Originally published in: Cognitive Science vol. 43 (2019) no. 7, 16 pp. DOI: *https://doi.org/10.1111/cogs.12768*

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2019) e12768

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12768

Speech Planning at Turn Transitions in Dialog Is Associated With Increased Processing Load

Mathias Barthel,[a] Sebastian Sauppe[b]

[a]*Language and Cognition Department, Max Planck Institute for Psycholinguistics*
[b]*Department of Comparative Linguistics, University of Zurich*

Received 25 September 2018; received in revised form 13 February 2019; accepted 21 May 2019

Abstract

Speech planning is a sophisticated process. In dialog, it regularly starts in overlap with an incoming turn by a conversation partner. We show that planning spoken responses in overlap with incoming turns is associated with higher processing load than planning in silence. In a dialogic experiment, participants took turns with a confederate describing lists of objects. The confederate's utterances (to which participants responded) were pre-recorded and varied in whether they ended in a verb or an object noun and whether this ending was predictable or not. We found that response planning in overlap with sentence-final verbs evokes larger task-evoked pupillary responses, while end predictability had no effect. This finding indicates that planning in overlap leads to higher processing load for next speakers in dialog and that next speakers do not proactively modulate the time course of their response planning based on their predictions of turn endings. The turn-taking system exerts pressure on the language processing system by pushing speakers to plan in overlap despite the ensuing increase in processing load.

Keywords: Turn taking; Dialog; Processing load; Task-evoked pupillary responses; Speech planning; Dual task

1. Introduction

Conversation is the most frequent form of human communication (Levinson, 2006), and taking turns at talk is a well-practiced task in which different speakers' contributions usually follow one another with only short gaps in between (Stivers et al., 2009). Planning a verbal response, however, is known to take between about 600 ms for single words (Indefrey, 2011; Strijkers & Costa, 2011) to well more than 1 second for short

Correspondence should be sent to Mathias Barthel, Language and Cognition Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. E-mail: mathias.barthel@mpi.nl

sentences (Griffin & Bock, 2000; Myachykov, Scheepers, Garrod, Thompson, & Fedorova, 2013), illustrating that timing a turn at talk in conversation is not a trivial task. To be able to quickly take their turn, next speakers need to start planning their response as early as possible, often in overlap with the incoming turn (Barthel, Meyer, & Levinson, 2017; Bögels, Magyari, & Levinson, 2015; Corps, Crossley, Gambi, & Pickering, 2018). Barthel, Sauppe, Levinson, and Meyer (2016) found that response planning was indeed done as early as the incoming turn's message could be conceived, even if the incoming turn did not end at that point.[1]

Planning the next turn while continuously monitoring the incoming turn for completion, and possibly for content, is a demanding dual-task situation. Both language comprehension and planning require allocation of central attention (Hagoort, Brown, & Osterhout, 1999; Kemper, Herman, & Lian, 2003; Kubose et al., 2006; Shitova, Roelofs, Coughler, & Schriefers, 2017), and both are known to interfere with concurrent non-linguistic tasks (Boiteau, Malone, Peters, & Almor, 2014; Roelofs & Piai, 2011; Sjerps & Meyer, 2015). The law of least mental effort proposes that humans try to make decisions and form strategies so as to minimize mental workload in order to achieve an efficient work–benefit ratio (Reichle, Carpenter, & Just, 2000; Zipf, 1949). It is thus a central question whether the language processing system is adapted to this highly frequent task or whether planning in overlap leads to increased processing load in the vicinity of turn transitions, the "crunch zone" of conversation (Roberts & Levinson, 2017). Using an auditory picture–word interference paradigm, Schriefers, Meyer, and Levelt (1990) compared the effects of concurrent noise versus concurrent speech on speech planning and found that naming latencies did not differ between a silent condition and a condition with distracting noise. With distracting words, however, naming latencies increased by 70 ms even when the words were unrelated to the picture names, indicating general interference of speech comprehension with speech planning. As participants were instructed to ignore any incoming speech and as their own utterances were independent of the presented speech input, the measured interference effects are effects of distraction rather than of the processes of integration of speech input, which is the task next speakers face in turn taking. Instead of trying to ignore incoming speech, interlocutors most of the time have to plan their next turn while concurrently listening to the incoming turn. Fargier and Laganaro (2016) studied picture naming performance with either a concurrent syllable or tone detection task and found longer response latencies and differences in ERP components in the syllable condition as compared to the tone condition, indicating increased interference between two concurrent linguistic tasks. Klaus, Mädebach, Oppermann, and Jescheniak (2017) made used of a dual-task paradigm combining sentence production as task 1 with a concurrent working memory as task 2. Participants were instructed to produce subject–verb–object sentences while they had to ignore auditory distractor words that were either phonologically or semantically related to either the subject or the object of the sentence. The concurrently performed working memory task was either visuospatial or verbal in nature. Under visuospatial load, both types of relatedness had effects on both the subject and object of the sentence. The pattern of results was similar under verbal load. Here, however, only phonological relatedness to the subject but not to the object affected sentence production performance, showing that verbal load reduced participants' phonological planning scope. These findings make it plausible to assume that next speakers postpone stages of formulation when planning in conversation in order to avoid inefficient processing due to interference. Barthel and Levinson (unpublished data), however, show that next speakers in a quiz-like situation engage in phonological planning as early as possible and in overlap with the incoming question. To date, evidence on the timing of the different processing stages in conversation is scarce, but the fact that response planning is frequently initiated in overlap with listening to the incoming turn is largely undisputed (but see Heldner & Edlund, 2010).

