Cognition 175 (2018) 77–95

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Early preparation during turn-taking: Listeners use content predictions to determine *what* to say but not *when* to say it☆

Ruth E. Corps[a,*], Abigail Crossley[a], Chiara Gambi[a,b], Martin J. Pickering[a]

[a] Department of Psychology, University of Edinburgh, United Kingdom
[b] School of Psychology, Cardiff University, United Kingdom

ARTICLE INFO

Keywords:
Prediction
Response preparation
Turn-taking
Question-answering
Conversation
Dialogue

ABSTRACT

During conversation, there is often little gap between interlocutors' utterances. In two pairs of experiments, we manipulated the content predictability of *yes/no* questions to investigate whether listeners achieve such coordination by (i) preparing a response as early as possible or (ii) predicting the end of the speaker's turn. To assess these two mechanisms, we varied the participants' task: They either pressed a button when they thought the question was about to end (Experiments 1a and 2a), or verbally answered the questions with either *yes* or *no* (Experiments 1b and 2b). Predictability effects were present when participants had to prepare a verbal response, but not when they had to predict the turn-end. These findings suggest content prediction facilitates turn-taking because it allows listeners to prepare their own response early, rather than because it helps them predict when the speaker will reach the end of their turn.

1. Introduction

Speaking and listening to speech are both extremely complex processes. Yet, during conversationinterlocutors are able to switch from one to the other exactly when they need to. In fact, speakers rarely overlap extensively, and the gap between their turns typically averages 200 ms (Stivers et al., 2009). To achieve such coordination, listeners must prepare their own response and articulate it at the appropriate moment. But how do they do so?

Current theories agree that interlocutors achieve such coordination in part by predicting the content of the speaker's incoming turn (i.e., what the speaker is likely to say next; e.g., Bögels & Levinson, 2017; Garrod & Pickering, 2015). Indeed, we know that comprehenders can predict upcoming language at different linguistic levels, including semantic, syntactic, and form-related information (e.g., Altmann & Kamide, 1999; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). However, it is currently unclear how these content predictions aid successful turn-taking.

Such predictions may ease processing of the incoming turn, allowing listeners to prepare an appropriate response (e.g., one which is semantically and syntactically appropriate) in good time, and thus respond earlier. But on its own, early preparation may not be sufficient for smooth turn-taking: Listeners must also articulate their response at the appropriate moment, so they do not overlap with the previous speaker nor leave a long gap. Content predictions may help listeners predict when the speaker's turn will end (see Corps, Gambi, & Pickering, 2018), so they can time their responses more precisely (i.e., clustered closer to the turn-end).

In principle, content predictions might support smooth turn-taking both by facilitating earlier response preparation and by allowing more precise turn-end prediction. Crucially, however, it is currently unclear how the process of determining what to say relates to the process of determining when to speak. One possibility is that listeners use content predictions to prepare a response early, hold this response in an articulatory buffer, and then launch articulation reactively when the speaker displays turn-final cues (e.g., drawl on the final syllable; Duncan, 1972). We term this the *early-planning hypothesis* (e.g., Levinson & Torreira, 2015), as it proposes that listeners determine what to say early, separately from determining when to say it. According to this hypothesis, content predictability facilitates turn-taking because listeners can prepare a response earlier when the content of the speaker's turn is more rather than less predictable. This account predicts that there is no role for prediction of the speaker's turn end because listeners use turn-final cues to determine when to speak, and so content predictability should only benefit the process of determining what to say and not the process of determining when to say it.

But turn-final cues are far from perfect predictors of a turn change (e.g., Gravano & Hirschberg, 2011). In addition, using production

☆ This research has been presented at a poster session at the 22nd Architectures and Mechanisms For Language Processing conference.

* Corresponding author at: Department of Psychology, 7 George Square, University of Edinburgh, Edinburgh EH8 9JZ, United Kingdom.
E-mail address: rcorps@exseed.ed.ac.uk (R.E. Corps).

https://doi.org/10.1016/j.cognition.2018.01.015
Received 29 August 2017; Received in revised form 26 January 2018; Accepted 29 January 2018
Available online 16 March 2018

processes to prepare and buffer a response is cognitively demanding and may interfere with the listener's ability to comprehend the speaker's unfolding utterance. Importantly, listeners could avoid such interference by beginning preparation only when they believe that they will soon have the opportunity to articulate their response (i.e., late in the turn; Sjerps & Meyer, 2015). According to this *late-planning hypothesis*, listeners use content predictions to predict the speaker's turn-end and only begin response preparation close to this moment (cf. Bögels & Levinson, 2017). If this is the case, then listeners should be more precise at predicting the speaker's turn-end when content is more rather than less predictable.

Note that although we present two opposing accounts in line with the literature, they are not necessarily mutually exclusive. The two mechanisms could work in parallel to some extent (see Bögels & Levinson, 2017). For example, listeners could use content prediction to prepare a response early and also to predict the speaker's turn-end in order to better time response articulation, in a way that would combine elements of both the early planning and the late planning account. Conversely, listeners may prepare late and also use turn-final cues (rather than turn-end prediction) to time articulation. However, it is an empirical question whether predictability affects only response preparation (early-planning), only turn-end prediction (late-planning), or indeed both.

To explore the role of predictability during turn-taking, we manipulated the content predictability of simple *yes–no* questions in two pairs of experiments, using two paradigms designed to capture different aspects of the turn-taking process. To isolate turn-end prediction, we first used a button-press task, in which listeners pressed a button as soon as they expected the speaker to reach the end of their turn (i.e., they were encouraged to predict this moment; De Ruiter, Mitterer, & Enfield, 2006). Since this paradigm encourages participants to precisely time their response, we analyzed absolute response precision (i.e., how close participants responded to the speaker's turn-end). While the early-planning hypothesis does not predict any difference in precision between predictable and unpredictable questions (because it assumes no role for turn-end prediction), the late-planning hypothesis predicts that listeners should be more precise (i.e., their responses should cluster closer to the speaker's turn-end) when they can predict question content than when they cannot.

To further explore the role of content predictability, we conducted two additional experiments using a question-answering task, which we assume captures response preparation in addition to turn-end prediction. Accordingly, we analyzed not only the precision of participants' responses (as in the button-press task), but also the signed response times (i.e., how early participants responded). Precision and response times are of course related measures but, crucially, changes in response times can influence response precision in different ways: If participants are slower to respond, their responses can become either less precise (if they occur after the end of the speaker's turn) or more precise (if they occur before the end of the speaker's turn). Moreover, changes in precision can occur independently of changes in response time (e.g., if the spread of responses increases without changes to the mean response time).

Thus, it is necessary to analyze both measures to determine whether content predictability affects precision (i.e., as predicted by the late-planning hypothesis) and whether it affects response timing (i.e., as predicted by the early-planning hypothesis). Early-planning proposes that listeners should respond earlier when they can predict question content than when they cannot (because content prediction helps listeners prepare earlier), but does not predict any difference in precision between predictable and unpredictable questions (because articulation is timed based on a different mechanism, namely reaction to turn-final cues). In contrast, the late-planning hypothesis proposes that responses should be more precise for predictable than unpredictable questions (because prediction helps listeners determine the turn-end more accurately), but does not predict any difference in signed response times between predictable and unpredictable questions (because listeners always begin preparation close to the turn end anyway).

We used the same items in both tasks to ensure comparability between the experiments. In the rest of the Introduction, we discuss evidence for and against both accounts, before describing the current study and formulating our predictions in more detail. We also distinguish two versions of the late-planning account that differ in what information they assume is used for turn-end prediction.

1.1. Evidence for early planning

Some research suggests that listeners prepare their own turns as early as possible. For example, in a question-answering task Bögels, Magyari, and Levinson (2015) found that participants responded earlier and showed activation in brain areas involved in speech production (e.g., Indefrey & Levelt, 2004) and motor response preparation (e.g., Babiloni et al., 1999) when the information (here, *007*) necessary for response preparation was available early in the turn (e.g., *Which character, also known as 007, appears in the famous movies*?) rather than late (e.g., *Which character from the famous movies is also called 007*?). These results suggest participants prepared their response further in advance when the critical information was available early rather than late. Importantly, they did so even though the question could have continued in a number of different ways (e.g., *appeared in Skyfall?, was recently played by Daniel Craig?*), meaning they could not necessarily predict the turn-end.

Barthel, Sauppe, Levinson, and Meyer (2016) provided further support for the early-planning account using a list-completion task, in which participants completed a confederate's pre-recorded utterances. Participants had to name any on-screen objects that the confederate had not already named, and so they could (in principle) prepare their response as soon as the confederate began uttering the last object name. The authors also manipulated whether participants could predict that the speaker's turn would end with a turn-final verb. Both eye-movements and response latencies suggested that participants planned their response as soon as possible. However, neither of these measures were influenced by the predictability of the speaker's turn-end, suggesting that listeners did not use such predictions to time response articulation. Participants may instead have launched articulation using turn-final cues (see Barthel, Meyer, & Levinson, 2017).

1.2. Problems with early planning

Although the evidence in Section 1.1 supports the early-planning hypothesis, this account faces two unresolved issues. First, it is unclear whether turn-final cues can explain all turn-taking behavior. In a corpus study of dyadic interactions, Gravano and Hirschberg (2011) assessed the role of seven turn-final cues (e.g., lengthening of the final word) and found that these cues were significantly more likely to occur in stretches of speech preceding speaker changes than in those preceding a continuation of the current speaker's turn. However, listeners were only 65% likely to take a turn when all seven cues were present. Although one of the cues considered by the authors was whether the turn was semantically and/or syntactically complete, they did not explore the role of content predictability, thus leaving open the possibility that other content-based mechanisms (such as turn-end prediction) are also at play.

Second, if addressees prepare their response as soon as possible, then production and comprehension processes must overlap. Since these processes recruit overlapping neural circuits (e.g., Segaert, Menenti, Weber, Petersson, & Hagoort, 2012) and most likely share resources, using production mechanisms to prepare and buffer a response in advance of the turn-end should be cognitively demanding and may interfere with the concurrent process of comprehending the speaker's turn. Indeed, previous research suggests all stages of preparation (e.g., lemma, word form, and phoneme selection; Cook &

Table 1
Example materials and possible completions for each of the four stimuli conditions.

Content predictability	Length predictability	Example question fragment	Possible completions
Predictable	Single	Are dogs your favourite…?	Animal
	Varied	Did The Titanic sink after…?	It hit an iceberg/hitting an Iceberg/crashing
Unpredictable	Single	Do you enjoy going to the…?	Supermarket/dentist/beach
	Varied	Do most students finish their…?	Dinner/studies after four years/exams on time

Meyer, 2008) require central processing capacity.

Crucially, listeners could avoid such interference by preparing a response only when they are sure the speaker is about to finish (i.e., late-planning hypothesis). Sjerps and Meyer (2015; see also Boiteau, Malone, Peters, & Almor, 2014) found results consistent with this account using a dual-task paradigm, in which participants completed a finger-tapping task while listening to pre-recorded picture descriptions. Even though participants knew which pictures they would later have to describe as soon as the speaker produced the first word of their utterance, participants' finger-tapping performance was affected only when the speaker began describing the last picture in their set (around two seconds after they had started speaking), suggesting that participants delayed response preparation. Contrary to Bögels et al. (2015), these studies support the late-planning hypothesis and suggest that listeners begin preparation towards the end of the speaker's turn.

1.3. Turn-end prediction: Dissociating content from length predictability

For the late-planning hypothesis to be correct, listeners must be able to determine when the speaker's turn will end so they can begin response preparation at the appropriate moment. However, it is still largely unclear how listeners predict turn-ends.

