in cognitive demands is indeed clearly distinguishable.

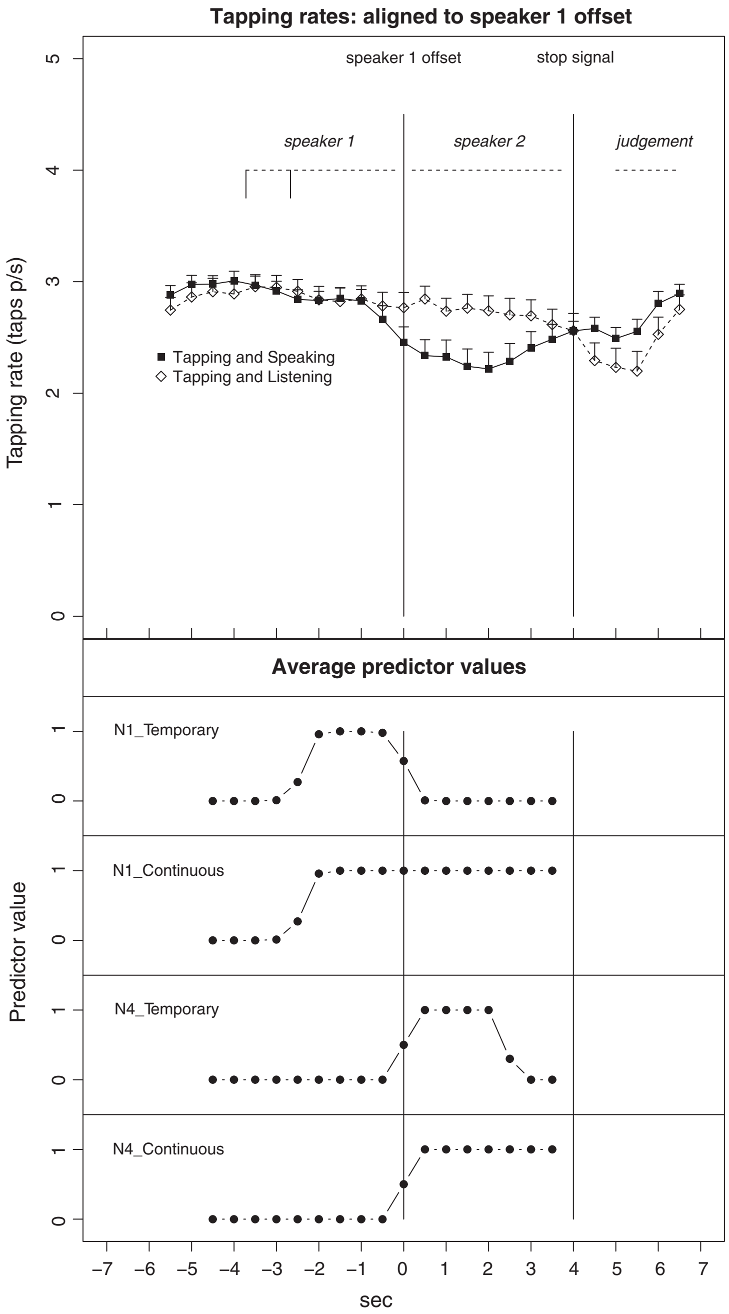

Fig. 3. Top panel: Average tapping rates across trials in Experiment 1, aligned to the offset of speaker 1. Error bars reflect the standard error of the by-participants mean. Bottom panel: Average predictor values for the analyses. Averages are displayed because the predictors were aligned to noun offsets and were therefore trial-specific.

decrease in tapping rate around the offset of the first speaker's turn than in the Tapping and Listening task. These results are in line with the Late Planning Hypothesis. Participants seem to have initiated their planning close to the end of the first speaker's turn. There was no evidence for a substantial decrease of interference after the initial planning phase.

In a second analysis we investigated the timing of interference around the time of the speaker switch in more detail. In this analysis we only included data points that were within 2 s of the end of the utterance of speaker 1. Four continuous models were created that were aligned differently to the offset of the first speaker's turn. The first model assumed a planning delay of 0.5 s, the second assumed no delay, and the third and fourth model assumed that planning preceded the offset of speaker 1 by 0.5 s and by 1 s, respectively. The bottom part of Table 1 reports the results. The optimal model assumed an increase in

Table 1
Model selection values for models that differ with respect to the temporal shape of the planning predictor. "Base" refers to the model without the planning predictor.

Predictor	*K*	*W*	Delta_AICc
Analysis 1: Planning across the trial			
N4_Continuous	26	1	0 (AICc = 125282.7)
N1_Continuous	26	0	238.48
N4_Temporary	26	0	414.92
N1_Temporary	26	0	971.28
Base	10	0	1042.87
Analysis 2: Alignment of planning to speaker 1 offset			
−0.5	26	1	0 (AICc = 65961.77)
0	26	0	15.89
−1.0	26	0	78.83
+0.5	26	0	88.52

Note. K refers to the number of free parameters. *W* reflects the relative weight of evidence for the model among the compared models. Delta_AICc reflects the information loss relative to the best model (which has a value of 0). Models are ordered based on Delta_AICc. Models having a difference in AICc < 2 compared to the model with the smallest AICc value have substantial support (evidence), those with 4 < *i* < 7 have considerably less support, and models having *i* > 10 have essentially no support, see Burnham and Anderson (2004).

interference from 0.5 s before the offset of the first speaker's turn. None of the other models received substantial support compared to this model. The optimal model had a significant value for the Intercept (*B* = 2.845, SE = 0.104, *p* < 0.001). No main effect was observed for the Planning predictor (*B* = −0.063, SE = 0.039, *p* = 0.107). No main effect was observed for Task (*B* = −0.002, SE = 0.070, *p* = 0.976). A significant main effect was observed for Block (*B* = 0.191, SE = 0.044, *p* < 0.001). A significant interaction was observed between Task and the Planning predictor (*B* = −0.462, SE = 0.077, *p* < 0.001). These results closely match those of the first analysis, with the exception that the main effect of Planning predictor was no longer significant in the smaller analysis window. This suggests that the decline in performance after speaker 1 offset in the Tapping and Listening task was especially strong towards the end of the speaker 2 part, i.e. the section that was included in the first but not in the second analysis.

2.2.2. Speech production data

Utterances that contained missing or incorrect object names and utterances where the last noun was not initiated at least 150 ms before the stop signal were coded as errors. The average error rates in the Tapping and Speaking and in the Speaking Only task were 17% and 8%, respectively (The pre-recorded Speaker 2 had an error rate of 22%, i.e., similar to that in the Tapping and Speaking task).

These errors were excluded from the following analyses, as were utterances with onset latencies outside of the range of 2.5 standard deviations from the grand mean (2.5% of the data). The average speech onset latencies for the remaining utterances were 452 ms (SE = 27 ms across participants) in the Tapping and Speaking Task and 372 ms (SE = 25 ms across participants) in the Speaking Only Task, respectively. This difference was tested with a model including a fixed effect for Task (with the levels Tapping and Speaking and Speaking Only; where the latter served as the reference level), and random slopes for participants and items on Task. The analysis revealed a significant intercept (*B* = 0.373, SE = 0.026, *p* < 0.001), indicating the occurrence of a reliable pause between turns. In addition, an effect of Task was observed (*B* = 0.078, SE = 0.025, *p* = 0.002), which shows that the gap between turns was significantly larger in the Tapping and Speaking than in the Speaking Only task.

Fig. 4 displays the average noun durations for each of the four positions in each speaking task. As can be seen the nouns were longer in the second and fourth than in the first and third position, and the durations in the two tasks were very similar. For the analysis of noun durations the model contained the fixed effects Task and Noun (with the levels Noun 1, Noun 2, Noun 3, Noun 4). Random slopes were included for the two main effects for both participants and items (the inclusion of interaction terms led to a failure to converge). A significant effect was observed for the Intercept (*B* = 0.312, SE = 0.008, *p* < 0.001), where the *B* value reflects the average duration for nouns in the Speaking Only task for nouns in the first position. No significant effect was observed for Task (*B* = −0.002, SE = 0.003, *p* = 0.543) as noun durations in the first position did not differ significantly across the two tasks. Significant effect were observed for each of the three positions (Noun 2: *B* = 0.047, SE = 0.010, *p* < 0.001; Noun 3: *B* = 0.029, SE = 0.010, *p* = 0.003; Noun 4: *B* = 0.100, SE = 0.013, *p* < 0.001) reflecting the fact that nouns in second, third and fourth position were longer in duration than those in first. No significant interaction effects were observed between Task and Noun for nouns in second and third position (Noun 2: *B* 0.0002, SE = 0.003, *p* = 0.942; Noun 3: *B* = −0.002, SE = 0.003, *p* = 0.518), which shows that nouns in these positions did not differ in length depending on the task. A significant interaction was observed between Task and Noun for the fourth position (*B* = 0.010, SE = 0.003, *p* = 0.001), reflecting the fact that in the fourth position nouns were uttered more slowly in the Tapping and Speaking than in the Speaking Only task. Note though, that this difference in noun durations was less than 10 ms.