The observation that planning in overlap is common can be accounted for in two ways. One account highlights the mechanisms of turn allocation and the time pressure at turn transitions. According to the simplest systematics of turn taking (Sacks, Scheglo, & Jefferson, 1974), the first participant who speaks up when a turn transition becomes relevant gains the right to take the next turn. While language production and comprehension are assumed to engage—at least partly—the same cognitive resources (Hagoort & Indefrey, 2014; Kempen, Olsthoorn, & Sprenger, 2012; Menenti, Gierhan, Segaert, & Hagoort, 2011; Silbert, Honey, Simony, Poeppel, & Hasson, 2014) potentially increased processing load due to parallel processing of the two might be traded for the benefit of early planning, leading to shorter turn-transition times (Barthel et al., 2016, 2017). The alternative account questions the assumption that the simultaneity of comprehension and production in conversation drastically increases processing load. Previous research shows that participants prefer to use parallel processing over serial processing in dual tasks (Hübner & Lehle, 2007). To investigate the reasons for this tendency, Lehle, Steinhauser, and Hübner (2009) instructed participants explicitly to apply either a parallel or a serial processing strategy when giving parity judgments on two numbers. Lehle et al. (2009) found that while a parallel processing strategy increased reaction times and error rates, it decreased processing load, which might be the main reason for preferring parallel over serial processing. Consequently, planning in overlap might not be associated with any significant increase in processing load, especially since turn taking is a highly practiced dual task and cognitive tasks become less demanding with increasing proficiency (Donovan & Radosevich, 1999; Hampton Wray & Weber-Fox, 2013; Neubauer & Fink, 2009; Van Selst, Ruthru, & Johnston, 1999; Weber-Fox, Davis, & Cuadrado, 2003).

Here, we test whether planning a response while simultaneously comprehending an interlocutor's turn imposes increased processing load on speakers as compared to non-overlapping response planning by analyzing task-evoked pupillary responses from an experiment employing a dialogic paradigm. Changes in pupil diameter in response to task-induced cognitive processes are a reliable indicator of processing load (Beatty, 1982; Beatty & Lucero-Wagoner, 2000; Sirois & Brisson, 2014). The analysis of task-evoked pupillary responses allows studying differences in task demands, that is, the amount of overall cognitive resources that need to be allocated in order to master a task (Hess & Polt, 1964; Kahneman, 1973; Laeng, Sirois, & Gredebäck, 2012). Most studies using task-evoked pupillary responses to measure processing load in language processing have focused on comprehension (Engelhardt, Ferreira, & Patsenko, 2010; Just & Carpenter, 1993; Koch & Janse, 2016; Kuchinke, Vo, Hofmann, & Jacobs, 2007; Schmidtke, 2014; Tromp, Hagoort, & Meyer, 2016; Zekveld, Kramer, & Festen, 2010, inter alia), and there are only a few studies that have investigated language production (Papesh & Goldinger, 2012; Sauppe, 2017). If planning in overlap leads to increased processing load, task-evoked pupillary responses should have larger amplitudes as compared to planning in silence, whereas they are not predicted to differ if overlap does not increase processing load during response planning.

We report a dialogic experiment in which participants took turns with a confederate describing arrays of objects. Participants' pupil diameter was measured as they listened and responded to pre-recorded critical utterances from the confederate. These utterances were designed to, on one hand, either allow for response planning in overlap or not, and on the other hand, to contain either a predictable or a non-predictable ending. In this way, the effects of planning in overlap as compared to planning in silence on task-evoked pupillary responses were tested in the context of predictable and non-predictable overlapping speech input.

2. Methods and materials

2.1. Participants

Forty-eight German native speakers (M_{age} = 26.3 years, SD = 7.6 years, 30 female) who reported to have normal hearing and vision participated in the experiment for payment. Eight participants were excluded from the analyses because they reported during a post-test questionnaire that they had noticed the presence of pre-recorded material. Two participants were excluded due to technical failures of recording equipment, leaving 38 participants for analysis. Participants gave informed consent and the experiment was approved by the Ethics Committee of the Faculty of Social Sciences, Radboud University Nijmegen.