So far in our discussion of the late-planning hypothesis, we have assumed that listeners use content predictions (i.e., lexico-semantic properties of upcoming words) to determine the speaker's turn-end. However, listeners may also predict the length of a turn by separately estimating the number of words until turn-end. Indeed, utterances are often predictable in length but unpredictable in content. To illustrate, the sentence fragment *Most people have two…* can be completed with many single words (e.g., *cars*, *dogs*, *siblings*), which overlap very little in their content. Conversely, utterances can be unpredictable in length but predictable in content. For example, the sentence fragment *The Titanic sank after…* can be completed with *it hit an iceberg, hitting an iceberg,* or *crashing*, which differ in length but overlap in content. Thus, listeners could predict a speaker's turn-end by predicting either its lexico-semantic content or its length (in number of words). Of course, being able to predict the length of the turn in number of words may not be sufficient to predict the turn-end accurately, as words differ in duration (e.g., number of syllables). However, such predictions would greatly constrain estimates of turn duration.

Given this distinction, one version of the late-planning hypothesis (*the length-prediction hypothesis*) proposes that turn-end prediction should be more precise when length is predictable rather than unpredictable, regardless of content predictability. For example, Magyari and De Ruiter (2012) found that turns that participants expected to be completed with more words (even though they could not predict the exact words) were those that elicited later button-press responses, suggesting that listeners can predict turn-ends by predicting the number of words the speaker will use.

The length-prediction hypothesis contrasts with a second version of the late-planning hypothesis, which we term the *content-prediction hypothesis*. This version maintains that length predictions are possible only when content is predictable. When content is unpredictable, listeners should not be able to predict how many words will follow. For example, Magyari, Bastiaansen, De Ruiter, and Levinson (2014) found that participants responded 70 ms before the end of predictable turns but 139 ms after the end of unpredictable turns. Together with concurrent EEG recordings, these results suggest that listeners used turn content to predict the speaker's turn-end.

However, previous studies have not manipulated length predictability independently from content predictability. In this study, we thus investigated whether participants predicted the length (in number of words) of the speaker's question, and whether they did so independently of predictions of content. To do so, we crossed our manipulation of content predictability (predictable vs. unpredictable; i.e., whether participants could predict the lexico-semantic content of upcoming words) with a manipulation of length predictability (single vs. varied; i.e., whether participants expected a single word completion or had no clear expectation about the number of words that would follow; see Table 1 for example stimuli) of simple questions.

Note that the early-planning hypothesis is not concerned with the distinction between content and length prediction, as it assumes no role for turn-end prediction. However, both versions of the late-planning account predict that listeners' button-press (Experiments 1a and 2a) and question-answering (Experiments 1b and 2b) responses should be more precise when content is predictable than when it is not. The content-prediction hypothesis predicts an interaction between content and length predictability, such that listeners should be more precise when length is predictable than when it is not, but only when content is also predictable. In contrast, the length-prediction hypothesis proposes that listeners should be more precise when length is predictable rather than unpredictable, regardless of content predictability. Finally, recall that since the early-planning hypothesis assumes that turn-end prediction does not play a role, it does not predict any effects of either content or length predictability on the precision of responses in any of the experiments.

1.4. Overview of experiments

In sum, we do not know how response preparation and articulation are interwoven during conversational turn-taking. Listeners may achieve such coordination by preparing a response early and launching articulation only after a turn-final cue (the early-planning hypothesis; Levinson & Torreira, 2015). Alternatively, they may begin preparation only when they know that the speaker is soon going to reach the end of their turn (the late-planning hypothesis; Sjerps & Meyer, 2015) and they may predict the turn-end either by predicting turn content (content-prediction hypothesis) or by predicting both turn content and turn length (length-prediction hypothesis).

To test these accounts, we conducted two pairs of experiments using button-press (Experiments 1a and 2a) and question-answering tasks (Experiment 1b and 2b). In Experiments 1a and 1b, we manipulated both the content (predictable vs. unpredictable) and length predictability (single vs. varied) of questions, to create four conditions. Experiments 2a and 2b were modelled on Experiments 1a and 1b, respectively, but included only three of the four conditions (predictable single, unpredictable single, unpredictable varied) which are sufficient to tease apart the content prediction and the length prediction hypotheses.

In the first pair of experiments, we strengthened participants' expectations about question length by having questions that were unpredictable in length end with a varied number of words (two or more);

questions whose length was predictable always ended with a single word. Since this approach made it difficult to compare content predictability across the single and varied conditions, in the second pair of experiments we selected single word completions for all questions (i.e., both those that were unpredictable and those that were predictable in length). Importantly, we found the same pattern of results across both pairs of experiments, suggesting that the length of completions chosen for the varied length conditions did not affect the results.

We analyzed both the response times (i.e., the signed deviation of listeners' responses from the turn-end) and absolute precision (i.e., how clustered around zero participants' response were) of responses in all experiments. However, precision is the most relevant measure for the button-press task, as participants are asked to respond exactly when they think the speaker will reach the end of their turn. In contrast, both response times and precision are relevant for the question-answering task, because this task captures both response preparation and turn-end prediction.

The early-planning account argues that listeners use prediction to prepare a response early, and so they should produce their verbal responses earlier when content is predictable rather than unpredictable. Since this account assumes no role for turn-end prediction, it makes no predictions regarding the precision of participants' responses. In contrast, the late-planning account argues that listeners use prediction to determine the speaker's turn-end, and so their responses should be more precise when the content (and possibly the length) of the speaker's turn is predictable rather than unpredictable. Since this account assumes no role for early preparation, it makes no predictions for effects on response times (see Table 2 for a summary of predictions).

2. Experiment 1a

Experiment 1a used a button-pressing task with four conditions. Stimuli in the single conditions were completed with a single word by the large majority of participants in a cloze pre-test, and were therefore predictable in length. Crucially, this word (in bold in the following examples) was either the same across participants (predictable single; e.g. *Are dogs your favourite **animal**?*), so that both content and length were predictable, or different (unpredictable single; e.g., *Do you enjoy going to the **supermarket**?*), so that length was predictable but content was not. Stimuli in the varied conditions were followed by completions that varied in length (i.e., their length was not predictable) and either did overlap in content (predictable varied; *Did The Titanic sink after **it hit an iceberg**?*), so that content was predictable while length was not, or did not overlap in content (unpredictable varied; *Do most students finish their **exams on time**?*), so that neither content nor length were predictable.

2.1. Method

2.1.1. Participants

Thirty native English speakers (3 males; *M*age = 20.23 years) at the University of Edinburgh participated in exchange for partial course credit or £4. Participants had no known speaking, reading, or hearing impairments.

2.1.2. Materials

We selected 116 questions (29 for each condition) using a norming task, in which 33 further participants from the same population (8 males; *M*age = 20.67) were presented with 160 question fragments and were instructed to "complete with the words or words that you think are most likely to follow the preceding context of the question" (i.e., we used a cloze task; Taylor, 1953).

We assessed length predictability by calculating the sample variance of the length (in number of words) of the completions for each fragment. In the single conditions, participants completed fragments with one word at least 90% of the time and so the length (i.e., a single word completion) was predictable. In contrast, different participants completed fragments in the varied conditions with different numbers of words (higher variance; $p < .001$, see Table 3), and so length was unpredictable. For these fragments, no more than 20% of pre-test participants provided a completion of the same length as the selected multiword completion (which was between two and eight words; $M = 3.22$).

We assessed content predictability using three different measures. First we calculated cloze probability (Taylor, 1953), which is the percentage of participants who provided a particular completion. We also computed Shannon entropy (i.e., $-\Sigma p_i \log_2(p_i)$, where p_i is the proportion of times each completion occurs for a given fragment; Shannon, 1948). Entropy is low (a minimum of 0) when completions are similar across participants, and high (a maximum of 5.04 when each of the 33 participants in the pre-test provided a different response) when responses are different. Note that both of these measures can only be

Table 2
Summary of predictions made by the accounts for the button-pressing task, which taps into turn-end prediction (Experiments 1a and 2a), and the question-answering task, which taps into turn-end prediction and response preparation (Experiments 1b and 2b).

Measures for which account makes predictions	Button-press task	Question-answering task
Signed response times	Early-planning hypothesis	
	The early planning account makes no predictions about the effects of content and length predictability during button-pressing	Content predictability: earlier responses for predictable than unpredictable questions The early-planning account makes no predictions about the effects of length predictability during question-answering
Precision	Late-planning hypothesis (content-prediction)	
	Content predictability: more precise when content is predictable than unpredictable Length predictability: no main effect on precision Content*Length predictability: more precise when length is predictable than when it is not, but only when content is predictable	Same predictions as for the button-press task
	Late-planning hypothesis (length-prediction)	
	Content predictability: more precise when content is predictable than unpredictable Length predictability: more precise when length is predictable than unpredictable	Same predictions as for the button-press task

Table 3
The means (*M*) and standard deviations (*SD*) of our measures of content predictability, length predictability, difficulty, plausibility, and duration (ms) for stimuli in Experiments 1a and 1b. The final column provides the number of utterances characterized by a pitch downstep in each condition.

Content	Length		Average variance of completion length	Completion length cloze[a]	Question fragment LSA[b]	Completion LSA[c]	Completion content cloze[d]	Question fragment entropy[e]	Question difficulty[f]	Question plausibility[f]	Question duration (ms)	Downstepped utterances
Predictable	Single	*M*	0.02	99%	0.91	0.95	93%	0.35	6.22	6.64	2398	29/29
		SD	0.04	3%	0.11	0.06	8%	0.36	0.48	0.35	646	
	Varied	*M*	1.18	19%	0.71	0.68	–	–	6.11	6.45	2996	27/29
		SD	0.82	14%	0.14	0.19	–	–	0.35	0.27	620	
Unpredictable	Single	*M*	0.11	92%	0.37	0.16	4%	3.01	6.17	6.52	1932	26/29
		SD	0.09	8%	0.12	0.08	2%	0.63	0.42	0.40	452	
	Varied	*M*	0.95	18%	0.35	0.23	–	–	6.24	6.48	2542	27/29
		SD	0.44	15%	0.11	0.12	–	–	0.40	0.39	597	

[a] Percentage of participants who provided the word length of the selected completion used in the main experiment (a single word in the single conditions; multiple words in the varied conditions) as a continuation in the cloze task.
[b] Average over all completion comparisons for that particular fragment.
[c] Average over comparisons between the selected completion and all other completions.
[d] Cloze percentages of the selected completion. If cloze percentage is higher, then participants converged on a completion.
[e] Entropy of question fragments presented to participants in the cloze task. If entropy is lower, then participants converged on a completion.
[f] Difficulty and plausibility ratings made on a scale of 1–7. 1 indicated that the question was very implausible/difficult to answer, while 7 indicated that the question was very plausible/easy to answer.

computed for stimuli in the single conditions, as completions in the varied condition may differ verbatim while having similar content (e.g., *it hit an iceberg* vs. *hitting an iceberg*). Stimuli in the predictable single condition had higher cloze probability ($p < .001$; see Table 3) and lower entropy ($p < .001$) than those in the unpredictable single condition ($p < .001$; see Table 3).

Finally, we computed Latent Semantic Analysis (LSA; Deerwester, Dumais, Furnas, Landauer, & Harsman, 1990) matrix comparisons using the general reading corpus. LSA determines the semantic similarity of words and phrases by calculating the extent to which they occur in the same context, and ranges from 1 (completions are identical) to −1 (completions are completely different). Importantly, it can be used to assess the similarity of completions that differ in number of words.

Using these LSA comparisons, we first calculated the content predictability of each fragment by averaging over the LSA scores for all pairwise comparisons between completions. Stimuli in the predictable content condition had higher fragment LSA than those in the unpredictable content conditions ($p < .001$; see Table 3). We also calculated the LSA value of each completion by averaging over the LSA scores for all comparisons between the chosen completion and every other completion to the same fragment. Completion LSA was higher in predictable than unpredictable conditions ($p < .001$).