The average durations of the sentences (onset of noun 1 to offset of noun 4) in the two speaking tasks were almost

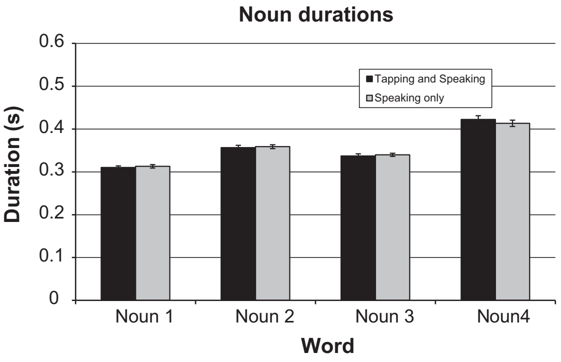

Fig. 4. Results of Experiment 1: Average noun durations (in seconds) for the four nouns for the Tapping and Speaking and the Speaking Only tasks. Error bars reflect the standard errors of the by-participant means.

identical (3.04 s and 3.02 s for the Tapping and Speaking and the Speaking only tasks, respectively). This is reflected in the analyses with only a significant effect for the Intercept ($B = 3.017$, SE = 0.031, $p < 0.001$), but not for Task ($B = 0.024$, SE = 0.022, $p = 0.279$).

2.2.3. Judgement task

On 2% of the trials no judgement response was given. On the remaining trials, participants almost always (on more than 94% of the trials in all tasks) indicated that correct utterances were correct. Incorrect utterances were identified as such on 70% of the trials in the Tapping and Speaking task, and on 60% of the trials in Tapping and Listening and the Speaking only task.

2.3. Discussion

The analyses showed that in the Tapping and Listening condition, tapping performance slightly deteriorated after the change of speaker (and especially at the end of the speaker 2 part). This effect is most likely a result of the monitoring task because during the Speaker 1 part, to which no monitoring task applied, no such decline was visible. Substantially more interference arose when the participants described the pictures themselves in the Tapping and Speaking task. This result is compatible with the intuition that speaking is harder than listening, and is in line with a number of previous reports (Almor, 2008; Boiteau et al., 2014; Recarte & Nunes, 2003). However, it is in contrast to the results of Kubose et al. (2006), who found that speaking and listening had very similar effects on performance in a secondary driving task. One reason for this difference in outcomes may be that the tasks used by Kubose et al. (2006) relied heavily on memory retrieval processes, as both the production and comprehension task involved creating mental maps of the positions of different buildings on the participants' campus. Memory retrieval processes, which involve strong dual-task costs (Naveh-Benjamin, Craik, Perretta, & Tonev, 2000), may have obscured potential differences between the comprehension and production tasks. Future research should address this issue further.

The top panel of Fig. 3 shows that the tapping rates in the Tapping and Speaking task were lowest about 2000 ms after speaker 1 offset, around the time that the participants were planning the third object name. This observation is surprising because earlier evidence suggests that the cognitive load for speakers is highest at utterance onset (Kemper et al., 2011). In our study, the speakers produced two clauses ("Put the A above the B and put the C above the D") on each trial, and they may have treated each of them as a planning unit (e.g., Smith & Wheeldon, 1999). This would predict approximately equal loads at utterance onset and around the middle of the utterance. The cognitive load may have been slightly lower before the onset of the first than before the onset of the second clause because encoding of the pictures for planning of the first clause benefitted from the preview of the objects at trial onset. In addition, when speakers were planning their first utterance fragment they were not simultaneously speaking and monitoring their own speech. However, as the participants were explicitly instructed to monitor and judge their own performance the cognitive load due to monitoring may have been higher than in everyday speaking tasks. Future studies might use the tapping task to examine the cognitive load prior and during the production of different types of utterances, and when speakers focus more or less on monitoring their speech.

The main research question for the current experiment was when participants would begin to engage in the capacity demanding processes of speech planning. According to the Early Hypothesis they should start preparing their utterance as soon as they knew which pictures they should describe. Therefore, tapping rates should decline soon after the offset of the first noun produced by speaker 1. According to the Late Hypothesis, speakers engage in the capacity demanding aspects of speech planning only at the end of the interlocutor's turn. The analyses unambiguously support the Late Hypothesis, as the continuous model assuming a decrease in tapping rate after the offset of the fourth noun was optimal. The optimal model showed that there was no difference in tapping rate between the two tasks before the offset of the fourth noun of speaker 1. The additional analysis focussing on the time period around the switch of speaker, showed that there was some overlap between listening and speech planning, as a difference arose from around 0.5 s before the offset of speaker 1 in the Tapping and Speaking task. Considering the complete duration of the turn of speaker 1 (about four seconds) and the fact that speakers already knew which pictures they should describe after about one second, this pattern is more consistent with the Late than the Early Hypothesis.

The analyses of the participants' speech onset latencies showed that they initiated their turn later in the Tapping and Speaking compared to the Speaking Only task. Thus, there was evidence for mutual interference between the two tasks. This pattern of results is in line with numerous earlier results demonstrating that performance in dual-task settings may reflect trade-offs between the tasks, leading to performance decrements in both tasks (e.g., Somberg & Salthouse, 1982). There was little effect of tapping on the spoken noun durations, which implies that tapping did not lead to significant motor interference. This observation suggests that the interference between tapping and speaking is mainly the result of interference at the level of speech planning and monitoring (Kunar et al., 2008; Oomen & Postma, 2002; Strayer & Johnston, 2001).

An additional observation was that the spoken durations of the nouns differed across the utterance positions (note that nouns were not counterbalanced across positions, but were very similar to each other with respect to a number of properties that may affect noun duration; see Appendix B for detail). The second noun was longer than the first and third one, and the fourth noun was the longest of all. This pattern can be readily explained by reference to the prosodic structure of the utterances: Speakers tend to lengthen phrase-final words, and utterance-final words receive additional lengthening (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991).

As noted, the increase in the speech onset latencies in the Tapping and Speaking compared to the Speaking Only task shows that tapping interfered with speech planning.

One possibility is that some or all of the cognitive processes involved in utterance planning (identifying the objects, selecting their names, retrieving the sound forms) were slowed down by the additional cognitive load imposed by the tapping task. Alternatively, speakers may have changed their planning strategy and initiated their speech planning later when they had to tap. We refer to the latter hypothesis as the Delayed Refocus Hypothesis. In other words, the additional cognitive load imposed by the tapping task may have biased the participants toward initiating their speech planning as late as possible. This would be interesting in its own right, but would imply that the tapping rates could not be seen as valid indicators of the time course of speech planning in the absence of tapping. One of the goals of Experiment 2 was to address this concern.

3. Experiment 2

In Experiment 2, the participants carried out the same tasks as in Experiment 1, but in addition to their speech and tapping performance their eye movements were recorded. This was done in order to obtain corroborating evidence about the onset of speech planning.

There is strong evidence from earlier studies demonstrating that listeners presented with descriptions of displays or scenes tend to look at the relevant objects. This was first shown in a seminal study by Cooper (1974) and confirmed in numerous experiments using the Visual World Paradigm (Huettig, Rommers, & Meyer, 2011 for review; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Similarly, many studies have shown that speakers producing descriptions of sets of objects or of events typically look at each of the objects they refer to shortly before naming it (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998). A likely basis for this strong link between linguistic processing and the listener's or speaker's eye gaze is that a person's point of gaze tends to correspond to their focus of visual attention (e.g., Irwin & Gordon, 1998) and that directing one's visual attention to an object facilitates processing it. Thus, listeners and speakers look at the relevant objects because this facilitates recognition of the objects and the retrieval of conceptual and linguistic information about them.

Based on the evidence from earlier eye tracking studies, we expected that the participants of the present experiment should look at the relevant objects in the order of mention. Thus, they should first look at the objects mentioned by speaker 1 and then at the objects mentioned by speaker 2. The most important question was when the participants in the speaking tasks would begin to fixate on the objects they had to name themselves. The Early Hypothesis predicts that this should happen as soon as they knew which objects they should describe, which was shortly after speaker 1 had named the first object. By contrast, the Late Hypothesis predicts that they should only turn to their own objects towards the end of the turn of speaker 1.