2.2. Apparatus

Participant and confederate were placed in separate sound-proof booths that were equipped with headphones and microphones with which they could communicate with one another. Visual stimuli were presented on a 21″ computer screen at a distance of approximately 60 cm. Participants' pupil size was recorded with an SMI RED-m remote eye tracker at 120 Hz sampling rate. Light conditions remained constant across participants.

2.3. Stimuli

2.3.1. Visual stimuli

Colored pictures of 468 objects were used to generate the visual stimuli. Ninety-six critical stimulus displays showing between three and five objects (32 displays each) were generated. Irrespective of the number of objects shown in an item display, each object filled approximately two degrees of visual angle and was located about 4 cm away from its neighbors, so that participants had to shift their gaze in order to foveally fixate individual objects. Between none and three of the objects had to be named by participants (24 displays each), and the remaining objects were named by the confederate (cf. Section 2.42.4).

2.3.2. Auditory stimuli

Each of the 96 critical stimulus displays was accompanied by a German sentence in one of the four conditions that were pre-recorded by the confederate and crossed according to whether the sentence ended in a verb or not (verb position) and whether it was predictable or not that the sentence would end with or without a final verb (end predictability; see Table 1). The presence of a sentence-final verb made planning in overlap possible, since all that participants needed to know to plan their response was which of the displayed objects they would have to name. When a sentence did not end in a verb, it ended in an object noun that was relevant for preparing the response, so that planning could only take place in silence after the turn ended. In predictable sentences, participants could know in advance whether the last word would be a verb or an object noun, since different verbs in second position (before the list of objects) either required another verb form in sentence-final position (such as the modal verb “can”) or not (such as the main verb “see”). In contrast, non-predictable sentences contained “have” in second position, which is ambiguous between being a main verb or an auxiliary and consequently either does or does not call for a sentence-final participle. Four pseudo-randomized lists were constructed, so that each item appeared in only one condition per list and the same number of items per condition appeared in each list.

2.4. Procedure

Prior to the experiment, participants were shown all objects in a booklet and asked to name them. Participants and the confederate were instructed as follows. In each trial, they would see a number of objects they could get and the confederate should tell the

Table 1
Example sentences of the four conditions used in the experiment. “I have/have gotten/see/can get a key, a kite, and a ruby.”

	End Predictability	
Verb Position	Unpredictable	Predictable
Not final	Ich habe einen Schlüssel, einen Lenkdrachen und einen Rubin	Ich sehe einen Schlüssel, einen Lenkdrachen und einen Rubin
Final	Ich habe einen Schlüssel, einen Lenkdrachen und einen Rubin besorgt	Ich kann einen Schlüssel, einen Lenkdrachen und einen Rubin besorgen

participant what objects she could get, so that the participant could tell the confederate what further objects he could get, only listing the objects that had not already been named by the confederate (all objects named by the confederate were also visible on the participant's display). Participants triggered the beginning of each trial by looking at a fixation cross at the center of the screen. Each trial began with a preview of 600–1,000 ms of the stimulus display before the critical sentence was played. The experiment started with 12 practice trials that were of the same structure as experimental trials. The eye tracker was (re-)calibrated four times at equal intervals during the experiment. The experiment lasted approximately 30 min and was followed by a computerized questionnaire asking participants whether they had noted the presence of pre-recorded material.

2.5. Data preprocessing and analyses

Preprocessing of pupil data and statistical analyses were carried out in R (R Core Team, 2018). Samples recorded with low validity (as indicated by SMI's recording software) and during blinks or saccades were treated as missing values and linearly interpolated separately for each eye. Pupil diameters of both eyes were averaged before time-locking to the offset of the last noun in the confederate turn. For each trial, pupil diameter was baselined by subtracting the mean diameter during a baseline period spanning the 500 ms preceding the offset of the last noun in the confederate turn. The mean task-evoked pupillary response amplitude was calculated for a time window of 3,000 ms after the time-lock point and peaks in pupil diameters were identified in this time window (Borchers, 2015).

The data set contained 2,736 trials in which both confederate and participant named at least one object. Trials in which participants did not name the correct objects or responded in overlap and trials with more than 30% missing values before interpolation in samples recorded between −500 and 3,000 ms relative to the offset of the confederate's last noun were excluded from statistical analyses (319 trials). Forty-three additional trials were excluded because their verbal response time was more than 3 SD longer than the participant's mean response time measured manually in Praat (Boersma & Weenink, 2015) from the offset of the incoming turn to the onset of the first object noun in the participants' turn. Additional items in which sentences were produced live by the confederate (see Barthel et al., 2016) were not considered for analyses (341 trials). On balance, 2,377 trials remained for analysis (13.12% of trials were excluded).