The four conditions were matched for average difficulty and plausibility (all $ps > .07$; see Table 3) using data collected in a second pretest, in which 15 new native English speakers (2 males; *M*age = 19.40) rated (i) how difficult they would find it to answer the question if asked, and (ii) whether the question made sense. Both ratings were made on a scale of 1 (very implausible/difficult to answer) to 7 (very plausible/easy to answer).

All questions were recorded by a native English male speaker, who was instructed to read the utterances as though "you are asking a question and expecting a response". Recordings were between 1317 and 5194 ms in duration (see Table 3). Utterances in the varied conditions were longer than those in the single conditions ($p < .001$), and those in the predictable condition were also longer than those in the unpredictable condition ($p < .001$; we return to this issue in the Results). All our questions had falling boundary tones, and 109 (see Table 3) were characterized by a pitch downstep, which occurs when the pitch of each syllable is lower than the previous syllable (Beckman & Pierrehumbert, 1986). Both judgments were validated by a second rater, who listened to 25% of the utterances (Cohen's kappa = 1, for both ratings).

2.1.3. Procedure

The experiment was controlled using E-Prime (version 2.0). Participants pressed a button to start audio playback of the question. A fixation cross (+) appeared 500 ms before question onset, and the screen turned red as audio playback began. Using a translation of the instructions used by De Ruiter et al. (2006), participants were told: "Press the button (using your dominant hand) when you believe the question will end. Do not wait until the speaker has finished the question and stopped speaking. Instead, you should press the button as soon as you expect the speaker to finish". Thus, they were encouraged to predict the turn-end, rather than simply wait for the speaker to reach the end of his utterance. Participants responded by pressing the middle button of a SR-box and audio playback stopped as soon as a response was recorded (as in De Ruiter et al., 2006).

Participants completed ten initial practice trials to familiarize themselves with the experimental procedure. The 116 stimuli were individually randomized, and participants were given the opportunity to take a break every 29 items.

2.2. Data analysis

Precision analyses are most relevant for this experiment, because the button-press task encourages participants to accurately predict the turn-end. The late-planning hypothesis predicts effects of content predictability (and possibly length predictability, depending on whether participants make separate content and length predictions) on the precision of participants' button-press responses, whereas the early-planning hypothesis does not predict any differences in precision. In addition, and for comparison with Experiment 1b, we also analyzed signed response times. Response times were defined with respect to question offset, and were negative when participants responded before the end of the speaker's question and positive when they responded after the end. We replaced 23 (0.66%) response times falling at least 2.5 standard deviations above the by-participant mean and 96 (2.76%) response times below the by-participant mean with the respective cut-off value. Note that, throughout our analyses, the results were the same regardless of whether or not responses were replaced with cut-off values. We evaluated the effects of content and length predictability on response times with linear mixed effects models (LMM; Baayen, Davidson, & Bates, 2008) using the *lmer* function of the *lme4* package (version 1.1–12; Bates, Maechler, Bolker, & Walker, 2015) in RStudio (version 0.99.896) with a Gaussian link function.

Precision was defined as the absolute value of response time. Before

taking the absolute value, we first standardized response time to have a mean of zero, so that we could assume a half-normal distribution or, equivalently (Leone, Nelson, & Nottingham, 1961), a normal distribution truncated at zero. Given that the distribution of response precision is truncated at the lower boundary of zero, the distributional assumptions of lmer are not met. Therefore, we used Bayesian mixed effects models (BMM) as implemented in the *brms* package (version 1.6.1; Bürkner, 2017). We initially fitted models using a normal distribution truncated at zero. However, such models did not converge, so we modelled our data using three other distribution families: the log-normal, the gamma, and the Weibull distribution (e.g., Pinder, Wiener, & Smith, 1978). In all cases, the Weibull was a better fit than either the log-normal or the gamma (assessed using LOO comparisons), and so we report parameters and credible intervals from models fitted using a Weibull distribution. We ran 4 chains per model, each for 1600 iterations, with a burn-in period of 800, and initial parameter values set to zero. All of the reported models converged with no divergent transitions (all $\hat{R}$ values ≤ 1.1); the number of effective samples for each estimate is reported in the Appendix.

Although the parameterization of the Weibull distribution implemented in *brms* is based on a scale and a shape parameter, we report and discuss only scale parameters; shape is most often used to model failure or mortality rates, which is not relevant to response precision (although full models are reported in the Appendix). The scale parameter, on the other hand, quantifies the spread of the distribution and is thus informative of the degree of precision in participants' responses. Note that scale parameters were fitted on the log scale (reported in the Appendix), but we report exponentiated estimates in the Results section as they are easier to interpret: The larger the exponentiated value of the scale parameter, the more spread out the probability mass of the distribution. All distributions were fitted using default *brms* priors.

In all instances, we fitted models using the maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013), except that correlations among random effects were fixed to zero to aid convergence. We fitted the full model where response times or precision was predicted by Content predictability (reference level: unpredictable vs. predictable), Length predictability (reference level: varied vs. single), and their interaction. These predictors were contrast coded (-0.5, 0.5) and centered. We also included Question Duration in our analyses (which was centered), since previous research suggests that longer turns tend to elicit earlier button-press responses (e.g., De Ruiter et al., 2006). To aid convergence, this predictor was included only as a main effect.

For the LMM analyses, we report coefficient estimates (b), standard errors (SE), and t values for each predictor. We assume that an absolute t value of 1.96 or greater indicates significance at the 0.05 alpha level (Baayen et al., 2008). For the BMM analyses, we report coefficient estimates of effect size (b), estimate errors (SE), and the 95% credible interval (CrI; i.e., under the model assumptions, there is a 95% probability that the parameter estimate is contained in this interval) for each predictor. If zero lies outside the credible interval, then we conclude there is sufficient evidence to suggest the estimate is different from zero.

2.3. Results

2.3.1. Analysis of response times

On average, participants responded 136 ms (see Fig. 1) before the end of the speaker's utterance, and 92% of the responses occurred within 1000 ms of the speaker's turn-end (see Fig. 2).

We found no significant effects of Content predictability ($b = -28.31$, $SE = 29.10$, $t = -0.97$) or Length predictability ($b = -19.25$, $SE = 34.00$, $t = -0.55$), and no interaction between the two ($b = -8.57$, $SE = 50.15$, $t = 0.17$; see the Appendix for full models). In contrast, Question Duration was a negative predictor of response times ($b = -152.17$, $SE = 15.04$, $t = -10.12$): Longer questions elicited earlier responses than shorter questions. Although there is a numerical difference in response times and response precision between the conditions in Fig. 1, note that these means are not adjusted for Question Duration, and our models show that this variable explains any differences in the observed means between conditions.

2.3.2. Precision analysis

Participants responded on average 303 ms away from the end of the speaker's turn (see Fig. 1 for a breakdown by condition). We found no evidence that either Content predictability ($b = -1.03$, $SE = 1.10$, CrI [-0.22, 0.16]), Length predictability ($b = -1.04$, $SE = 1.12$, CrI [-0.25, 0.17]), or the interaction between the two ($b = -1.28$, $SE = 1.20$, CrI[-0.60, 0.10]) affected the scale parameter of the distribution. However, Question Duration had a positive effect on scale ($b = 1.19$, $SE = 1.05$, CrI[0.07, 0.27]), such that the spread of the distribution was greater when questions were longer.

2.4. Discussion

In Experiment 1a, we investigated whether turn-end prediction plays a role in conversational turn-taking, as predicted by the late-planning hypothesis (e.g., Sjerps & Meyer, 2015; see Table 2). Specifically, we examined whether listeners predict the speaker's turn-end by predicting its content and length independently of one another (length-prediction hypothesis; Magyari & De Ruiter, 2012), or whether they predict length only if content is predictable (content-prediction hypothesis; Magyari et al., 2014). Recall that the early-planning hypothesis assumes that turn-end prediction does not play a role in turn-taking, and so makes no predictions for this task (see Table 2).

Inconsistent with the late-planning hypothesis, we found no effects of content or length predictability when analyzing the precision of participants' button-press responses. Instead, responses were influenced by question duration: Longer questions elicited less precise (and earlier) responses than shorter questions, as in previous research using the button-press paradigm (e.g., De Ruiter et al., 2006). There were also no content and length effects on signed response times; this contrasts with previous findings using the button-press paradigm (e.g., Magyari & De Ruiter, 2012; Magyari et al., 2014), which have shown that listeners respond earlier to predictable than unpredictable turns, even when conditions are matched for average duration.

This duration effect could be interpreted in line with previous research using reaction time experiments (see also Magyari, De Ruiter, & Levinson, 2017), which has found that response times are longer when the interval between a warning signal (alerting participants to the forthcoming reaction stimulus) and the reaction stimulus is shorter (e.g., Näätänen, 1971). When the utterance is longer, the interval between the warning signal and the reaction stimulus (i.e., between turn onset and turn-end) is also longer, and since the probability of the reaction stimulus (the turn-end) occurring continuously increases (Sanders, 1966), the listener is more likely to respond earlier when the utterance is longer in duration.

Another possibility is that longer turns elicit earlier responses because they typically contain more points of possible turn completion (see Sacks, Schegloff, & Jefferson, 1974), and the listener may simply be more likely to mistake one of these points of completion for the actual turn-end. For example, consider the long question (2761 ms) *Did The Titanic sink after hitting an iceberg*?. It contains two plausible completion points: One after *sink*, and another after *iceberg*. Now compare it to the short question (1729 ms) *Are dogs your favourite animal*?, which contains only one plausible completion point (after *animal*) that coincides with the end of the question. Listeners may respond earlier to the first turn because there is an additional point of possible turn completion, before the actual turn-end.

In sum, the results of Experiment 1a did not provide any evidence to suggest that participants used either content or length predictability to determine the speaker's turn-end. Following Dienes (2014), we

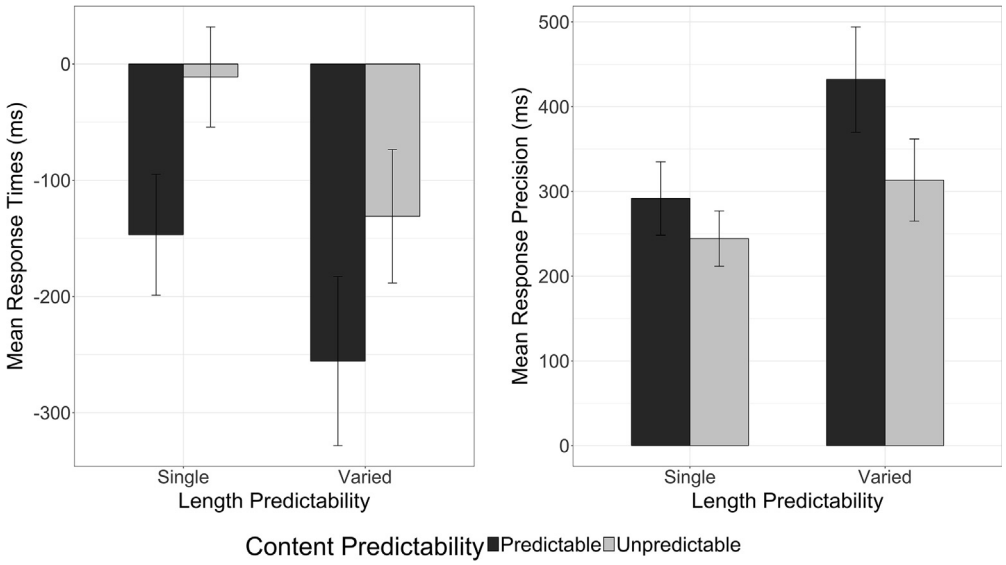

Fig. 1. Observed means of response times (left) and precision (right) for the four conditions in Experiment 1a. Error bars represent ± 1 standard error from the mean.

compared the null effect of content predictability with an hypothesized effect size distribution ranging between 0 and twice the mean condition difference reported by Magyari et al. (2014): 209 ms. The resulting Bayes factor was less than 0.33 (B = 0.11), indicating strong evidence in favor of the null hypothesis. (Note that we could not compute Bayes factors for the effect of Length predictability because we lack a measure of effect size.) These findings are more consistent with the early-planning hypothesis, which suggests listeners use predictions of turn content to prepare a response, but not to predict the speaker's turn-end. Since our conclusions are based on null results, however, we conducted Experiment 1b (a question-answering task) to test further predictions of the latter hypothesis, namely that listeners use content predictions to prepare a response as early as possible.