Given the results of Experiment 1, we expected to obtain further support for the Late Hypothesis. Thus, with regard to the speech and tapping performance we expected to replicate the results of Experiment 1. Additionally, we expected that the late decline in the participants' tapping performance in the Tapping and Speaking task would be accompanied by late shifts of gaze to the objects they had to describe.

As noted above, it is possible that participants used different planning strategies in the Speaking Only and in the Tapping and Speaking task, initiating their speech planning later when they had to tap. If this is the case, the shift of gaze to the speaker 2 objects should occur later in the Tapping and Speaking than in the Speaking Only task. Therefore, the registration of the speakers' eye movements allowed us not only to obtain further evidence about the coordination of listening and speech planning, but also to evaluate how the onset of speech planning was affected by the tapping task.

In Experiment 1, the participants only received a few minutes of training for the tapping task. The data from several participants had to be excluded because they could not combine tapping with the linguistic task. In Experiment 2, participants received more extensive tapping training. This should reduce the data loss in the experiment, and, since the variability in tapping performance should be lower, might improve the sensitivity of the tapping rate as a measure of cognitive load.

3.1. Method

3.1.1. Participants

Eighteen right-handed participants (four men) from the participant pool at the Max Planck Institute for Psycholinguistics were tested. All participants were university students, with a mean age of 21 years. They had not participated in Experiment 1. Participants received a small financial reward for their participation.

3.1.2. Apparatus

An Eyelink 100 eye-tracker (SR Research) along with the software packages Experiment Builder and Data Viewer (SR Research) was used to record the participants' eye movements. Apart from this, the same equipment was used as in Experiment 1. The movements of the right eye were recorded with a sampling frequency of 500 Hz.

3.1.3. Materials and design

The same design and materials were used as in Experiment 1. However, as a different lab was used and, due to the use of the eye tracker, the participants sat slightly further away from the screen than in Experiment 1, the visual angle of the line drawings was smaller (maximally 4 deg compared to 6 deg) and the same held for the arrows (2.5 deg instead of 4.5 deg).

3.1.4. Procedure

As in Experiment 1, the participants were initially familiarised with the pictures. Then the tapping training was conducted, which was more extensive than in the first experiment. There were four training tasks. The first and fourth tasks were identical to the tasks of Experiment 1; that is, participants had to tap until they had performed

50 consecutive correct taps. The second task consisted of solving 38 math problems while tapping the complex pattern. Participants first saw a math problem (e.g. "11 × 9") and after 8 s an answer (e.g. "97"). They had to indicate whether or not the answer was correct by saying *goed (correct)* or *fout (wrong)*. The third task was to engage in an informal 8-min conversation with the experimenter while continuously tapping the complex pattern. The experimenter initiated a number of topics, such as "favourite holiday" and "hobbies". Between the third and the fourth training part, the participant's knowledge of the picture names and their handedness was assessed as in Experiment 1. Different training tasks were selected in order to render the training session as stimulating as possible for the participants and to obtain pilot data concerning the usefulness of the tapping task in different research context. The procedure during the main experiment was identical to the procedure in Experiment 1, except that the participants' eye movements were recorded in addition to their speech and tapping. The participants sat in front of the eyetracker, approximately 1 m away from the screen. Head movements were restricted through the use of chin and forehead rests. The eye tracker was calibrated before each test block.

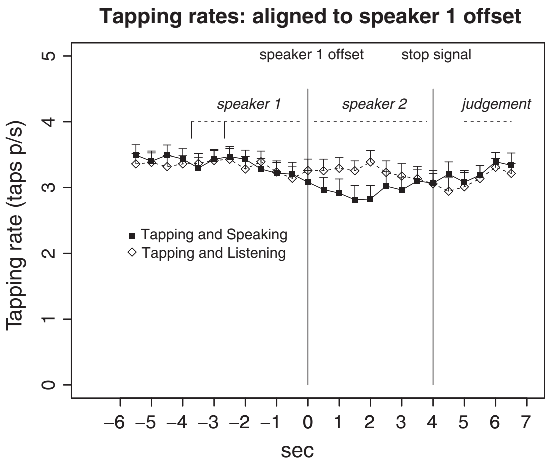

Fig. 5. Results of Experiment 2: Average tapping rates aligned to the offset of speaker 1.

3.2. Results

3.2.1. Tapping data

The tapping data were analysed in the same way as for Experiment 1. Invalid taps were excluded (24% of all taps), as were incorrect taps (9% of the valid taps). The results for the remaining data, shown in Fig. 5, were similar to those of Experiment 1. Again, the tapping rates in the Tapping and Speaking and Tapping and Listening task were very similar before and during the speaker 1 part, but differed in the speaker 2 part. As for Experiment 1, we examined the pattern of interference until the offset of the speaker 2 window.

The analysis approach was identical to that of Experiment 1. The models again included fixed main effects for Task and Block. Four models were built that differed in a Planning predictor representing specific planning hypotheses (Early Temporary, Late Temporary, Early Continuous, Late Continuous). A fifth, baseline model did not contain a planning predictor. Model fit was evaluated using AICc values. Interaction terms were included for the interactions between the Planning predictors and Task. Random slopes and intercepts for participants and items were included for the interaction between Task and the Planning predictors.

The top part of Table 2 contains the model comparison values. As in Experiment 1, the optimal model was "N4_Continuous". This model contained the predictor that assumed an increase of cognitive demand at speaker 1 offset, and this demand remained until the end of the speaker 2 part. The model had a significant intercept ($B = 3.32$, $SE = 0.154$, $p < 0.001$) reflecting the average overall tapping rate in the Tapping and Listening task. A marginal effect was observed for the Planning predictor ($B = -0.079$, $SE = 0.047$, $p = 0.095$), suggesting a trend for tapping performance to decline from the speaker 1 part to the speaker 2 part in the Tapping and Listening task. This trend may be related to the monitoring task, which focussed on the speaker 2 part. No main effect was observed for Task ($B = 0.044$, $SE = 0.127$, $p = 0.724$), indicating that, during the speaker 1 part (where the planning predictor had value 0 and where therefore the main effect of Task was estimated), Tapping and Speaking and Tapping and Listening led to a similar rate of tapping. In contrast to the results of Experiment 1, no significant main effect was observed for Block ($B = 0.076$, $SE = 0.080$, $p = 0.341$) showing that the increased training in Experiment 2 prevented a further increase in tapping rate over the experiment. A significant interaction was observed between Task and the Planning predictor ($B = -0.338$, $SE = 0.111$, $p = 0.002$). Similar to Experiment 1 this reflects the observation that in the Tapping and Speaking task there was significant interference between speech planning and tapping right around the offset of the first speaker's turn.

As for Experiment 1, a second analysis was carried out to examine the timing of interference in a window ranging between −2 and +2 s from speaker 1 offset. The bottom part of Table 2 reports the results. In the optimal model the planning predictor was aligned exactly to the offset

Table 2
Model selection values for tapping rates in Experiment 2. Models differ with respect to the shape of the Planning predictor over time. "Base" refers to the model without the planning predictor.

Predictor	*K*	*W*	Delta_AICc
Analysis 1: Planning across the trial			
N4_Continuous	26	1	0 (AICc = 66510.55)
N4_Temporary	26	0	106.07
N1_Continuous	26	0	126.34
N1_Temporary	26	0	252.23
Base	10	0	274.54
Analysis 2: Alignment of planning to speaker 1 offset			
0	26	0.97	0 (AICc = 35351.28)
+0.5	26	0.02	7.34
−0.5	26	0	13.53
−1.0	26	0	42.56

of the first speaker. Minimal additional support was found for the model that assumed an increase in interference within 0.5 s after the offset of the first speaker. The optimal model had a significant value for the Intercept ($B = 3.260$, SE = 0.158, $p < 0.001$). No main effect was observed for the Planning predictor ($B = 0.035$, SE = 0.046, $p = 0.451$), for Task ($B = 0.033$, SE = 0.155, $p = 0.831$) or Block ($B = -0.005$, SE = 0.091, $p = 0.950$). A significant interaction was observed between Task and the Planning predictor ($B = -0.437$, SE = 0.113, $p < 0.001$). This model displays very similar results as the previous analysis on a more extended time window.

3.2.2. Speech production data

Utterances that contained missing or incorrect object names or utterances for which the last noun was not initiated at least 150 ms before the stop signal were coded as errors. The error rates in the Tapping and Speaking and in the Speaking Only task were 9% and 6%, respectively. Errors were excluded from the following analyses, as were utterances with onset latencies deviating by more than 2.5 standard deviations from the mean (3% of the data). The average speech onset latency was longer in the Tapping and Speaking task (390 ms (SE = 34)) than in the Speaking Only task (329 ms (SE = 38)). This difference was tested with a model including a fixed effect for Task (with the levels Tapping and Speaking and Speaking Only; where the latter served as the reference level) and random slopes for participants and items on Task. The analysis revealed a significant intercept ($B = 0.330$, SE = 0.038, $p < 0.001$), indicating the occurrence of a pause between turns for the Speaking Only task. In addition, an effect of Task was observed ($B = 0.059$, SE = 0.021, $p = 0.004$) which shows that the pause between turns was significantly longer in the Tapping and Speaking task compared to the Speaking Only task.