Three linear mixed effects regression models were fitted (Bates, Mächler, Bolker, & Walker, 2015) with the mean amplitude, peak amplitude, and peak latency as dependent variables. The underlying assumption is that differences in the mean and peak amplitudes, and peak latency relate to differences in processing load and reflect differences in task difficulty (Beatty & Lucero-Wagoner, 2000). While the peak amplitude is a good measure for processing load, accurate peak detection is not straightforward, as the location of peaks is susceptible to noise in the recorded signal (Luck, 2014). The mean amplitude is a more conservative measure for processing load, since it takes into account the whole analysis window and is thus less susceptible to noise. Differences in the latency of peaks between conditions relate to differences in task difficulty, reflecting differences in the time it takes to do the necessary computations in order to give a response. Converging results in these measures is desirable when drawing inferences on cognitive demand on the basis of task-evoked pupillary responses. Verb position and sentence end predictability as well as their interaction were the predictors of interest. Their statistical significance was assessed using F-tests with Kenward–Roger approximations of degrees of freedom (Fox & Weisberg, 2011; Halekoh & Hojsgaard, 2014; Kenward & Roger, 1997). The maximal random effects structures as justified by design which allowed models to converge were used (Barr, 2013; Barr, Levy, Scheepers, & Tily, 2013). A number of nuisance variables were included in the fixed effects structure of the models (Sassenhagen & Alday, 2016): the duration of the confederate turn, since the pre-recorded sentences differed in complexity; the number of objects to be named by the participant, since task difficulty increases with the number of choices (Hick, 1952); trial number, to account for changes over the course of the experiment; and a binary variable indicating whether the sentence structure of the confederate turn was reused in the response turn, since processing load might be influenced by structural priming (Pickering & Ferreira, 2008; Segaert, Menenti, Weber, Petersson, & Hagoort, 2012). The statistical significance of nuisance variables was not assessed. Categorical predictors were deviation coded (−0.5 and 0.5) and continuous predictors were mean centered.

3. Results

Average task-evoked pupillary responses are shown in Fig. 1 and descriptive statistics are presented in Table 2.

Linear mixed effects regressions revealed that task-evoked pupillary responses in the verb-final conditions had statistically significantly higher mean amplitudes, higher peak amplitudes, and greater peak latencies than in non-final conditions. Neither the main effect of predictability nor its interaction with verb position reached statistical significance in any of the three models (Table 3).

4. Discussion

We investigated the level of processing load in next speakers in the vicinity of turn transitions in dialog to answer the question whether planning a turn at talk in overlap with the incoming turn leads to higher processing load than planning it in silence. Task-evoked pupillary responses recorded during a dialogic list-completion task were analyzed, and mean amplitudes and peak amplitudes were found to be higher and peak latencies to be longer when planning was done in overlap than when it was done in silence. While the sentences in conditions that allowed for early planning in overlap were often slightly more complex than the sentences in conditions that did not allow for early planning, the differences in sentence complexity were much greater within than between conditions.

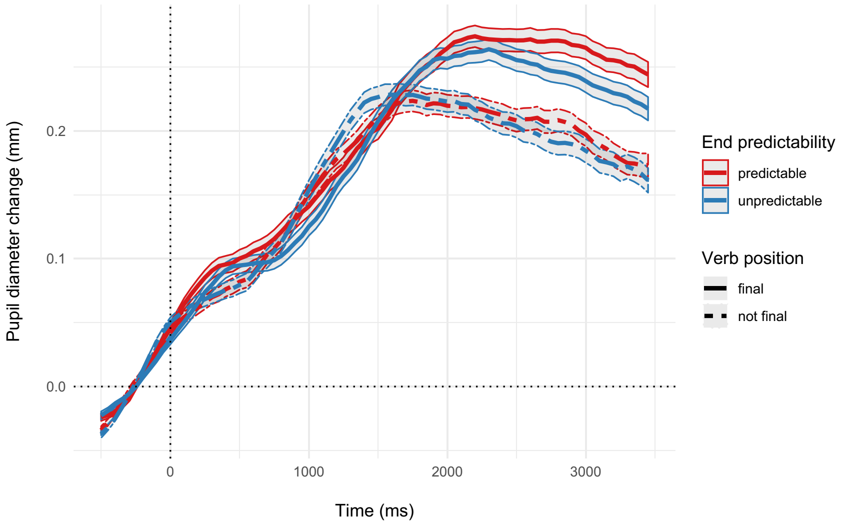

Fig. 1. Grand average changes in pupil diameter (task-evoked pupillary responses) in millimeters, time-locked to the offset of the last noun of the confederate's turn (dashed vertical line). Ribbons indicate 95% confidence intervals. The analysis time window ranged from 0 to 3,000 ms. For plotting only, samples were averaged into 50 ms bins within each trial to align time steps across trials before grand averaging.