3. Experiment 1b

Experiment 1b was identical to Experiment 1a, with the exception that participants verbally answered each question either *yes* or *no*. If the early-planning hypothesis is correct, then we expected participants to answer earlier when question content was predictable rather than unpredictable. Since we found no evidence to suggest listeners used content or length predictability to predict turn-endings in Experiment 1a, we did not predict any effects of content or length predictability on the precision of participants' verbal responses.

3.1. Method

3.1.1. Participants

Thirty new participants from the same population as in Experiment 1a (4 males, *M*age = 19.43) participated on the same terms.

3.1.2. Materials and procedure

The materials and procedure were identical to those used in Experiment 1a, with the exception that participants were told: "Answer as quickly as possible. Do not wait until the speaker has finished the question and has stopped speaking. Instead, you should answer as soon as you expect the speaker to finish the question". Thus, participants were encouraged to prepare a response as soon as possible (rather than simply wait for the speaker to finish) and articulate it close to the speaker's turn-end. Participants spoke into the microphone, and playback stopped as soon as a response was recorded using a voicekey.

3.2. Data analysis

Response times and precision were calculated using the same procedure as Experiment 1a. Of the 3468 responses, 188 (5.42%) were discarded because they could not be categorized as *yes* or *no*. We removed a further 12 (0.35%) response times greater than 10,000 ms, as they were clear outliers. We then replaced 45 response times (1.37%) at

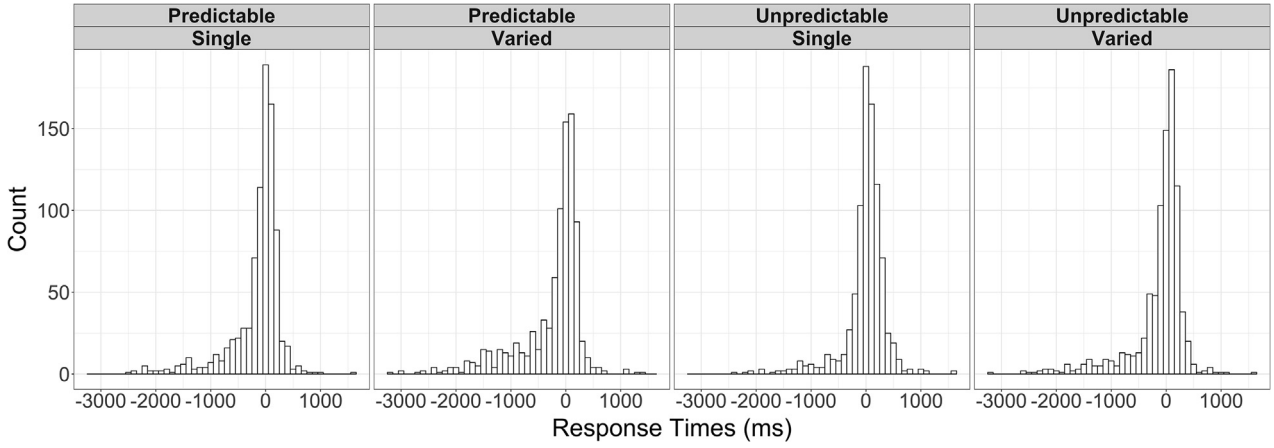

Fig. 2. The distribution of response times in the four conditions in Experiment 1a. Trials are placed into 100 ms time bins.

the upper limit and 27 (0.37%) at the lower limit.

We fitted models using the same procedure as in Experiment 1a. However, we included two further predictors to account for possible answer characteristics. *Yes* responses are usually produced faster than *no* responses (e.g., Strömbergsson, Hjalmarsson, Edlund, & House, 2013), and so we included Answer Type (reference level: no vs. yes) in our analyses. Since some of our questions were fact-based (e.g., *Did The Titanic sink after hitting an iceberg*?) while others were opinion-based (e.g., *Are dogs your favourite animal*?) we also included Agreement, which was the absolute difference between the percentage of participants who answered yes and the percentage who answered no. We assume that fact-based questions are likely to have a clear answer, and so Agreement will be high (a maximum of 100 when all participants provide the same answer). Thus, participants may need less time to determine what to say. For opinion-based questions, however, both *yes* and *no* are equally plausible answers, and thus Agreement will be low (a minimum of 0 when half of the participants answer *yes* and half answer *no*). As a result, participants may need more time to decide what to say.

3.3. Results

3.3.1. Response time analysis

On average, participants responded 379 ms after the end of the speaker's turn (see Fig. 3), and 90% of responses occurred within 1000 ms of the speaker's turn-end (see Fig. 4).

Participants answered earlier when content was predictable rather than unpredictable ($b = -153.01$, $SE = 34.08$, $t = -4.49$). However, there was no effect of Length predictability ($b = 10.89$, $SE = 33.25$, $t = 0.33$), and no interaction between Content and Length predictability ($b = -110.21$, $SE = 63.75$, $t = -1.73$). Inconsistent with previous research (e.g., Strömbergsson et al., 2013), response times were not affected by Answer Type ($b = -21.86$, $SE = 16.46$, $t = -1.33$): Participants were equally fast to respond *yes* and *no,* which may suggest that having participants interact with a pre-recorded speaker, rather than an actual interlocutor, reduces the social bias against "no" responses. However, Agreement was a significant negative predictor of response times ($b = -55.21$, $SE = 15.17$, $t = -3.64$): As expected, questions with higher agreement elicited earlier response times than those with lower agreement. In addition, longer questions elicited earlier responses than shorter questions ($b = -72.88$, $SE = 17.25$, $t = -4.23$), as in Experiment 1a.

3.3.2. Precision analysis

On average, participants answered 509 ms away from the end of the speaker's turn (see Fig. 4 for a breakdown by condition). We found no evidence for an effect of either Content predictability ($b = 1.05$, $SE = 1.13$, CrI[−0.17, 0.28]), Length predictability ($b = 1.02$, $SE = 1.08$, CrI[−0.14, 0.18]), or their interaction ($b = -1.20$, $SE = 1.15$, CrI[−0.47, 0.09]). Precision was not influenced by Answer Type ($b = -1.01$, $SE = 1.04$, CrI[−0.10, 0.07]) or Agreement ($b = -1.06$, $SE = 1.03$, CrI[−0.13, 0.00]), but the spread of the distribution was greater when questions were longer in duration ($b = 1.16$, $SE = 1.04$, CrI[0.08 0.22]), as in Experiment 1a.

3.4. Comparison analysis with Experiment 1a

To determine whether the effect of content predictability in Experiment 1b was significantly different from Experiment 1a, we conducted a cross-experiment comparison. We used the same analysis structure as in Experiment 1b, but included an interaction between Content predictability, Length predictability, and Experiment (reference level: question-answering vs. button-pressing). Experiment was contrast coded (−0.5, 0.5), centered, and included as by-items random slopes. Since the size of the estimates suggested that Question Duration had a larger effect in Experiment 1a ($b = -152.17$) than 1b ($b = -72.88$), we included a Question Duration by Experiment interaction in the fixed effects structure of the model. Although we did not include Answer Type (yes or no) as a main effect because this variable was participant-specific (i.e., different participants answered yes or no to different items), we did include Agreement, since this variable was item-specific.

Importantly, when analyzing response times, we found a significant effect of Content predictability ($b = -86.88$, $SE = 29.75$, $t = -2.92$), Experiment ($b = -491.56$, $SE = 79.38$, $t = -6.19$), and a significant interaction between the two ($b = 156.80$, $SE = 39.69$, $t = 3.95$), confirming that Content predictability affected the timing of participants' verbal responses more than the timing of their turn-end predictions. In addition, there was no effect of Length predictability, and this predictor did not interact (either two-way or three-way) with any other predictors (all $ts < 1.96$).

When analyzing the precision of participants' responses, we found an effect of Experiment ($b = -1.90$, $SE = 1.22$, CrI[−1.04, −0.24]), but no effect of Content predictability ($b = -1.05$, $SE = 1.07$, CrI [−0.19, 0.10]), Length predictability ($b = -1.01$, $SE = 1.07$, CrI

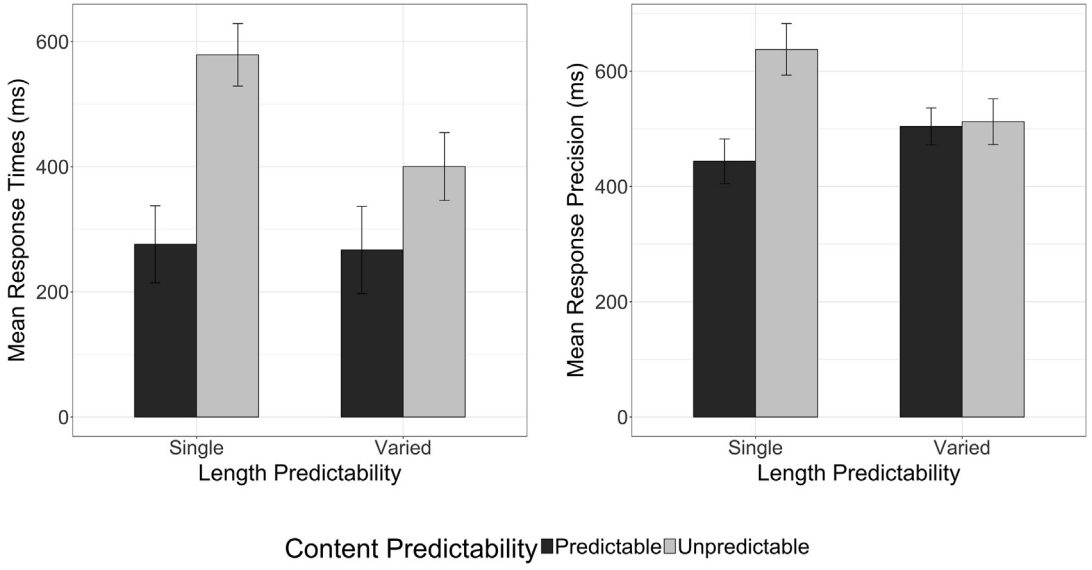

Fig. 3. Observed means of response times (left) and precision (right) for the four conditions in Experiment 1b.