Fig. 6 displays the average noun durations. For the analysis of noun durations the model contained the fixed effects Task and Noun (with the levels Noun 1, Noun 2, Noun 3, Noun 4). Random slopes were included for the two main effects and their interaction for both participants and items, keeping the random effects structure maximal. Model comparisons led to the exclusion of the fixed effects interaction term between Task and Noun ($\chi^2 = 2.487$, ΔDf = 3, $p = 0.478$). A significant effect was observed for the Intercept ($B = 0.337$, SE = 0.010, $p < 0.001$), where the B value reflects the average duration for nouns in the Speaking Only task for nouns in the first position. No significant effect was observed for Task ($B = 0.0007$, SE = 0.003, $p = 0.815$) as nouns did not differ significantly in duration across the two tasks. Significant effects were observed for each of the three positions (Noun 2: $B = 0.044$, SE = 0.012, $p < 0.001$; Noun 3: $B = 0.031$, SE = 0.011, $p = 0.003$; Noun 4: $B = 0.110$, SE = 0.013, $p < 0.001$) reflecting the fact that nouns in second, third, and fourth position were articulated more slowly than those in first position. The average duration of the speakers' utterances (onset of first noun to offset of last noun) was 2.62 s in both speaking tasks. All of these results closely replicate the findings of Experiment 1.

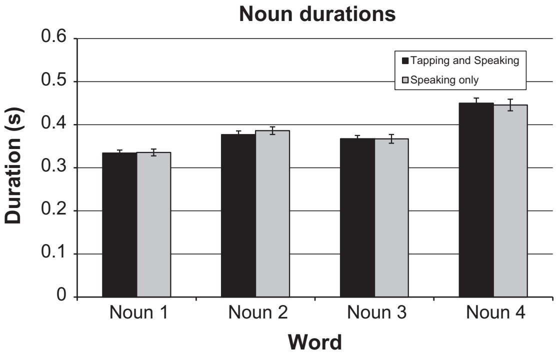

Fig. 6. Results of Experiment 2: Average noun durations per task. Error bars reflect the standard error of the by-participant means.

3.2.3. Judgement task

On 2% of the trials no judgement response was given. For the remaining trials, participants almost always (on more than 98% of the trials in all tasks) categorised correct responses as such. The rates of detecting errors were 64%, 68%, and 66% for the Tapping and Speaking, the Tapping and Listening, and the Speaking Only task, respectively.

3.2.4. Eye movement data

To analyse the eye movements we defined regions of interest for each of the eight objects (0.5 cm around edges of the objects) and categorised the participants' fixations as falling onto any of the objects mentioned by speaker 1 or onto any of the objects mentioned by speaker 2 or elsewhere. Fixations with durations below 80 ms were discarded as spurious. Subsequent fixations onto the same objects (i.e., fixations to different parts of the same object) were combined into gazes.

For each task, Fig. 7 displays the proportions of all gazes across the trial that were directed to speaker 1 objects (top panel) or to speaker 2 objects (bottom panel). Gazes are aligned to trial onset. At the top of each panel indications are provided for the minimal and maximal length of speaker 1's utterances (indicated with small vertical marks). Due to the variable length of speaker 1's utterances, the onset and offset of the speaker 2 part also varied. The earliest and latest start and end times are also indicated with small vertical marks. The figure shows that shortly after trial onset, the participants had no preference for speaker 1 objects or speaker 2 objects. This is because they could not anticipate which row of objects the first speaker would describe. After about 4 s (1 s after speaker 1 onset) they focussed mostly on the speaker 1 objects. Towards the end of the turn of speaker 1, they tended to shift their gaze towards the objects of speaker 2. Finally, after the end of the speaker 2 part (when the objects were no longer present), the participants' looks moved away from the interest areas, often to the stop signal displayed in the centre of the screen. These observations confirm that the participants mostly looked at the task-relevant objects.

As long as the objects were in view, the majority of the gazes (83%) fell within the pre-defined interest areas, and therefore the fixation patterns for speaker 1 objects and

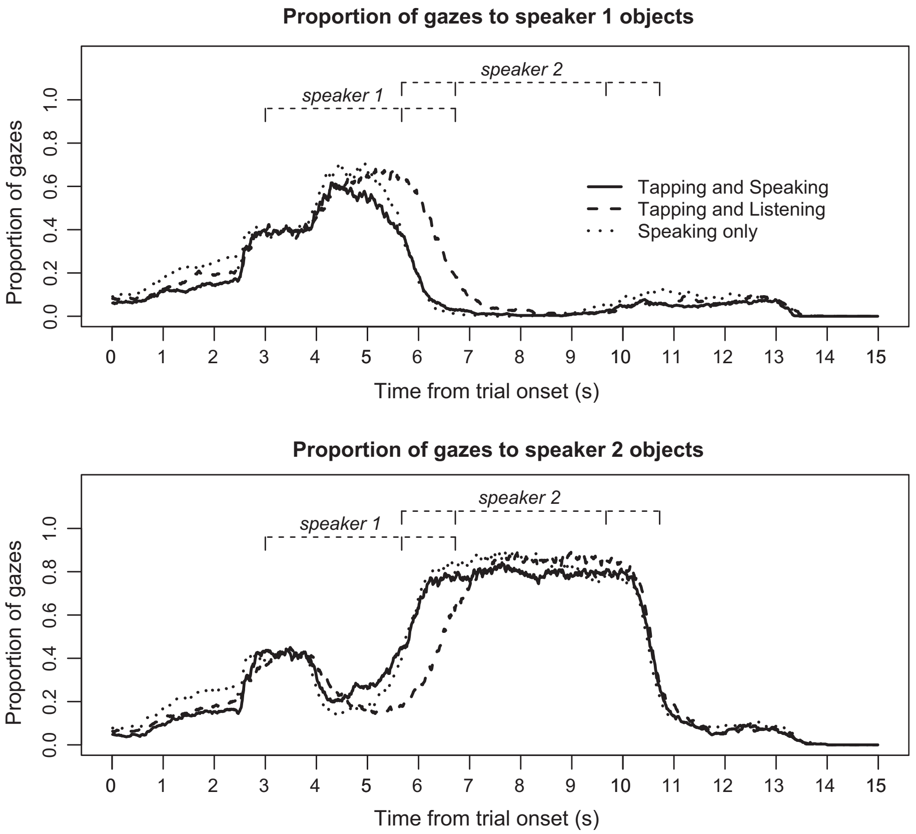

Fig. 7. Results of Experiment 2: Proportions of gazes to the objects of speaker 1 (top panel) or the objects of speaker 2 (bottom panel), in the three tasks.

speaker 2 objects were largely complementary: More gazes to speaker 1 objects were accompanied by fewer gazes to speaker 2 objects. Since we were primarily interested in gazes to speaker 2 objects, we report analyses of the proportions of gazes to speaker 2 objects (proportions out of gazes to speaker 1 objects and speaker 2 objects combined).

The gaze data were analysed in a similar way as the tapping data. Two predictor variables were created, namely an Early Planning predictor, which started at a value of 0 and was set to a value of 1 after the offset of the first noun of speaker 1, and a late predictor, which also started at a value of 0, but was set to a value of 1 only after the offset of the fourth noun of speaker 1. Since speakers can be assumed to keep fixating objects that they are describing we did not consider the "temporary" variants of the predictors. Models of gaze behaviour were fitted, and the model with the best fit to the data was selected by means of model comparisons.

The models included the fixed effects Task (with the levels Tapping and Speaking, Tapping and Listening, and Speaking Only; with Tapping and Speaking as the reference level), and the Planning predictors, along with their interaction. The interaction and main effects were also included in the random effects structure, with slopes and intercepts for participants and intercepts for items (inclusion of slopes led to over-fitting). Gaze data were analysed in the window between 3 s (picture onset) and 10 s after trial onset.