Table 2
Means (standard deviations in parentheses) of peak and mean amplitudes, and peak latencies by condition

Condition	Peak Amplitude in mm	Mean Amplitude in mm	Peak Latency in ms
No final verb/unpredictable	0.409 (0.231)	0.167 (0.198)	1,850 (782)
No final verb/predictable	0.412 (0.222)	0.165 (0.185)	1,868 (825)
Final verb/unpredictable	0.432 (0.226)	0.179 (0.188)	2,030 (756)
Final verb/predictable	0.446 (0.241)	0.190 (0.196)	2,041 (782)

Whether a sentence ended in a verb or not influenced pupillary responses beyond the influence of sentence duration, which was included as a nuisance variable to account for the length of a sentence and thereby its complexity. Taken together, the presented results show that planning in overlap is more demanding than planning in silence.

In their analyses of eye movements from the experiment here, Barthel et al. (2016) found that participants started to plan their response as early as possible, that is, as soon as they had identified the last noun of the incoming turn irrespective of another verb form following before the end of the turn or not. Consequently, participants generally started planning their response in overlap with the incoming turn in verb-final conditions and in silence in conditions without a final verb. When planning in overlap, the time gained by starting to plan early was not fully reflected in the reduction of turn-transition times. When participants planned their response in overlap, planning overlapped with turn-final

Table 3
Linear mixed effects regression models predicting mean task-evoked pupillary response amplitude (in mm), peak task-evoked pupillary response amplitude (in mm), and peak task-evoked pupillary response latency (in ms). Statistical significance based on Type II *F*-tests with Kenward–Roger degrees of freedom (Kenward & Roger, 1997)

	Mean Amplitude (mm)				Peak Amplitude (mm)				Peak Latency (logarithm of ms)			
	$\hat{\beta}$	$\|t\|$	F	p	$\hat{\beta}$	$\|t\|$	F	p	$\hat{\beta}$	$\|t\|$	F	p
Intercept	0.175	12.366			0.425	21.313			7.397	182.183		
Verb position (= final)	0.023	2.788	7.847	0.006**	0.034	3.647	13.327	<0.001***	0.114	3.507	12.323	<0.001***
End predictability (= predictable)	0.006	0.789	0.617	0.438	0.009	1.182	1.393	0.238	−0.012	0.374	0.140	0.708
Verb position × End predictability	0.009	0.674	0.453	0.501	0.004	0.268	0.072	0.789	0.012	0.185	0.034	0.853
Structural priming (= yes)	0.009	1.116			0.016	1.754			−0.011	0.317		
Sentence duration (*z*)	−0.002	0.235			<0.001	0.022			0.081	2.716		
Trial number (*z*)	−0.015	4.270			−0.016	3.858			−0.060	3.844		
Delta of objects (*z*)	0.007	0.700			0.013	1.191			0.099	3.329		

**$p < .01$; ***$p < .001$.

verbs which were about 600 ms long. In these cases, however, gaps between turns were shorter by only approximately 100 ms. This means that participants spent considerably more time planning their response when planning started in overlap than when planning was done in silence. The reported pattern of task-evoked pupillary responses sheds light on the cause of this discrepancy: The increase in planning time was due to higher processing load in planning in overlap as compared to planning in silence.

Given that planning in overlap is the norm in conversation, the finding that it is a more demanding strategy as compared to planning in silence shows that the requirements of the systematics of turn taking in conversation (Sacks et al., 1974) receive precedence over the minimization of mental effort. The culturally developed turn-taking system exerts pressure on the cognitive mechanisms of language processing, enforcing strategies that raise processing load in order to meet the requirements set by the rules of turn allocation and the semiotics of turn timing. Increased processing load for the sake of finely attuned temporal alignment of turns thus appears to be a cornerstone in the organization of turn allocation: If you want to take a turn at talk, you need to push your language processor in order to speak up before other participants. Trading high processing load for shorter turn transitions is a prerequisite for the timing of turns to become a meaningful source of information. If the next speaker does not claim her turn in time, she can be interpreted as lacking interest in the conversation, its topic, or her interlocutor, as having trouble understanding the previous turn or parts of it (Kendrick, 2015; Scheglo, Jefferson, & Sacks, 1977), as being unwilling to comply with a request or as preparing to disagree with an assessment (Kendrick & Torreira, 2014; Roberts & Francis, 2013; Roberts, Margutti, & Takano, 2011). In that way, turn timing is meaningful in itself, irrespective of the content of the following turn, with a long gap before a turn leading the recipient to expect a dispreferred response, for example, a rejection of an invitation (Bögels, Kendrick, & Levinson, 2015). With the timing of turn taking being a source of information that is analyzed by listeners, more information can be inferred from a single unit of talk. This enriches social interaction in conversation but comes at the cost of increased processing load for the individual speaker.