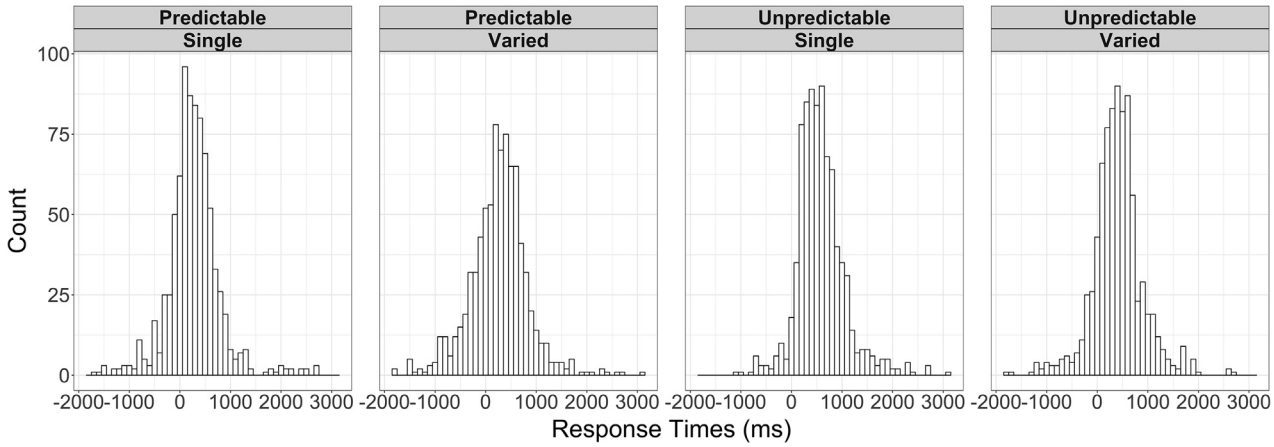

Fig. 4. The distribution of observed response times in the four conditions in Experiment 1b.

[−0.14, 0.12]), and no interaction between any of these predictors (all CrIs included 0). Response times and precision were influenced by Agreement and Question Duration in the same way as in the individual analyses; in addition, Agreement had a negative influence on the precision of responses in the comparison analysis ($b = -1.08$, $SE = 1.03$, CrI[−0.13, −0.03]), even though it did not in the individual experiment analyses. These results suggest that the lack of predictability effects on the precision of participants' responses was comparable in the question-answering and button-pressing tasks. Along with the individual experiment analyses, these results confirm there was an effect of content predictability in the question-answering task, but not in the button-pressing task. Thus, participants used content predictions to prepare their response, but not to predict the speaker's turn-end.

3.5. Discussion

In Experiment 1b, we investigated whether early response preparation occurs during turn-taking. Participants answered earlier when question content was predictable rather than unpredictable, suggesting they used predictions of turn content to prepare a verbal response. In contrast, we found no effects of content or length predictability on the precision of participants' responses. Together with Experiment 1a and our cross-experiment comparisons, these results suggest that listeners in our experiments used content predictions to prepare their verbal response as early as possible but not to predict the turn-end, and are thus consistent with the early-planning hypothesis.

However, in both Experiments 1a and 1b, our measures of content predictability were not comparable across the single and varied length conditions. Since we used multi-word completions in the varied conditions, the predictability of completions was assessed at an earlier point in the varied than in the single conditions. For example, the unpredictable varied question *Do most students finish their exams on time*? was cut off three words before question end (*Do most students finish their*…) in the pre-test, whereas the unpredictable single question *Do you enjoy going to the supermarket?* was cut off just one word before question end (*Do you enjoy going to the*…). But the content predictability of the utterance may well increase with each additional word the speaker produces. For instance, the listener cannot predict what the speaker will say after the words *Do most students finish their*… (and so the predictability of question content is fairly low at this point), but may be able to predict *time* after hearing *Do most students finish their exams on*…).

Indeed, when we conducted a cloze post-test to assess the content predictability of the final word of the questions in the varied conditions, in which 33 participants from the same population as Experiment 1 (8 males; *M*age = 20.15) completed the same procedure as previous pretests, we found that stimuli in the two varied conditions had significantly higher content predictability (predictable varied completion cloze: 76%, unpredictable varied completion cloze: 68%; predictable varied completion LSA: 0.83, unpredictable varied completion LSA: 0.73) than those in the unpredictable single condition (completion cloze: 4%; completion LSA: 0.16; all *ps* < 0.001). Thus, even though the predictable and unpredictable single conditions demonstrate that listeners can use content predictions to prepare their responses early, our measures of content predictability in the varied conditions were not comparable to those in the single conditions. This may have affected our length predictability manipulation, and so we conducted two further experiments (Experiments 2a and 2b) in which all stimuli had single word completions to provide a further test of the length prediction hypothesis.

4. Experiment 2a

Experiment 2a was identical to Experiment 1a, in that participants were instructed to press a button when they thought the speaker had reached the end of their turn, but we selected single word completions for all stimuli to ensure content predictability was comparable across the conditions. We also discarded the predictable varied condition from Experiment 1 because most of these stimuli were completed with a single word most of the time, and so a single word completion would have been predictable in this condition.

Importantly, discarding the predictable varied condition does not affect our ability to disentangle late from early-planning, as we can still examine effects of content predictability across the button-press and the question-answering paradigm. It also does not affect our ability to determine whether participants predicted the speaker's turn-end by predicting the length of the speaker's utterance separately from its content, as we can still compare the two unpredictable content conditions. The content-prediction hypothesis predicts no difference in response precision in the two unpredictable content conditions; the length-prediction hypothesis predicts that responses should be more precise for unpredictable utterances whose length is predictable (i.e., unpredictable single condition) rather than unpredictable (i.e., unpredictable varied condition).

To minimize any confounding effect of Question Duration (as occurred in Experiment 1a), we followed Magyari et al. (2014) and matched the average duration of the three stimulus conditions. Since we also used the same stimuli in Experiment 2b, we matched the average Agreement of the three conditions.

4.1. Method

4.1.1. Participants

Thirty new native English speakers (10 males; *M*age = 22.20) at the University of Edinburgh participated on the same terms as previous experiments.

Table 4
The means (*M*) and standard deviations (*SD*) of our measures of content predictability, length predictability, difficulty, plausibility, answer agreement, and duration (ms) for stimuli in Experiments 2a and 2b. The final column provides the number of utterances characterized by a pitch downstep in each condition.

Condition		Average variance of completion length	Completion length cloze[a]	Question fragment LSA[b]	Completion LSA[c]	Completion content cloze[d]	Question fragment entropy[e]	Question difficulty[f]	Question plausibility[f]	Answer agreement	Question duration (ms)	Downstepped utterances
Predictable single	*M*	0.03	98%	0.90	0.94	91%	0.43	6.34	5.78	53%	2284	23/28
	SD	0.04	3%	0.11	0.07	9%	0.37	0.52	0.64	36%	632	
Unpredictable single	*M*	0.05	97%	0.37	0.15	5%	2.96	6.00	5.58	37%	2021	22/28
	SD	0.05	3%	0.12	0.07	2%	0.68	0.76	0.56	27%	560	
Unpredictable varied	*M*	0.88	38%	0.34	0.20	–	–	6.21	5.68	43%	2031	15/28
	SD	0.59	21%	0.10	0.14	–	–	0.47	0.52	27%	489	

[a] Percentage of participants who provided the word length of the selected completion used in the main experiment (a single word in the single conditions; multiple words in the varied conditions) as a continuation in the cloze task.
[b] Average over all completion comparisons for that particular fragment.
[c] Average over comparisons between the selected completion and all other completion.
[d] Cloze percentages of the selected completion. If cloze percentage is higher, then participants converged on a completion.
[e] Entropy of question fragments presented to participants in the cloze task. If entropy is lower, then participants converged on a completion.
[f] Difficulty and plausibility ratings made on a scale of 1–7. 1 indicated that the question was very implausible/difficult to answer, while 7 indicated that the question was very plausible/easy to answer.

4.1.2. Materials

We constructed 141 question fragments, sometimes by re-using materials from Experiment 1. Note that we pre-tested both old and new fragments to ensure consistency across the item set. We selected completions for these fragments using the same pre-test procedure as in Experiment 1, with 33 new native English speakers (2 males, *M*age = 20.03 years). Using these responses, we selected 28 stimuli for each of the three conditions (84 stimuli in total).

We calculated content and length predictability as in Experiment 1. However, we selected single word completions for all fragments in all conditions. This completion length was used by at least 90% of participants in the single conditions, and by no more than 72% of participants in the unpredictable varied condition (see Table 4). Questions in the predictable and unpredictable single conditions were matched for average completion length variance ($p = .15$), and both conditions had lower variance than questions in the unpredictable varied condition (all $ps < .001$; see Table 4).

Stimuli in the predictable single condition had higher fragment LSA than the two unpredictable content conditions (all $ps < .001$). In addition, the predictable single condition had higher cloze probability and lower entropy than the unpredictable single condition (all $ps < .001$). The LSA values for the two unpredictable conditions were matched (all $ps > .13$; see Table 4).

We matched the mean difficulty, plausibility, and answer agreement (all $ps > .09$) of the three conditions using data from a separate pre-test, in which participants (31 native English speakers; 5 males, *M*age = 20.58) answered each question either *yes* or *no* and rated the difficulty and plausibility of questions, as in Experiment 1a. Questions were recorded by the same native English speaker as in Experiment 1a, and were matched for average duration (all $ps > .21$; see Table 4). When analyzing the pitch contours of these questions, six (7%) had creaky voice, all had falling boundary tones, and sixty (71%) had a downstep in pitch (see Table 4). Both judgments were again validated the same second coder as in Experiment 1a, who rated 25% of the stimuli. This resulted in a Cohen's kappa of 1 for boundary tone judgements and 0.72 for downstep judgements, which is considered "good" agreement (see Cicchetti, 1994; Landis & Koch, 1977). Note that, if listeners use downsteps to determine the speaker's turn-end (e.g., Cutler & Pearson, 1986), then we would expect them to be more precise at timing their response in the unpredictable varied condition (where there are more downsteps) than in either the unpredictable or the predictable content single conditions. However, this is the opposite of the predictions made by the content- or length-prediction hypotheses.

4.1.3. Procedure

The procedure was identical to Experiment 1a, except that breaks occurred after every 28 stimuli.

4.2. Data analysis

Response times and precision were analyzed as in Experiment 1a. We replaced 12 response times (0.48%) above the upper limit, and 66 (2.62%) below the lower limit with the cut-off value. Data analysis, predictors, and random effects structure were identical to those used in Experiment 1a. However, we defined two orthogonal Helmert contrasts to capture effects of Content and Length predictability. The Content contrast compared the mean of the two unpredictable conditions (1/3) to the predictable condition (−2/3, reference level), and the Length contrast compared the unpredictable varied condition (0.5) to the unpredictable single condition (−0.5, reference level). Since the two contrasts are orthogonal, no interaction term was included. Even though we balanced Question Duration, we still included it as an additional main effect to ensure our results could not be attributed to any residual differences. All predictors were centered.

4.3. Results and discussion

4.3.1. Analysis of response times

Participants responded 117 ms before the end of the speaker's turn (see Fig. 5) and 93% of responses occurred within 1000 ms of the end of the speaker's question (see Fig. 6).

As in Experiment 1a, we found no significant effect of Content ($b = 0.39$, $SE = 35.60$, $t = 0.01$) or Length predictability ($b = 18.75$, $SE = 41.74$, $t = 0.45$). The Bayes factor for the null effect of content predictability was 0.05, again indicating strong evidence in favor of the null hypothesis. Question Duration was still a negative predictor of response times ($b = -125.00$, $SE = 41.74$, $t = -8.98$).