Table 3 reports the fit information for the two models. The optimal model assumes alignment to Noun 4 offset. This shows that across the three tasks, the shift of gaze was more closely aligned to the offset of the fourth noun than to that of the first noun of speaker 1. The optimal model had a significant effect for the Intercept ($B = 0.458$, $SE = 0.024$, $p < 0.001$), reflecting the overall proportion of looks to speaker 2 objects in the Tapping and Speaking Task before the offset of the fourth noun. A main effect was found for the Planning predictor ($B = 0.507$, $SE = 0.028$, $p < 0.001$) indicating that, in the Tapping and Speaking task, looks to speaker 2 objects increased close to the offset of the fourth noun of speaker 1. A main effect was observed for Task at the level of Tapping and Listening ($B = -0.122$, $SE = 0.012$, $p < 0.001$), indicating that before the offset of the fourth noun, there were fewer looks to speaker 2 objects in the Tapping and Listening task than in the Tapping and Speaking task. This suggests that participants started looking at the speaker 2 objects later when they did not have to describe the objects themselves. In addition, there was an interaction between Task at the level of Tapping and Listening and the Planning predictor ($B = 0.089$, $SE = 0.015$, $p < 0.001$), reflecting a steeper slope of shifting gaze from speaker 1 objects to speaker 2 objects in the Tapping and Listening than in the Tapping and

Table 3
Model selection values for models of proportions to speaker 2 objects that only differ with respect to the timing of the Planning predictor.

Alignment	*K*	*W*	Delta_AICc
Noun 4	18	1	0 (AICc = 12462.19)
Noun 1	18	0	12490.89
Base	8	0	13553.96

Table 4
Model selection values for models of gazes to speaker 2 objects that differed with respect to the timing of the Planning predictor (−1 to +0.5 s relative to speaker 1 offset).

Alignment	*K*	*W*	Delta_AICc
Tapping and speaking			
−0.5	7	1	0 (AICc = 623.94)
0	7	0	849.38
−1.0	7	0	1398.45
+0.5	7	0	2243.41
Tapping and listening			
0	7	1	0 (AICc = 2487.52)
+0.5	7	0	450.07
−0.5	7	0	1326.84
−1.0	7	0	2791.76
Speaking only			
−0.5	7	1	0 (AICc = −643.80)
0	7	0	1890.45
−1.0	7	0	2145.70
+0.5	7	0	3856.45

Speaking task (i.e., the shift from speaker 1 objects to speaker 2 objects was slightly more gradual in the Tapping and Speaking task than in the Tapping and Listening task. This may partly be due to the lower proportion of looks to speaker 2 objects in the Tapping and Listening part during the speaker 1 part). A weak main effect was observed for Task at the level of Speaking Only ($B = -0.044$, SE = 0.020, $p = 0.033$), where the negative B weight reflects the fact that, before the offset of the fourth noun of speaker 1, there were *fewer* looks to speaker 2 objects in the Speaking Only task than in the Tapping and Speaking task. This observation provides direct evidence against the Delayed Refocus Hypothesis because on this hypothesis there should be more looks to speaker 2 objects in the Speaking Only task before the offset of the fourth noun of speaker 1.

In addition there was a weak interaction between Task at the level of Speaking only and the Planning predictor ($B = 0.052$, SE = 0.024, $p = 0.034$) reflecting a steeper slope of shifting gaze from speaker 1 objects to speaker 2 objects in the Speaking Only than in the Tapping and Speaking task.

Visual inspection of the results shown in Fig. 7 suggested that the shift of gaze was tightly linked to the offset of the speech of speaker 1. To investigate the time course of the gaze shift in more detail, an analysis was carried out where the Late Planning predictor was aligned exactly with the offset of the utterance or shifted by +0.5 s (i.e. occurred slightly later in time), by −0.5 s, and by −1.0 s. For each of the three tasks, models were compared with each of these four predictors. Table 4 presents the results. It can be observed that for both the Tapping and Speaking and the Speaking only task, the shift of gaze towards speaker 2 objects was best aligned to the predictor that changed from 0 to 1 half a second before the offset of the fourth noun. In the Tapping and Listening task, however, looks were more closely aligned to the offset of the fourth noun itself.

The final set of analyses investigated the relationship between gazes and tapping performance. Specifically, we explored whether in the Tapping and Speaking task the increase in the proportion of looks to the objects the participants had to name coincided with an increase in interference, as captured in the tapping performance. As described above, at trial onset the participants looked at both rows of objects about equally often because they did not know which row speaker 1 would describe. Therefore, the early gazes to the speaker 2 objects were probably not directly related to the participants' speech planning. For the analysis of the alignment of the tapping rates to the "looks for naming", we determined for each trial in the Tapping and Speaking task the onset of the gaze to a speaker 2 object that was closest in time to the participant's speech onset. For the Tapping and Listening condition we determined for every trial the onset of the gaze that was closest in time to the onset of the utterance of the pre-recorded speaker 2. In the Tapping and Speaking task, the participants typically looked at the object before naming it. In the Tapping and Listening task, however, they sometimes only looked at the object just *after* it had been mentioned by speaker 2. Tapping rates were aligned to the fixation onset time that was *closest* to the onset of the object name (i.e., irrespective of whether it preceded or followed name onset, for both tasks). Fig. 8 displays the average tapping rates in nine 0.5 s time windows around the initiation of the looks for naming (indicated by the vertical line at 0). As can be observed, the onset of the first look for naming was followed by an immediate decrease in tapping rate in the Tapping and Speaking task, but there was no analogous decrease in the Tapping and Listening task.

To statistically assess these patterns four Planning predictors were created for the time window from 2 s before until 2 s after the onset of the first look for naming. The predictors started at value 0 and changed to 1 from −0.5 s before the onset of the critical gaze, exactly at gaze onset (0 s), or +0.5 s, or +1 s after gaze onset and onwards. With each of these predictors an lmer model was fitted that included the fixed factors Planning predictor, Task, and their interaction. An additional base model was fitted that had no Planning predictor in its terms. Table 5 displays the results. The optimal model assumed an increase in dual-task interference from 0.5 s after the onset of the critical look. However, some support was also found for the model assuming an increase in interference immediately at the onset of the critical gaze.

The optimal model had a significant effect for the Intercept ($B = 3.323$, SE = 0.155, $p < 0.001$). No main effect was found for the Planning predictor ($B = 0.006$, SE = 0.051, $p = 0.912$) reflecting the fact that there was no change in tapping after the +0.5 time point in the Tapping and Listening task. There was no main effect for the factor Task ($B = -0.037$, SE = 0.159, $p = 0.813$) indicating that there

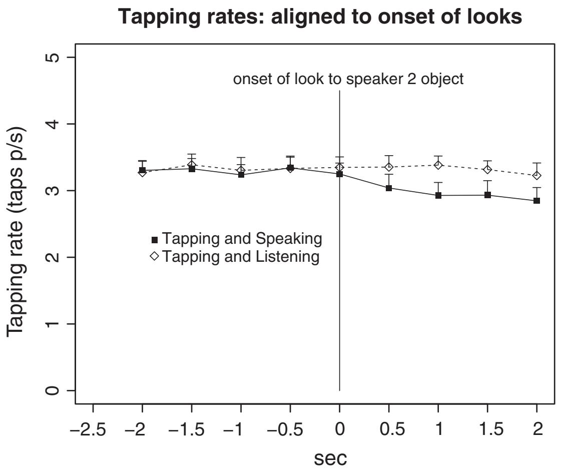

Fig. 8. Results of Experiment 2: Average tapping rates aligned to the first look for naming of speaker 2 objects. Error bars reflect the standard error of the by-participants means.

was no overall difference in tapping rates between the two tasks before the onset of looks for naming. A significant interaction was observed between Task and the Planning predictor ($B = -0.398$, SE = 0.119, $p < 0.001$), which shows that there was a significant decrease in tapping rate for the Tapping and Speaking task around 0.5 s after the first look for naming.

3.3. Discussion

In Experiment 2, the participants received more extensive training in the tapping task than in Experiment 1. This reduced the effect of the tapping task on speech onset latencies in the comparison between the Speaking Only and the Tapping and Speaking task (from 80 ms to 61 ms), suggesting that ample practice reduced the load imposed by the tapping task and the interference with speech planning. Future experiments could test whether more extreme training regimes would eliminate the interference effect (see e.g., Spelke, Hirst, & Neisser, 1976; Strobach, Liepelt, Pashler, Frensch, & Schubert, 2013). However, to measure cognitive load in turn-taking settings, a moderate amount of practice appears to be optimal, as it leads to good performance in the task, little data loss due to participants' inability to perform the task, and reliable interference effects.

With respect to the participants' performance in the tapping task in different parts of the trial, the main results of Experiment 1 were replicated. Fig. 5 shows that in the Tapping and Listening task, tapping performance during the two listening parts differed little from each other, and only declined in the judgement phase, when participants had to indicate whether or not the utterance produced by speaker 2 was correct. In the Tapping and Speaking task, tapping declined only at the end of the turn of speaker 1. This pattern confirms that participants postponed the cognitively demanding aspects of preparing their turn until just before they took over the turn, consistent with the Late Hypothesis.