As processing load is high at turn transitions due to time pressures, next speakers might develop strategies to distribute processing load evenly over time when planning their turn. Based on findings that participants in dual tasks can to some degree choose to apply different processing strategies (Hübner & Lehle, 2007; Miller, Ulrich, & Rolke, 2009; Navon & Gopher, 1979; Navon & Miller, 2002; Tombu & Jolicœur, 2005), one conceivable way to avoid high peaks in processing load would be to apply a “proactive planning” strategy in cases when incoming turns contain highly predictable turn-final words. If predictability of a turn-final word leads to effective changes in response planning, processing load in sentences with predictable turn ends should be lower than in sentences with unpredictable turn ends. However, none of the analyzed pupillary response measures (peak amplitude, mean amplitude, and peak latency) were significantly affected by predictability, lending no support to the hypothesis that participants applied a proactive planning strategy in order to keep processing load low at turn transitions. We take this as evidence that next speakers did not utilize the predictability of incoming verbal material to adapt the time course of their response planning (cf. also Huettig & Mani, 2016). In order to meet turn timing requirements, next speakers seem to aim to plan their contribution as early and fast as possible, accepting increased processing loads during response planning to avoid risking the consequences of being too slow to take their turn.

By planning their response in overlap with comprehending the incoming turn, participants' behavior agrees with the general tendency to choose parallel processing over serial processing in dual tasks (Hübner & Lehle, 2007); they do not postpone encoding processes until after a predictable final word. In our experiment, however, the reason for this choice cannot have been reduced processing load, as our analyses of task-evoked pupillary responses show that planning a response in overlap induces higher processing load than planning in silence. Instead, participants' motivation was more likely to reduce the length of gap after the incoming turn. Intending to take a well-timed turn, next, speakers employed a planning strategy that at the same time took them longer to plan their response and was more demanding as compared to delaying response planning. While it remains possible that the choice of processing strategy is a question of preference of individual speakers (Bögels, Casillas, & Levinson, 2018) or the demands of the dual-task situation (Lehle & Hübner, 2009; Reissland & Manzey, 2016), parallel processing appears to be the standard strategy in dialog.

In sum, the turn-taking system requires next speakers to accept higher processing loads induced by planning in overlap in order to be able to respond as fast as possible to an incoming turn so as to avoid the social consequences ensuing from noticeable gaps between turns of talk. In the words of Kahneman (1973), participants in a conversation are forced to trade *efficiency* in terms of processing load for *effectiveness* in terms of short gaps between turns. This means that the turn-taking system is not optimized for next speakers' processing, but for overall effectiveness in social interaction. While putting pressure on cognitive processing in individual speakers, the turn-taking system allows for a dense semiotics of turn timing that organizes and enriches social interaction in conversation. In addition to viewing the turn-taking system as shaping the evolution of aspects of grammar (Auer, 2005; Ford & Thompson, 2003; Roberts & Levinson, 2017), the need to meet the timing demands in turn taking might also be shaping the design of the cognitive system. The study presented in this paper shows that examining task-evoked pupillary responses during speech planning is a promising technique to further investigate the mechanisms of speech processing in conversation.

Acknowledgments

We thank Freya Materne for her work as our confederate, Ronald Fisher and Tanja Marton for their technical support, and Antje S. Meyer and three anonymous reviewers for their comments on an earlier version of this manuscript. This research was funded by the Max Planck Society for the Advancement of Science (through the Max Planck Institute for Psycholinguistics, Nijmegen).

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data availability statement

Raw data and analysis scripts are available from https://osf.io/pf2br/.

Note

1. The study reported here presents pupillometric data from Barthel et al. (2016), which focused on eye movements.

References

Auer, P. (2005). Projection in interaction and projection in grammar. *Text - Interdisciplinary Journal for the Study of Discourse*, *25*(1), 7–36. https://doi.org/10.1515/text.2005.25.1.7

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*(328), 1–2. https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final 'go-signals'. *Frontiers in Psychology*, *8*, 393. https://doi.org/10.3389/fpsyg.2017.00393

Barthel, M., Sauppe, S., Levinson, S. C., & Meyer, A. S. (2016). The timing of utterance planning in task-oriented dialogue: Evidence from a novel list-completion paradigm. *Frontiers in Psychology*, *7*(1858). https://doi.org/10.3389/fpsyg.2016.01858

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson(Eds.), *Handbook of psychophysiology* (2nd ed., pp. 142–162). Cambridge, UK: Cambridge University Press.

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer [Computer program]. Version 5.3.56. Retrieved from http://www.praat.org/. Accessed Jan, 2019.

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

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never say no... How the brain interprets the pregnant pause in conversation. *PLoS ONE*, *10*(12), e0145474. https://doi.org/10.1371/journal.pone.0145474

Bögels, S., Magyari, L., & Levinson, S. C. (2015). Neural signatures of response planning occur midway through an incoming question in conversation. *Scientific Reports*, *5*(12881), 1–11. https://doi.org/10.1038/srep12881

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

Borchers, H. W. (2015). pracma: Practical Numerical Math Functions [R package]. Version 2.1.4. Retrieved from www.cran.r-project.org. Accessed Jan, 2019.