4.3.2. Precision analysis

Participants responded 297 ms away from the end of the speaker's question on average (see Fig. 5). We found no evidence for an effect of either Content predictability ($b = 1.26$, $SE = 1.16$, CrI[−0.07, 0.53]) or Length predictability ($b = 1.11$, $SE = 1.30$, CrI[−0.41, 0.62]). However, the spread of the distribution was again greater when

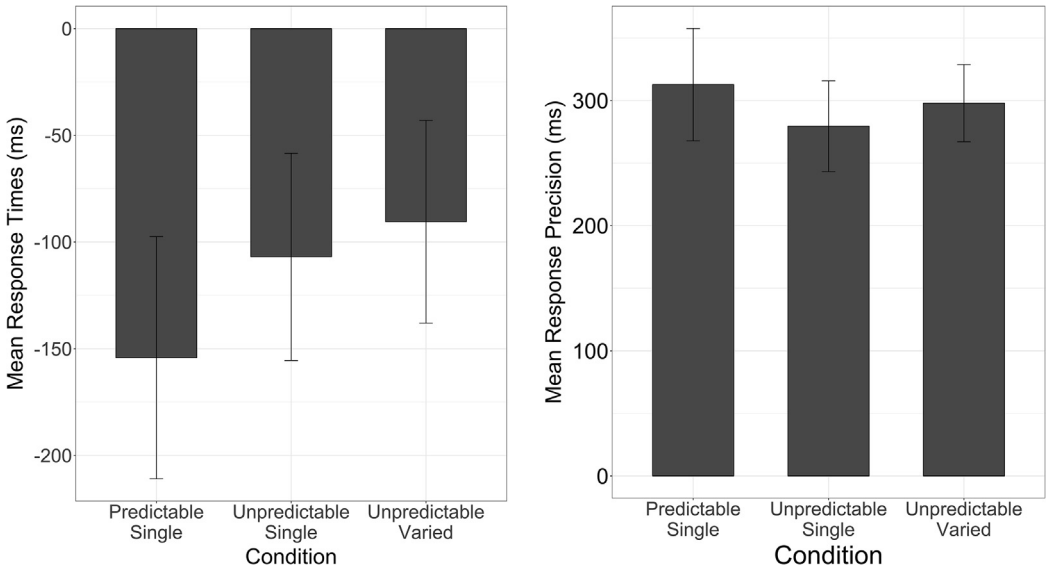

Fig. 5. Observed means of response times (left) and precision (right) for the three conditions in Experiment 2a.

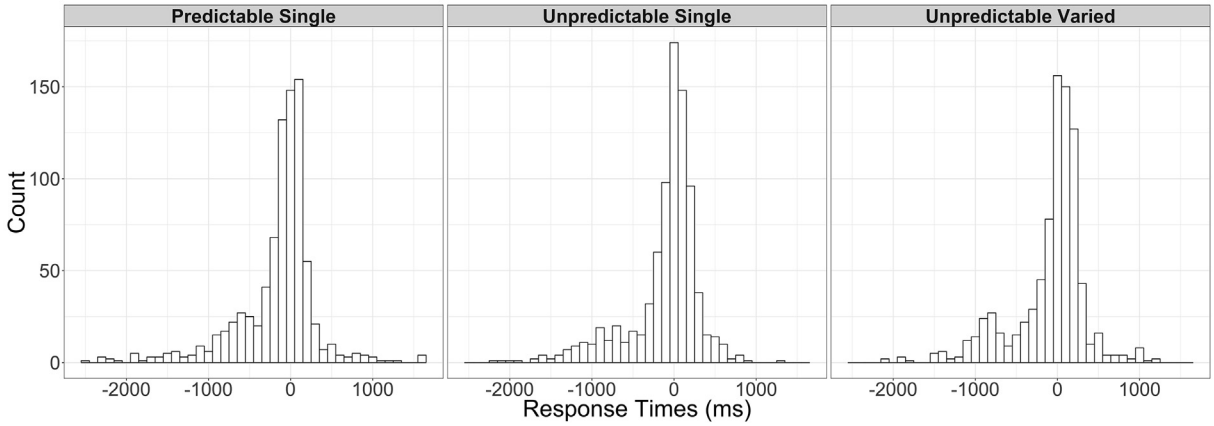

Fig. 6. The distribution of observed response times in the three conditions in Experiment 2a.

questions were longer in duration (b = 1.26, SE = 1.05, CrI[0.13, 0.32]). These results are consistent with Experiment 1a, and provide no support for the idea that listeners used content or length predictability to predict the speaker's turn-end.

5. Experiment 2b

Experiment 2b was identical to Experiment 1b, in that participants verbally answered each question either *yes* or *no*, but we used the same stimuli from Experiment 2a. If participants use content predictions to prepare a verbal response, then we expect them to answer earlier when question content is predictable rather than unpredictable. Since we found no evidence to suggest listeners used content or length predictability to determine the end of the speaker's turn in any of the previous experiments, we did not expect either of these variables to influence response precision.

5.1. Method

5.1.1. Participants

Thirty new participants from the same population in the previous three experiments (10 males; *M*age = 22.20) took part on the same terms.

5.1.2. Materials and procedure

The materials were identical to those used in Experiment 2a, and the procedure was identical to that used in Experiment 1b.

5.2. Data analysis

We discarded 39 responses (1.58%) because they could not be clearly categorized as *yes* or *no*. We discarded nine (0.36%) response times greater than 10,000 ms, and then replaced 39 response times (1.58%) at the upper limit and 30 (1.21%) at the lower limit. We analyzed response times and precision using the same procedure as Experiment 2a, but in addition we also included Answer Type (reference level: no vs. yes) and Answer Agreement as main effects.

5.3. Results and discussion

5.3.1. Analysis of response times

Participants responded 484 ms after the end of the speaker's turn (see Fig. 7) and 89% of responses occurred within 1000 ms of question end (see Fig. 8).

Participants answered earlier when question content was predictable rather than unpredictable (b = 95.78, SE = 34.54, t = 2.77). However, there was no effect of Length predictability (b = 28.19, SE = 36.81, t = 0.77). These results replicate Experiment 1b, and suggest that participants prepared their answer as early as possible.

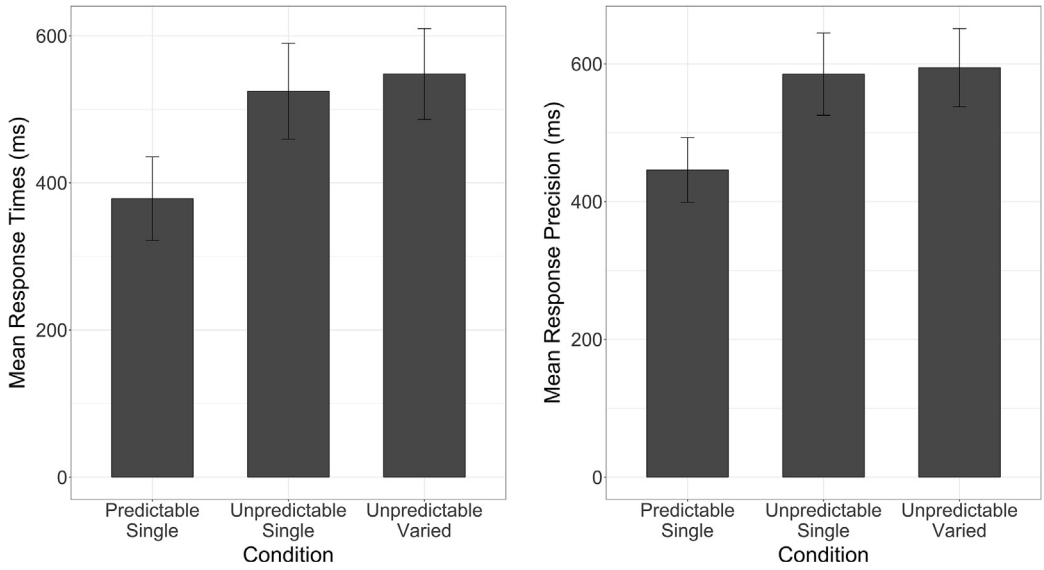

Fig. 7. Observed average response times (left) and precision (right) for the three conditions in Experiment 2b.

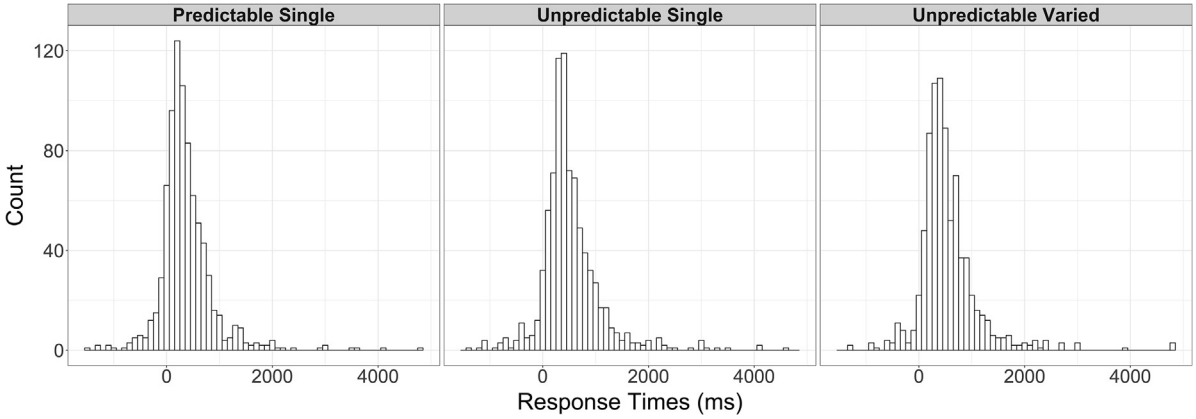

Fig. 8. The distribution of observed response times in the three conditions in Experiment 2b.

Unlike Experiment 1b, participants answered *yes* earlier than *no* ($b = -143.43$, $SE = 19.92$, $t = -7.20$). This replicates previous studies (e.g., Stivers et al. 2009; Strömbergsson et al., 2013) and suggests that the lack of an effect of Answer Type in Experiment 1b cannot be attributed to the fact that our participants interacted with a pre-recorded speaker rather than an actual interlocutor. In addition, participants answered questions with higher agreement earlier than those with lower agreement ($b = -35.81$, $SE = 15.66$, $t = 2.29$). Finally, questions longer in duration elicited earlier response times than those shorter in duration ($b = -59.85$, $SE = 15.67$, $t = -3.82$). Together with Experiment 1b, these results suggest that Answer Type, Agreement, and Question Duration all influence response times during a question-answering paradigm.

5.3.2. Precision analysis

Participants responded 542 ms away from the end of the speaker's question (see Fig. 7). Response precision was not influenced by Content predictability ($b = 1.13$, $SE = 1.11$, CrI[−0.08, 0.33]), Length predictability ($b = 1.01$, $SE = 1.16$, CrI[−0.28, 0.31]), Answer Type ($b = 1.00$, $SE = 1.05$, CrI[−0.09, 0.09]), or Answer Agreement ($b = 1.02$, $SE = 1.04$, CrI[−0.05, 0.09]). However, the spread of the distribution was greater when questions were longer in duration ($b = 1.12$, $SE = 1.04$, CrI[0.04, 0.18]). These results replicate Experiment 1b, and suggest participants did not time response articulation by predicting the content or the length of the speaker's question.

5.4. Comparison analysis with Experiment 2a

As in Experiments 1a and 1b, we conducted a cross-experiment comparison between Experiments 2a and 2b. We used the same analysis structure as in the previous cross-experiment comparisons, but with predictors defined as in Experiment 2b. Recall that Content and Length predictability were implemented as orthogonal contrasts in Experiment 2b; therefore, we included two three-way interactions between Content predictability, Experiment, and Question Duration and between Length predictability, Experiment, and Question Duration, but no four-way interaction.

We could not analyze the precision of participants' responses because the model did not converge ($\hat{R}$ values > 1.1), but note that we found no effects of either Content or Length predictability on precision in either Experiment 2a or 2b. Below, we report only a cross-experiment comparison of the analysis of response times.