Table 5
Model selection values for models of the tapping rates that differed with respect to the timing/presence of the Planning predictor. Alignment is to the onset of the first looks for naming.

Alignment	*K*	*W*	Delta_AICc
+0.5	25	0.86	0 (AICc = 32587.27)
0	25	0.14	3.65
+1.0	25	0	35.24
−0.5	25	0	41.64
Base	9	0	91.24

In the analyses of the participants' speech we again observed a delay in speech onset in the Tapping and Speaking compared to the Speaking Only task, and no difference in noun durations between these tasks. This confirms that the tapping task interfered with the preparation of the utterance occurring before speech onset, but did not strongly interfere with the execution of the articulatory commands. The influence of tapping on speech preparation is in line with the findings of Boiteau et al., despite the differences in setting. Boiteau et al., did not measure gap durations but reported the occurrence of overlaps (speaker 2 starts speaking before the end of the turn of speaker 1). They found that the dual-task situation led to a decrease of overlaps, and the average gap duration and the number of overlaps are likely to be strongly related (smaller gaps on average increase the likelihood of overlaps). In both studies, therefore, interference of the secondary task on speech production was especially observed at the start of the speaker's turn.

The analyses of the participants' eye movements showed that they looked preferentially at task-relevant parts of the display. In the Tapping and Listening task, they looked at the objects named by speaker 1 and speaker 2 in the order of mention. This replicates numerous earlier findings from studies using the Visual World Paradigm demonstrating that listeners tend to look at objects being named (Huettig et al., 2011). Most importantly, we found that participants initiated "looks for naming" late during the turn of speaker 1. Thus, the analyses of the participants' tapping performance and of their eye movements both support the Late Hypothesis. Moreover, we established that looks to objects for naming were immediately followed by a decrease in tapping performance. In other words, the cognitive load increased as soon as the participants directed their visual attention to the objects they had to name. To our knowledge, this is a novel finding, which we consider important in its own right.

Finally, one might expect that speakers would, perhaps strategically, postpone their speech planning in the Tapping and Speaking compared to the Speaking Only task because of the additional cognitive load imposed by the tapping task (as per the Delayed Refocus Hypothesis). The eye movement analyses did not support this view. In fact, the eye movement data suggested that participants started looking at the objects they had to name *earlier* in the Tapping and Speaking task than in the Speaking Only task. Appendix C shows that in the Tapping and Speaking task participants gazed at the third and fourth objects of speaker 1 less frequently than in the Speaking Only task. That is, without a dual task setting participants may even be more inclined to adopt a late planning strategy.

4. General discussion

Two experiments were carried out using a dual-task paradigm involving turn-taking along with a continuous complex tapping task. In both experiments there were three tasks: In the Tapping and Speaking task, participants first listened to a pre-recorded speaker and then took over the turn while performing a tapping task. In the Tapping and Listening task, they overheard two pre-recorded speakers taking turns while tapping the complex pattern. In the Speaking Only task they listened to the first speaker and then took over the turn without tapping. In Experiment 2, participants received more extensive tapping training than in Experiment 1 and their eye movements were recorded in addition to their speech and tapping performance. Tapping performance was recorded to assess variations in the cognitive load imposed by the task, and the participants' eye movements provided information about their allocation of visual attention.

The main objective of the study was to investigate when participants would begin to plan their own turn, which should be accompanied by a decline in tapping performance and a shift of visual attention to the objects to be named. It is generally assumed that in natural conversations speakers often begin to plan their utterances while listening to their interlocutor (De Ruiter et al., 2006; Sacks et al., 1974). Given that inter-turn intervals tend to be quite short (Heldner & Edlund, 2010; Stivers et al., 2009), often shorter than the time a speaker would need to plan a complex utterance from scratch (Indefrey & Levelt, 2004; Strijkers & Costa, 2011), this assumption is plausible. However, so far it is not substantiated by much empirical evidence. In fact, to our knowledge, only a single study (Boiteau et al., 2014) had investigated when, during the turn of the interlocutor or afterwards, speakers begin to plan their utterances. The present study complements the work by Boiteau et al. (2014) by using a different experimental setup. We created a situation where the content of the utterances the participants heard and produced was tightly controlled and where they could plan their utterances either early, as soon they knew which objects they had to describe, or late, close to the offset of the first speaker's turn. We asked which planning strategy the participants would prefer.

The experiments provide a clear answer to this question: The analyses of the participants' tapping performance and of their eye movements showed that they began to plan their utterance just before the end of the preceding speaker's turn. Thus, listening and speech planning overlapped, as many authors had suspected. However, the overlap was small. Although the participants knew as soon as speaker 1 had named the first of four objects which objects they should describe, they typically postponed utterance planning by about 2 s, until speaker 1 had initiated the last of the four object names.

Why did the participants opt for late utterance planning? A simple answer is that this was the easiest way to accomplish the task. As discussed in the Introduction, many studies have shown that speech comprehension and production, require processing capacity (Kemper et al., 2003; Kubose et al., 2006). If participants aim to keep their mental load at an even and relatively low level throughout the trial, they should minimise the temporal overlap between speaking and listening.

Against this, one may argue that the participants' tapping performance indicated that the listening task was low in central capacity demands. Hence, it should be easy to combine listening and speech planning. However, this argument overlooks the possibility that an increase in processing load will arise when speakers plan their utterance early and then have to hold it in working memory until the end of the first speaker's turn. In addition, the processing load may also increase substantially due to domain-specific interference between speech planning and listening. Numerous studies using the picture–word interference paradigm have shown that it takes speakers longer to name objects in the presence of spoken or written distracter words than in the absence of such distracters (Glaser & Düngelhoff, 1984; Schriefers, Meyer, & Levelt, 1990). This delay occurs because competition arises and needs to be resolved between representations activated by the distracter words and by the targets. Resolving this competition requires processing resources (Piai, Roelofs, & Schriefers, 2014). In other words, though listening to speaker 1 was not particularly demanding, planning an utterance while doing so may induce a considerable load because of interference between the representations activated during listening and speech planning.

In light of these considerations, one might wonder why participants did not initiate their speech planning even later, after the offset of the utterance of Speaker 1. The participants' timing of utterance planning may have been affected by the need to complete the utterance before the stop-signal; and in further research one might assess how the timing of utterance planning is affected when speakers are given more or less stringent response deadlines. However, we think that initiating utterance planning just before the end of the interlocutor's turn is likely to be a default strategy that speakers use in many situations. It offers a good balance between keeping the cognitive load of concurrent listening and speaking low and being able to respond promptly to the interlocutor.

Support for our view that speakers typically – not just in our experiment – initiate their utterance planning shortly before the anticipated end of the interlocutor's turn comes from the similarity of the results of the present study to those obtained by Boiteau and colleagues. As described in the Introduction, Boiteau et al. (2014) also used a dual-task paradigm to track the capacity demands arising during turn taking, but the two studies differed in the secondary task (tracking a visual target versus finger tapping) and the structure of their linguistic tasks (engaging in actual conversation versus picture description). In spite of these important differences, both studies led to the same main conclusion, namely that speakers begin to plan their utterance shortly before the offset of the preceding turn.

Though linking utterance planning to the anticipated end of the preceding turn may be the speaker's default planning strategy, it is, of course, not mandatory. There are undoubtedly situations where speakers have fully planned their turn much earlier (e.g., when they are interrupted by the interlocutor) and situations where they are literally speechless long after the preceding turn has ended. There are many variables that could potentially affect the speakers' timing of utterance planning, including, for instance, the pressure to respond fast (Roberts & Francis, 2013; Swets et al., 2013), the intelligibility and complexity of the preceding turn, the speaker's linguistic ability, and working memory capacity. An important task for further research is to find out which, if any, of these variables actually affect the onset of speech planning and the speed and efficiency of planning. Moreover, the proposal that speech planning is linked to the anticipated end of the interlocutor's turn presupposes that the speaker can indeed predict when the turn will end. This may be easier in some situations than in others. Thus, another important research question is how well speakers can predict the ends of turns, which variables they use to do so (see e.g., De Ruiter et al., 2006), and how the predictability of ends of turns affects their speech planning.

Evidently, using a Late Planning strategy implies that the speaker cannot fully plan a complex utterance before the end of the interlocutor's turn. So far, very little is known about speakers' planning spans in natural conversations. Studies have shown that speakers are highly flexible in their use of advance planning strategies and planning increments (Ferreira & Swets, 2002; Konopka, 2012; Swets et al., 2013), and that they often initiate utterances based on partial utterance plans, corresponding, for instance, to a subject noun phrase. Analyses of corpora of casual speech also indicate that speakers often have not fully planned their turns, but need to buy time for further planning by producing fillers such as "ehm", which accounted for about 3.5% of all spoken words in a sample by Torreira, Adda-Decker, and Ernestus (2010; see also Clark & Fox Tree, 2002). These observations are consistent with the view that late planning is a default strategy speakers use in many situations. We predict that speakers, rather than beginning to plan early, initiate complex utterances on the basis of partial utterance plans.