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

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. https://doi.org/10.1037/0021-9010.84.5.795

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *Quarterly Journal of Experimental Psychology*, *63*(4), 639–645. https://doi.org/10.1080/17470210903469864

Fargier, R., & Laganaro, M. (2016). Neurophysiological modulations of non-verbal and verbal dual-tasks interference during word planning. *PLoS ONE*, *11*(12), e0168358. https://doi.org/10.1371/journal.pone.0168358

Ford, C. E., & Thompson, S. A. (2003). Social interaction and grammar. In M. Tomasello(Ed.), *The new psychology of language*(Vol. 2). Mahwah, NJ: Lawrence Erlbaum.

Fox, J., & Weisberg, S.(2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

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

Hagoort, P., Brown, C. M., & Osterhout, L. (1999). The neurocognition of syntactic processing. In C. M. Brown & P. Hagoort (Eds.), *Neurocognition of language* (pp. 273–361). Oxford, UK: Oxford University Press.

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual Review of Neuroscience*, *37*(1), 347–362. https://doi.org/10.1146/annurev-neuro-071013-013847

Halekoh, U., & Hojsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models—The R package pbkrtest. *Journal of Statistical Software*, *59*(9), 1–30.

Hampton Wray, A., & Weber-Fox, C. (2013). Specific aspects of cognitive and language proficiency account for variability in neural indices of semantic and syntactic processing in children. *Developmental Cognitive Neuroscience*, *5*, 149–171. https://doi.org/10.1016/j.dcn.2013.03.002

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

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26. https://doi.org/10.1080/17470215208416600

Hübner, R., & Lehle, C. (2007). Strategies of flanker coprocessing in single and dual tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 103–123. https://doi.org/10.1037/0096-1523.33.1.103

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31. https://doi.org/10.1080/23273798.2015.1072223

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 255. https://doi.org/10.3389/fpsyg.2011.00255

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, *47*(2), 310–339. https://doi.org/10.1037/h0078820

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kempen, G., Olsthoorn, N., & Sprenger, S. (2012). Grammatical workspace sharing during language production and language comprehension: Evidence from grammatical multitasking. *Language and Cognitive Processes*, *27*(3), 345–380. https://doi.org/10.1080/01690965.2010.544583

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 or noise. *Psychology and Aging*, *18*(2), 181–192. https://doi.org/10.1037/0882-7974.18.2.181

Kendrick, K. H. (2015). The intersection of turn-taking and repair: The timing of other-initiations of repair in conversation. *Frontiers in Psychology*, *6*, 250. https://doi.org/10.3389/fpsyg.2015.00250

Kendrick, K. H., & Torreira, F. (2014). The timing and construction of preference: A quantitative study. *Discourse Processes*, *52*(4), 1–35. https://doi.org/10.1080/0163853X.2014.955997

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558

Klaus, J., Mädebach, A., Oppermann, F., & Jescheniak, J. D. (2017). Planning sentences while doing other things at the same time: Effects of concurrent verbal and visuospatial working memory load. *Quarterly Journal of Experimental Psychology*, *70*(4), 811–831. https://doi.org/10.1080/17470218.2016.1167926

Koch, X., & Janse, E. (2016). Speech rate effects on the processing of conversational speech across the adult life span. *The Journal of the Acoustical Society of America*, *139*(4), 1618–1636. https://doi.org/10.1121/1.4944032

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

Kuchinke, L., Vo, M., Hofmann, M., & Jacobs, A. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*(2), 132–140. https://doi.org/10.1016/j.ijpsycho.2007.04.004

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Lehle, C., & Hübner, R. (2009). Strategic capacity sharing between two tasks: Evidence from tasks with the same and with different task sets. *Psychological Research Psychologische Forschung*, *73*(5), 707–726. https://doi.org/10.1007/s00426-008-0162-6

Lehle, C., Steinhauser, M., & Hübner, R. (2009). Serial or parallel processing in dual tasks: What is more effortful? *Psychophysiology*, *46*(3), 502–509. https://doi.org/10.1111/j.1469-8986.2009.00806.x

Levinson, S. C. (2006). On the human "interaction engine." In N. Enfield & S. C. Levinson(Eds.), *Roots of human sociality: Culture, cognition and interaction*(pp. 39–69). Oxford, UK: Berg.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: The MIT Press.