Importantly, when analyzing response times, we found no significant effect of Content predictability ($b = 40.15$, $SE = 33.24$, $t = 1.21$) or Length predictability ($b = 52.22$, $SE = 61.83$, $t = 0.84$). There was a significant effect of Experiment ($b = -597.21$, $SE = 15.33$, $t = -38.97$), such that participants responded earlier in the button-press than question-answering task. As in Experiment 1, there wasan interaction between Content predictability and Experiment ($b = -132.56$, $SE = 36.08$, $t = -3.67$). But there was no interaction between Length predictability and Experiment ($b = -13.52$,

$SE = 67.79$, $t = -0.20$). Response times were influenced by Answer Agreement in the same way as in the individual experiment analyses ($b = -36.80$, $SE = 14.22$, $t = -2.59$). Together with the individual analyses, these results suggest that the effect of content predictability was stronger in the question-answering than button-pressing experiment. In other words, these results provide further evidence to suggest listeners used content predictions to prepare a verbal response, but not to predict the speaker's turn-end.

6. General discussion

In two pairs of experiments, we used button-press (Experiments 1a and 2a) and question-answering (Experiments 1b and 2b) tasks to investigate how interlocutors use prediction to achieve finely coordinated turn-taking. We contrasted two different hypotheses: (i) the early-planning hypothesis, which proposes that listeners use content predictions to prepare an early response but not to predict the speaker's turn-end (e.g., Levinson & Torreira, 2015), and (ii) the late-planning hypothesis, which proposes that listeners use content predictions (content-prediction hypothesis) and possibly length predictions (in number of words; length-prediction hypothesis) to determine the speaker's turn-end, and only begin preparation close to this moment (e.g., Sjerps & Meyer, 2015). In all experiments, we manipulated both the content (i.e., the predictability of the words of the speaker's turn) and length predictability (i.e., the predictability of the number of words needed to complete the turn) of simple *yes/no* questions.

There were no predictability effects on the precision of participants' button-presses or verbal responses (i.e., how closely participants responded to the speaker's turn-end), suggesting that listeners did not use linguistic information (either about content or length) to predict the speaker's turn-end. However, we did find effects of content predictability on response times in the question-answering tasks: Participants answered earlier when the final word(s) of the question were predictable (e.g., *Are dogs your favourite animal*?) rather than unpredictable (e.g., *Do you enjoy going to the supermarket*?). These results are consistent with findings from studies in language comprehension, which have shown that listeners can use the content of the speaker's utterance to predict how it continues (e.g., Altmann & Kamide, 1999), and suggest that listeners used such predictions to prepare their own response early during the speaker's turn.

Our findings are consistent with previous research that supports early planning during turn-taking (e.g., Barthel et al., 2016, 2017; Bögels et al., 2015) and suggest that listeners used content predictions to prepare their response early, but not to predict when they could launch articulation of this response. In contrast, our findings are inconsistent with the late-planning hypothesis, which suggests that listeners delay preparation until they know that they will soon have the opportunity to launch articulation. Specifically, Sjerps and Meyer (2015) found that listeners delayed preparation until near the end of the speaker's utterance. However, it may be that this discrepancy is due to their use of the dual-task paradigm: If participants had prepared a response early then they would have had to carry out three simultaneous tasks (i.e., comprehending the speaker's turn, preparing their own response, and finger tapping). Thus, their participants may have delayed preparation because they used cognitive resources to carry out an additional attention-demanding task, which is normally absent during conversation. Sjerps and Meyer addressed this issue in their second experiment, in which they found that participants looked towards to-be-named objects only shortly before producing their response. However, listeners may have given preference to looking for comprehension, and thus did not look earlier at the objects that they themselves had to name.

Our results are inconsistent with both the length-prediction hypothesis, which proposes that listeners predict the speaker's turn-end by predicting the length (in number of words) of the speaker's utterance, even when content is unpredictable (e.g., Magyari & De Ruiter, 2012), and the content-prediction hypothesis (Magyari et al., 2014), which instead suggests that listeners predict length only when content is predictable. However, there are a number of notable differences between our experiments and previous studies that have manipulated the content or length predictability of turns. First, neither Magyari and De Ruiter nor Magyari et al. included utterance duration as a control variable in their analyses. Duration was a strong predictor of response times in both of our button-press experiments (and those reported by De Ruiter et al., 2006): We found that questions longer in duration elicited less precise and earlier responses than those shorter in duration. Thus, it is possible that previous findings can be attributed to residual differences in duration, even if those studies matched the average duration of turns across conditions.

But other studies, which have fully controlled for duration, demonstrated turn-end prediction does play a role in turn-taking, and specifically that being able to understand the content of the speaker's utterance is important for determining the speaker's turn-end (e.g., De Ruiter et al., 2006; Riest, Jorschick, & De Ruiter, 2015). It is less clear, however, whether these studies demonstrate that the predictability of this content is important. In fact, Riest et al. found no difference between a condition in which participants could preview a transcript of the turn and one in which they were exposed to the turn for the first time. They interpreted this as evidence that speakers predicted the turn-end in both conditions, but it could also be interpreted as evidence that predictability does not affect how early participants respond in the button-press paradigm (there was no separate assessment of turn predictability, so it is difficult to determine how predictable the turns were when participants heard them for the first time).

Another difference between our study and previous ones is that our questions were produced by a pre-recorded speaker, while those in previous studies (e.g., De Ruiter et al., 2006) were taken from natural conversation. Thus, we may have failed to replicate their effects of content predictability because certain characteristics (i.e., changes in pitch, intonation, etc.) that are present in natural stimuli may have been absent in our recorded stimuli. We also note that both of our experiments used an explicit task, in which participants were encouraged to predict the speaker's turn-end (Experiments 1a and 2a) and answer quickly (Experiments 1b and 2b). But in natural conversation, listeners are unlikely to predict turn-ends explicitly or be aware of the explicit pressure to respond quickly. Nevertheless, these tasks allow us to tap into some of the mechanisms underlying coordination during turn-taking.

In sum, our results suggest that listeners can and do prepare their response early. Future research could explore what aspects of their response listeners prepare in advance. It is possible that they prepare the lexical content of their response and hold this response in an articulatory buffer until they can launch articulation (see Piai, Roelofs, Rommers, Dahlslätt, & Maris, 2015). But assuming that production and comprehension share resources (e.g., Segaert et al., 2012), how does the listener manage to prepare and buffer a response while comprehending the speaker's unfolding turn? If the listener can predict what the speaker is going to say, then it may matter less that they fully comprehend the speaker's unfolding turn because they have already comprehended enough of the utterance to predict the speaker's message and prepare a response. Although some comprehension must be necessary, in case any prediction is inaccurate, the listener may manage the capacity demands of concurrent production and comprehension by allocating fewer resources to comprehending their interlocutor's turn. Further research could investigate this issue.

Regardless, listeners must still ensure they articulate their pre-prepared response at the appropriate moment. Listeners may rely on a number of mechanisms to do so (e.g., Bögels & Levinson, 2017; see also Wilson & Wilson, 2005). One possibility is that listeners launch articulation of their response reactively, after they have encountered one or more turn-final cues (e.g., falling boundary tone). This more reactive strategy (Duncan, 1972; Heldner & Edlund, 2010) may still be

compatible with short inter-turn intervals because launching articulation does not take as long as preparing a response from scratch (articulation takes around 145 ms; Indefrey & Levelt, 2004). Note that listeners are likely to be sensitive to a collection of such cues (e.g., Bögels & Torreira, 2015), and could use multiple cues to determine points of possible turn completion.

Importantly, these cues could work in parallel with a turn-end prediction mechanism, and this may well explain why turn-final cues are not necessarily perfect predictors of a speaker switch (e.g., Gravano & Hirschberg, 2011). For example, in instances when the listener is able to predict that the speaker will soon reach the end of their turn, they may allocate more processing resources to paying attention to possible turn-final cues, so that they are quicker to launch articulation when the speaker displays such cues. But in instances when such predictions are not possible, the listener may process such cues much less efficiently, resulting in longer gaps between turns.

In conclusion, we have shown that participants in a question-answering task were sensitive to the predictability of final words in questions: Participants answered earlier when such words were predictable rather than unpredictable . However, we found no evidence that participants used their ability to predict the final word to estimate when the speaker's turn would end. Thus, we conclude that content predictability helps listeners prepare a verbal response early, but does not help them determine when they should launch articulation of this response.

Acknowledgements

Ruth Corps is supported by the Economic and Social Research Council [grant number ES/J500136/1]. Chiara Gambi is supported by a Leverhulme Research Project Grant [RPG-2014-253] to Martin Pickering and Hugh Rabagliati. We thank Alice Turk for invaluable advice on analyzing the pitch contours of our stimuli, Max Dunn for acting as second coder for both experiments, and Nigel Corps for patiently recording stimuli audio.

Appendix A

Linear mixed effects model output for the response time analysis of all four individual experiments (see Table A1).

Table A1
Linear mixed effects model output for the analysis of response times in all four experiments. RE var = Random effects variance; (p) stands for random effects by participants; (i) stands for random effects by items. All predictors are defined in the Data Analysis section for each experiment.

	Experiment 1a				Experiment 1b				Experiment 2a				Experiment 2b			
Predictor	Coeff.	SE	t	RE var	Coeff.	SE	t	RE var	Coeff.	SE	t	RE var	Coeff.	SE	t	RE var
Intercept	−136.21	55.53	−2.45	(p) 87806 (i) 13,859	380.26	57.60	6.60	(p) 93329 (i) 18687	−117.23	51.58	−2.27	(p) 71600 (i) 19603	483.65	60.19	8.04	(p) 101914 (i) 12201
Duration	−152.17	15.04	−10.12	–	−72.88	17.25	−4.23	–	−124.99	13.92	−8.98	–	−59.85	15.67	−3.82	–
Answer	–	–	–	–	−21.86	16.46	−1.33	–	–	–	–	–	−143.43	19.92	−7.20	–
Agreement	–	–	–	–	−55.21	15.17	−3.64	–	–	–	–	–	−35.81	15.66	−2.29	–
Content	−28.31	29.00	−0.97	(p) 3498	−153.01	34.08	−4.49	(p) 5909	0.39	35.60	0.01	(p) 0	−81.68	39.07	−2.09	(p) 54
Length	−19.25	34.00	−0.57	(p) 10589	10.89	33.25	0.33	(p) 1171	18.75	41.74	0.45	(p) 3064	28.19	36.81	0.77	(p) 0
Content * length	−8.57	50.15	−0.17	(p) 0.00	−110.21	63.75	−1.73	(p) 17340	–	–	–	–	–	–	–	–

Appendix B

Bayesian mixed model output for the precision analysis of all four individual experiments (see Table B1).

Table B1
Model output for precision analyses in all experiments. Estimates are on the log scale (linear estimates in-text). We report fixed effects and the variance (var) explained by random effects (RE). (f) = fixed effect, (p) = RE by participants, (i) = RE by items.