The current report and the report by Boiteau et al. demonstrate that both complex tapping and the visual-motor tracking task can be used to track the capacity demands of speaking and listening over time. One may ask which of these tasks is preferable for a specific research question. The visual-motor tracking task has better temporal resolution than the tapping task (here 0.5 s), although the temporal resolution of complex tapping can be improved by increasing the number of observations. Another advantage of the tracking task is that it is easy to manipulate task difficulty by changing the speed of the moving target. The tapping task is somewhat more versatile as it is not dependent on exogenous stimuli. Therefore it can, for instance, be used when participants are engaged in natural conversation with mutual eye contact or when they need to look at a screen to carry out a linguistic task. Depending on the research question, a tracking task or complex tapping may be preferred. The study by Boiteau and colleagues and the present study have extended the set of available research tools, now allowing for further investigations of the time course of cognitive demands in dialog settings.

The linguistic tasks in the two studies differed markedly in the constraints on the utterance content set for the participants. Whereas the participants in the study by Boiteau et al. engaged in informal conversations about everyday topics, the participants of the present study heard and produced descriptions of simple drawings. In spite of this marked difference in the way the utterances were elicited the main conclusion of both studies was the same. Clearly, the speakers' tendency to initiate utterance planning time-locked to the offset of the preceding speaker's turn is quite pervasive and is seen regardless of how the utterances are elicited. Other aspects of utterance planning and resource allocation during speaking and listening may well be sensitive to the way utterances are elicited. Researchers need to decide which way of eliciting utterances is most suitable for assessing their hypotheses, and in particular how important it is in the specific research context to control which utterances participants hear and produce.

Perhaps the most striking difference between the studies lies in the social context provided in the experiments, i.e. whether the participants produced utterances directed at a confederate or friend (Experiments 1 and 2, respectively, in the study by Boiteau and colleagues), or listened to and responded to pre-corded utterances (the study reported here). In the present set of studies, these tasks led to very similar conclusions. Again, however, other aspects of sentence planning may very well depend on the social situation in which the utterances are produced. As most experimental psycholinguistics has focussed on monologues, hardly anything is known about the way language comprehension and production processes are affected by the presence or absence of an interlocutor. At the moment, researchers need to rely on their best judgement to decide whether or not their specific research question is best addressed using procedures that do or not involve a confederate or testing participants interaction with each other. Ideally, important theoretical questions should be studied in more than one way.

5. Conclusions

It has often been proposed that smooth transitions between turns in natural dialogue require speakers to plan their utterances while listening to the preceding speaker. Our experiments confirmed that this view is correct, as there was indeed some temporal overlap between listening and speech planning. However, our results also showed that speakers only initiated the cognitively demanding aspects of speech planning when their interlocutor's turn was almost completed. We propose that such late utterance planning may be a default strategy speakers adopt in many situations. Planning while listening may therefore be less pervasive than one might think.

Acknowledgements

We thank Johan Weustink for building the tapping-box, Tim Korbmacher and Samantha van den Hoogen for assistance with testing participants and annotating the production data, Joost Rommers and Lies Cuijpers for lending their voices for the stimuli. We also thank the members of the MPI Dialogue Project for helpful discussions of the research reported here.

Appendix A. Materials

Dutch names with English translation equivalents. Lists were created such that semantic overlap between lists was minimal.

List 1: vlinder (butterfly), leeuw (lion), spin (spider), fles (bottle), hond (dog).
List 2: appel (apple), banaan (banana), peer (pear), doos (box), aardbei (strawberry).
List 3: zaag (saw), schaar (scissors), lepel (spoon), kaars (candle), vork (fork).
List 4: muur (bricks), dak (roof), sleutel (key), vlag (flag), tent (tent).
List 5: bril (glasses), helm (helmet), schoen (shoe), gitaar (guitar), pet (hat).
List 6: clown (clown), voet (foot), robot (robot), spiegel (mirror), ballon (balloon).
List 7: cactus (cactus), blad (leaf), boom (tree), pijp (pipe), bloem (flower).
List 8: fiets (bicycle), bus (bus), trein (train), ring (ring), kaas (cheese).

Appendix B. Item properties

See Table B1.

Appendix C

See Fig. C1.

Table B1
Average length and log frequency of the dominant name, taken from the picture corpus by Severens et al. (2005), for the names of the pictures in the four positions on the speaker 2 row of objects (standard deviations between brackets).

Measure	Position			
	Noun 1	Noun 2	Noun 3	Noun 4
Number of syllables	1.24 (0.34)	1.26 (0.44)	1.29 (0.46)	1.28 (0.45)
Number of phonemes	4.0 (0.86)	4.01 (0.98)	4.02 (1.02)	4.06 (1.09)
Log frequency	1.47 (0.50)	1.47 (0.53)	1.37 (0.48)	1.37 (0.55)

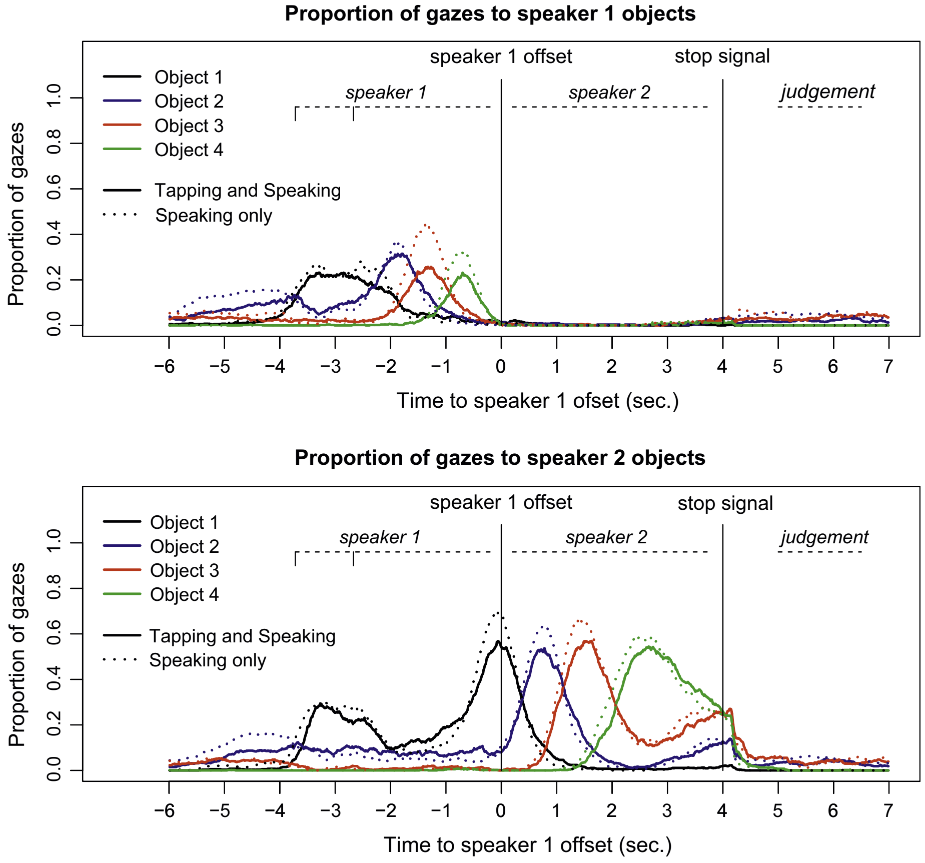

Fig. C1. Experiment 2: Proportions of gazes to the four objects of speaker 1 (top panel) and the four objects of speaker 2 (bottom panel) in the Tapping and Speaking (solid lines) and the Speaking Only (dotted lines) task. Gazes are aligned to speaker 1 offset (at 0 s).

References

Almor, A. (2008). Why does language interfere with vision-based tasks? *Experimental Psychology, 55*(4), 260–268.

Baddeley, A. D. (1976). *The psychology of memory*. New York: Basic Books.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47–89.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4*, 328.

Bates, D., Maechler, M., & Bolker, B. (2013). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-2*.

Becic, E., Dell, G. S., Bock, K., Garnsey, S. M., Kubose, T., & Kramer, A. F. (2010). Driving impairs talking. *Psychonomic Bulletin & Review, 17*(1), 15–21.

Belke, E. (2008). Effects of working memory load on lexical-semantic encoding in language production. *Psychonomic Bulletin & Review, 15*(2), 357–363.