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological Science*, *22*(9), 1173–1182. https://doi.org/10.1177/0956797611418347

Miller, J., Ulrich, R., & Rolke, B. (2009). On the optimality of serial and parallel processing in the psychological refractory period paradigm: Effects of the distribution of stimulus onset asynchronies. *Cognitive Psychology*, *58*(3), 273–310. https://doi.org/10.1016/j.cogpsych.2006.08.003

Myachykov, A., Scheepers, C., Garrod, S., Thompson, D., & Fedorova, O. (2013). Syntactic flexibility and competition in sentence production: The case of English and Russian. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1601–1619. https://doi.org/10.1080/17470218.2012.754910

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, *86*(3), 214–255. https://doi.org/10.1037/0033-295X.86.3.214

Navon, D., & Miller, J. (2002). Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive Psychology*, *44*(3), 193–251. https://doi.org/10.1006/cogp.2001.0767

Neubauer, A. C., & Fink, A. (2009). Intelligence and neural efficiency. *Neuroscience & Biobehavioral Reviews*, *33*(7), 1004–1023. https://doi.org/10.1016/j.neubiorev.2009.04.001

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. https://doi.org/10.3758/s13414-011-0263-y

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459. https://doi.org/10.1037/0033-2909.134.3.427

R Core Team. (2018). *R: A language and environment for statistical computing [Computer software manual]*. Vienna: Austria. Retrieved from https://www.R-project.org/. Accessed Jan, 2019.

Reichle, E. D., Carpenter, P. A., & Just, M. A. (2000). The neural bases of strategy and skill in sentence-picture verification. *Cognitive Psychology*, *40*(4), 261–295. https://doi.org/10.1006/cogp.2000.0733

Reissland, J., & Manzey, D. (2016). Serial or overlapping processing in multitasking as individual preference: Effects of stimulus preview on task switching and concurrent dual-task performance. *Acta Psychologica*, *168*, 27–40. https://doi.org/10.1016/j.actpsy.2016.04.010

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. https://doi.org/10.1121/1.4802900

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. https://doi.org/10.1080/0163853X.2011.558002

Roberts, S. G., & Levinson, S. C. (2017). Conversation, cognition and cultural evolution: A model of the cultural evolution of word order through pressures imposed from turn taking in conversation. *Interaction Studies*, *18*(3), 402–442. https://doi.org/10.1075/is.18.3.06rob

Roelofs, A., & Piai, V. (2011). Attention demands of spoken word planning: A review. *Frontiers in Psychology*, *2*, https://doi.org/10.3389/fpsyg.2011.00307

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

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language*, *162*, 42–45. https://doi.org/10.1016/j.bandl.2016.08.001

Sauppe, S. (2017). Symmetrical and asymmetrical voice systems and processing load: Pupillometric evidence from sentence production in Tagalog and German. *Language*, *93*(2), 288–313. https://doi.org/10.1353/lan.2017.0015

Scheglo, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382. https://doi.org/10.2307/413107

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, *5*, 137. https://doi.org/10.3389/fpsyg.2014.00137

Schriefers, H., Meyer, A. S., & Levelt, W. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*, 86–102. https://doi.org/10.1016/0749-596X(90)90011-N

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension—An fMRI study. *Cerebral Cortex*, *22*(7), 1662–1670. https://doi.org/10.1093/cercor/bhr249

Shitova, N., Roelofs, A., Coughler, C., & Schriefers, H. (2017). P3 event-related brain potential reflects allocation and use of central processing capacity in language production. *Neuropsychologia*, *106*, 138–145. https://doi.org/10.1016/j.neuropsychologia.2017.09.024

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43), E4687–E4696. https://doi.org/10.1073/pnas.1323812111

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(6), 679–692. https://doi.org/10.1002/wcs.1323

Sjerps, M. J., & Meyer, A. S. (2015). Variation in dual-task performance reveals late initiation of speech planning in turn-taking. *Cognition*, *136*, 304–324. https://doi.org/10.1016/j.cognition.2014.10.008

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

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), 1–16. https://doi.org/10.3389/fpsyg.2011.00356

Tombu, M., & Jolicœur, P. (2005). Testing the predictions of the central capacity sharing model. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(4), 790–802. https://doi.org/10.1037/0096-1523.31.4.790

Tromp, J., Hagoort, P., & Meyer, A. S. (2016). Pupillometry reveals increased pupil size during indirect request comprehension. *Quarterly Journal of Experimental Psychology*, *69*(6), 1093–1108. https://doi.org/10.1080/17470218.2015.1065282

Van Selst, M., Ruthru, E., & Johnston, J. C. (1999). Can practice eliminate the Psychological Refractory Period effect? *Journal of Experimental Psychology: Human Perception and Performance*, *25*(5), 1268–1283. https://doi.org/10.1037/0096-1523.25.5.1268

Weber-Fox, C., Davis, L. J., & Cuadrado, E. (2003). Event-related brain potential markers of high-language proficiency in adults. *Brain and Language*, *85*(2), 231–244. https://doi.org/10.1016/S0093-934X(02)00587-4

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251S

Zipf, G. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, UK: Addison-Wesley Press.