Bergen, B., Medeiros-Ward, N., Wheeler, K., Drews, F., & Strayer, D. (2013). The crosstalk hypothesis: Why language interferes with driving. *Journal of Experimental Psychology: General, 142*(1), 119.

Bock, K., Dell, G. S., Garnsey, S. M., Kramer, A. F., & Kubose, T. T. (2007). 2 Car talk, car listen. In A. S. Meyer, L. Wheeldon, & A. Krott (Eds.), *Automaticity and control in language processing* (pp. 21–42).

Bodwell, J. A., Mahurin, R. K., Waddle, S., Price, R., & Cramer, S. C. (2003). Age and features of movement influence motor overflow. *Journal of the American Geriatrics Society, 51*(12), 1735–1739.

Boersma, P., & Weenink, D. (2009). *Praat: Doing phonetics by computer (Version 5.1)*.

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.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304.

Caplan, D., & Waters, G. (2013). Memory mechanisms supporting syntactic comprehension. *Psychonomic Bulletin & Review, 20*(2), 243–268.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speech. *Cognition, 84*, 73–111.

Cleland, A. A., Tamminen, J., Quinlan, P. T., & Gaskell, M. G. (2012). Spoken word processing creates a lexical bottleneck. *Language and Cognitive Processes, 27*(4), 572–593.

Cook, A. E., & Meyer, A. S. (2008). Capacity demands of phoneme selection in word production: New evidence from dual-task experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 886.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*(1), 84–107.

Crowther, J. E., & Martin, R. C. (2014). Lexical selection in the semantically blocked cyclic naming task: The role of cognitive control and learning. *Frontiers in Human Neuroscience, 8*, 1–20.

De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 515–535.

D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature, 378*(6554), 279–281.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review, 87*(3), 272.

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1187.

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

Ford, M., & Holmes, V. M. (1978). Planning units and syntax in sentence production. *Cognition, 6*(1), 35–53.

Fraser, S. A., Li, K. Z. H., & Penhune, V. B. (2010). Dual-task performance reveals increased involvement of executive control in fine motor sequencing in healthy aging. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 65*(5), 526–535.

Garrett, M. F. (1982). Production of speech: Observations from normal and pathological language use. In A. W. Ellis (Ed.), *Normality and pathology in cognitive functions* (pp. 19–76). London: Academic Press.

Glaser, W. R., & Düngelhoff, F.-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: Human Perception and Performance, 10*(5), 640–654.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language, 57*(4), 544–569.

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology, 25*(1), 1–42.

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

Grosjean, F., Grosjean, L., & Lane, H. (1979). The patterns of silence: Performance structures in sentence production. *Cognitive Psychology, 11*(1), 58–81.

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

Hiscock, M., Cheesman, J., Inch, R., Chipuer, H. M., & Graff, L. A. (1989). Rate and variability of finger tapping as measures of lateralized concurrent task effects. *Brain and Cognition, 10*(1), 87–104.

Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(1), 196–205.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151–171.

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

Irwin, D. E., & Gordon, R. D. (1998). Eye movements, attention and trans-saccadic memory. *Visual Cognition, 5*(1–2), 127–155.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: An interdisciplinary perspective* (pp. 166–196). Clevedon, UK: Multilingual Matters.

Kahneman, D. (1973). *Attention and effort*. New York: Prentice Hall.

Kemper, S., Herman, R. E., & Lian, C. H. T. (2003). The costs of doing two things at once for young and older adults: Talking while walking, finger tapping, and ignoring speech of noise. *Psychology and Aging, 18*(2), 181–192.

Kemper, S., Herman, R. E., & Nartowicz, J. (2005). Different effects of dual task demands on the speech of young and older adults. *Aging, Neuropsychology, and Cognition, 12*(4), 340–358.

Kemper, S., Schmalzried, R. L., Herman, R., & Mohankumar, D. (2011). The effects of varying task priorities on language production by young and older adults. *Experimental Aging Research, 37*(2), 198–219.

Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language, 66*(1), 143–162.

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

Kunar, M. A., Carter, R., Cohen, M., & Horowitz, T. S. (2008). Telephone conversation impairs sustained visual attention via a central bottleneck. *Psychonomic Bulletin & Review, 15*(6), 1135–1140.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences, 9*(2), 75–82.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General, 133*(3), 339–354.

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press.

Levinson, S. C. (2013). Action formation and ascription. In T. Stivers & J. Sidnell (Eds.), *The handbook of conversation analysis* (pp. 103–130). Malden, MA: Wiley-Blackwell.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences, 9*(6), 296–305.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology, 59*(3), 203–243.

Mazerolle, M. J. (2013). *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c) (Version R package version 1.35.)*.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: I. Basic mechanisms. *Psychological Review, 104*(1), 3–65.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66*(2), B25–B33.

Naveh-Benjamin, M., Craik, F. I. M., Perretta, J. G., & Tonev, S. T. (2000). The effects of divided attention on encoding and retrieval processes: The resiliency of retrieval processes. *The Quarterly Journal of Experimental Psychology: Section A, 53*(3), 609–625.

Oldfield, R. C. (1971). Assessment and analysis of handedness – Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113.

Oomen, C. C. E., & Postma, A. (2002). Limitations in processing resources and speech monitoring. *Language and Cognitive Processes, 17*(2), 163–184.

Papesh, M. H., & Goldinger, S. D. (2012). Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, & Psychophysics, 74*(4), 754–765.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance, 10*(3), 358–377.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116*(2), 220–244.

Piai, V., Roelofs, A., & Schriefers, H. (2014). Locus of semantic interference in picture naming: Evidence from dual-task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 147–165.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America, 90*, 2956–2970.

R_Core_Team (2013). *R: A language and environment for statistical computing (Version 3.0.1)*. R Foundation for Statistical Computing, Vienna, Austria.

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied, 9*(2), 119–137.

Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America, 133*(6), EL471–EL477.

Roberts, F., Margutti, P., & Takano, S. (2011). Judgments concerning the valence of inter-turn silence across speakers of American English, Italian, and Japanese. *Discourse Processes, 48*(5), 331–354.

Roelofs, A., & Piai, V. (2011). Attention demands of spoken word planning: A review. *Frontiers in Psychology, 2*, 307.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735.

Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language, 29*(1), 86–102.

Seth-Smith, M., Ashton, R., & McFarland, K. (1989). A dual-task study of sex differences in language reception and production. *Cortex, 25*(3), 425–431.

Severens, E., Van Lommel, S., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica, 119*(2), 159–187.

Shao, Z., Roelofs, A., & Meyer, A. S. (2012). Sources of individual differences in the speed of naming objects and actions: The contribution of executive control. *The Quarterly Journal of Experimental Psychology, 65*(10), 1927–1944.

Simon, T. J., & Sussman, H. M. (1987). The dual task paradigm: Speech dominance or manual dominance? *Neuropsychologia, 25*(3), 559–569.

Smith, M., & Wheeldon, L. (1999). High level processing scope in spoken sentence production. *Cognition, 73*(3), 205–246.

Somberg, B. L., & Salthouse, T. A. (1982). Divided attention abilities in young and old adults. *Journal of Experimental Psychology: Human Perception and Performance, 8*(5), 651–663.

Spelke, E., Hirst, W., & Neisser, U. (1976). Skills of divided attention. *Cognition, 4*(3), 215–230.

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.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied, 9*(1), 23.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science, 12*(6), 462–466.

Strijkers, K., & Costa, A. (2011). Riding the lexical speedway: A critical review on the time course of lexical selection in speech production. *Frontiers in Psychology, 2*, 356.

Strobach, T., Liepelt, R., Pashler, H., Frensch, P. A., & Schubert, T. (2013). Effects of extensive dual-task practice on processing stages in simultaneous choice tasks. *Attention, Perception, & Psychophysics, 75*, 900–920.

Swets, B., Jacovina, M. E., & Gerrig, R. J. (2013). Effects of conversational pressures on speech planning. *Discourse Processes, 50*(1), 23–51.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.

Theeuwes, J. (1991). Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Attention, Perception, & Psychophysics, 49*(1), 83–90.

Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication, 52*, 201–212.

Van de Velde, M., Meyer, A. S., & Konopka, A. E. (2014). Message formulation and structural assembly: Describing "easy" and "hard" events with preferred and dispreferred syntactic structures. *Journal of Memory and Language, 71*(1), 124–144.

Wagner, V., Jescheniak, J. D., & Schriefers, H. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 423–440.

Watanabe, K., & Funahashi, S. (2014). Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature Neuroscience, 17*(4), 601–611.

Wickens, C. D. (1980). *The structure of attentional resources*. In R. Nickerson (Ed.). *Attention and performance VIII* (pp. 239–257). Hillsdale, NJ: Erlbaum.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review, 12*, 957–968.