(Exp. 1a) Predictor	Estimate	SE	CrIs	Effective sample
Intercept	(f) −0.82; (p) 0.72; (i) 0.41	(f) 0.14; (p) 0.11; (i) 0.04	(f) −1.08, −0.54; (p) 0.54, 0.97; (i) 0.34, 0.49	(f) 347; (p) 873; (i) 1283
Shape_Intercept	(f) 0.40; (p) 0.34; (i) 0.23	(f) 0.07; (p) 0.05; (i) 0.02	(f) 0.27, 0.53; (p) 0.26, 0.45; (i) 0.34, 0.49	(f) 629; (p) 932; (i) 1318
Duration	(f) 0.17	(f) 0.05	(f) 0.07, 0.27	(f) 1647
Shape_Duration	(f) −0.15	(f) 0.03	(f) −0.21, 0.09	(f) 1698
Content	(f) −0.03; (p) 0.18	(f) 0.10; (p) 0.06	(f) −0.22, 0.16; (p) 0.05, 0.31	(f) 1384; (p) 544
Shape_Content	(f) 0.00; (p) 0.11	(f) 0.06; (p) 0.04	(f) −0.11, 0.10; (p) 0.02, 0.19	(f) 1612; (p) 944
Length	(f) −0.04; (p) 0.27	(f) 0.11; (p) 0.06	(f) −0.25, 0.17; (p) 0.16, 0.40	(f) 1617; (p) 1577
Shape_Length	(f) −0.07; (p) 0.06	(f) 0.06; (p) 0.04	(f) −0.19, 0.04; (p) 0.00, 0.13	(f) 1645; (p) 1004
Content * Length	(f) −0.24; (p) 0.25	(f) 0.18; (p) 0.12	(f) −0.60, 0.10; (p) 0.02, 0.50	(f) 1560; (p) 831
Shape_Content * Length	(f) −0.04; (p) 0.07	(f) 0.10; (p) 0.06	(f) −0.23, 0.16; (p) 0.00, 0.21	(f) 1785; (p) 1578
(Exp. 1b) Predictor	Estimate	SE	CrIs	Effective sample
Intercept	(f) −0.44; (p) 0.33; (i) 0.25	(f) 0.07; (p) 0.05; (i) 0.03	(f) −0.58, 0.30; (p) 0.15, 0.44; (i) 0.20, 0.31	(f) 443; (p) 569; (i) 1096
Shape_Intercept	(f) 0.19; (p); 0.19; (i) 0.06	(f) 0.04; (p) 0.03; (i) 0.03	(f) 0.10, 0.26; (p) 0.14, 0.26; (i) 0.00, 0.11	(f) 571; (p) 954; (i) 469
Duration	(f) 0.15	(f) 0.04	(f) 0.08, 0.22	(f) 1405
Shape_Duration	(f) 0.04	(f) −0.02	(f) −0.07, 0.00	(f) 2675
Answer	(f) −0.01	(f) 0.04	(f) −0.10, 0.07	(f) 3200
Shape_Answer	(f) 0.02	(f) 0.03	(f) −0.04, 0.07	(f) 3200
Agreement	(f) −0.06	(f) 0.03	(f) −0.13, 0.00	(f) 1181
Shape_Agreement	(f) 0.02	(f) 0.02	(f) −0.01, 0.05	(f) 2713
Content	(f) 0.05; (p) 0.52	(f) 0.12; (p) 0.08	(f) −0.17, 0.28; (p) 0.38, 0.71	(f) 587; (p) 676

(continued on next page)

Table B1 (*continued*)

(Exp. 1a) Predictor	Estimate	SE	CrIs	Effective sample
Shape_Content	(f) 0.10; (p) 0.25	(f) 0.06; (p) 0.05	(f) −0.02, 0.21; (p) 0.16, 0.35	(f) 797; (p) 1227
Length	(f) 0.02; (p) 0.21	(f) 0.08; (p) 0.06	(f) −0.14, 0.18; (p) 0.10, 0.34	(f) 1281; (p) 1327
Shape_Length	(f) −0.01; (p) 0.05	(f) 0.04; (p) 0.04	(f) −0.08, 0.06; (p) 0.00, 0.13	(f) 2363; (p) 1231
Content*Length	(f) −0.19; (p) 0.33	(f) 0.14; (p) 0.13	(f) −0.48, 0.09, (p) 0.07, 0.58	(f) 1344; (p) 614
Shape_Content*Length	(f) 0.01; (p) 0.26	(f) 0.08; (p) 0.09	(f) −0.16, 0.17; (p) 0.07, 0.45	(f) 1576; (p) 720
(Exp. 2a) Predictor	Estimate	SE	CrIs	ES
Intercept	(f) −0.74; (p) 0.72, (i) 0.49	(f) 0.14; (p) 0.10; (i) 0.05	(f) −1.02, 0.47; (p) 0.55, 0.94; (i) 0.41, 0.60	(f) 456; (p) 1032; (i) 1430
Shape_Intercept	(f) 0.44; (p) 0.31, (i) 0.25	(f) 0.07; (p) 0.05; (i) 0.03	(f) 0.31, 0.57; (p) 0.24, 0.41; (i) 0.21, 0.31	(f) 876; (p) 1070; (i) 1697
Duration	(f) 0.23	(f) 0.05	(f) 0.13, 0.32	(f) 1739
Shape_Duration	(f) −0.10	(f) 0.03	(f) −0.16, −0.04	(f) 2082
Content	(f) 0.23; (p) 0.28	(f) 0.15; (p) 0.08	(f) −0.07, 0.53; (p) 0.11, 0.45	(f) 1292; (p) 802
Shape_Content	(f) 0.01; (p) 0.10	(f) 0.08; (p) 0.05	(f) −0.14, 0.17; (p) 0.01, 0.20	(f) 1327; (p) 802
Length	(f) 0.10; (p) 0.23	(f) 0.26; (p) 0.14	(f) −0.41, 0.62; (p) 0.01, 0.54	(f) 1330; (p) 1010
Shape_Length	(f) 0.20; (p) 0.18	(f) 0.14; (p) 0.10	(f) −0.14, 0.17; (p) 0.01, 0.37	(f) 1596; (p) 993
(Exp. 2b) Predictor	Estimate	SE	CrIs	ES
Intercept	(f) −0.71; (p) 0.58; (i) 0.23	(f) 0.11; (p) 0.08; (i) 0.03	(f) −0.94, −0.50; (p) 0.44, 0.76; (i) 0.17, 0.30	(f) 238; (p) 588; (i) 1279
Shape_Intercept	(f) 0.28; (p) 0.58; (i) 0.10	(f) 0.06; (p) 0.08; (i) 0.02	(f) 0.16, 0.41; (p) 0.24, 0.41; (i) 0.06, 0.14	(f) 335; (p) 677; (i) 1066
Duration	(f) 0.11	(f) 0.04	(f) 0.04, 0.18	(f) 1434
Shape_Duration	(f) 0.00	(f) 0.02	(f) −0.04, 0.04	(f) 1955
Answer	(f) 0.00	(f) 0.05	(f) −0.09, 0.09	(f) 3200
Shape_Answer	(f) 0.13	(f) 0.03	(f) 0.06, 0.19	(f) 3200
Agreement	(f) 0.02	(f) 0.04	(f) −0.05, 0.09	(f) 1689
Shape_Agreement	(f) 0.02	(f) 0.02	(f) −0.01, 0.06	(f) 2090
Content	(f) 0.12; (p) 0.35	(f) 0.10; (p) 0.08	(f) −0.08, 0.33; (p) 0.20, 0.52	(f) 1168; (p) 768
Shape_Content	(f) −0.17; (p) 0.23	(f) 0.07; (p) 0.05	(f) −0.30, −0.04; (p) 0.13, 0.35	(f) 1087; (p) 1194
Length	(f) 0.01; (p) 0.18	(f) 0.15; (p) 0.12	(f) −0.28, 0.31; (p) 0.01, 0.44	(f) 1328; (p) 942
Shape_Length	(f) −0.08; (p) 0.18	(f) 0.08; (p) 0.12	(f) −0.25, 0.08; (p) 0.00, 0.25	(f) 2268; (p) 1508

Appendix C

Lists of stimuli used in all four experiments. Completions chosen from the pre-test are *italicized* (see Tables C1 and C2).

Table C.1
Stimuli used in Experiments 1a and 1b.

Content predictability	Length predictability	Simulus
Predictable	Single	Have you passed your driving *test*?
		Do you celebrate Christmas on the twenty fifth of *December*?
		Can most fish breathe under *water*?
		To cook a cake, will I need to put it in the *oven*?
		Is red your favourite color?
		If I wear sunglasses, will they keep the sun out of my *eyes*?
		Do dogs have four legs?
		Have you ever forgotten your keys and been locked out of the *house*?
		Are pandas the colors black and *white*?
		Have you ever seen a spider with less than eight *legs*?
		Is David Cameron the prime *minister*?
		At University, are you a psychology *student*?
		Do you regularly borrow books from the *library*?
		Is a piano a musical *instrument*?
		Should I go to the zoo if I want to see a lot of different *animals*?
		Is a baby kangaroo called a *joey*?
		Do you think surfers are scared of being bitten by a *shark*?
		Do you think most students will pass their *exams*?
		Is a Dalmatian dog black and *white*?
		While eating, have you ever accidentally bitten your *tongue*?
		To pay for your tuition fees, did you have to take out a student *loan*?
		Are dogs your favourite *animal*?
		Is Andy Murray a tennis *player*?
		Either at university or school, have you ever failed an *exam*?
		Should I buy my friend a present for her *birthday*?
		Did you wake up before 9o'clock this *morning*?
		To keep the sun out of my eyes, should I wear *sunglasses*?
		Is spring your favourite season of the *year*?
Predictable	Varied	If my feet are cold, should I put on *some socks*?
		To pay for your studies, did you take *out a loan*?
		Have you ever forgotten about an assignment and handed it in *having done it on the way to class?*
		Did The Titanic sink after *it hit an iceberg*?
		Have you ever taken the blame even though you weren't *at fault*?
		When eating, do you cut your food with *a knife and fork*?

(*continued on next page*)

Table C.1 (*continued*)

Content predictability	Length predictability	Simulus
		Do you see your parents *at the weekend*?
		When you go to restaurants, do you leave *a ten percent tip*?
		To communicate with others, do deaf people have to *watch and lip read*?
		Is summer your favourite *season of the year?*
		Do people become werewolves when *they see a full moon*?
		I don't have a watch, so could you *tell me the time please?*
		Have you ever been to a casino and lost *a lot of money*?
		Have you ever broken your leg and used *a crutch*?
		There are no clean plates left, so could you *wash some up*?
		When it is cold outside, should I wear a scarf to keep *myself warm*?
		Does the dentist always tell you to brush *your teeth more*?
		Should I make an optician's appointment if I think I need *new glasses?*
		As well as being a student, do you also have a *part time job*?
		This coffee is too hot, so before I drink it should I let it *cool down a little*?
		In your tea, would you like *milk and sugar*?
		There's a hole in my sock, so could you get *me new ones*?
		The dishes need cleaning, so could you *help me clean them*?
		I'm struggling to see, so should I get *a pair of glasses*?
		During the night, have you ever woken up after *a nightmare*?
		I'm going to cut my hair myself, so can you get me *a pair of scissors*?
		In the past, have you ever been late when *you had an appointment*?
		After an argument, have you ever slammed *a door shut*?
		My toaster is broken, so could you *fix it please*?
Unpredictable	Single	Have you ever visited the city of *Paris*?
		Are you in your third year of *marriage*?
		Are there a lot of females in your *apartment*?
		Do you enjoy going to the *supermarket*?
		Today, do you think I should wear a *tie*?
		Do you need to go to the supermarket to buy some *crisps*?
		In the past, have you had a lot of different *cars*?
		Would you like to see a picture of my *spider*?
		Have you ever injured your *eye*?
		Have you ever seen a wild *bear*?
		Do you like to eat a lot of *crisps*?
		During the summer, do you like spending time at the *library*?
		Do you live far away from the *beach*?
		Are you really looking forward to *tonight*?
		Would you like to take an evening *class*?
		Is an orange the same color as a *tiger*?
		If you could get a pet, would you like to get a *tortoise*?
		Should I buy a new suit for my *dance*?
		Do you have any lectures on *mathematics?*
		Are you very scared of *ghosts*?
		Do you think you are good at *singing*?
		Do most people have two *siblings*?
		Do you have a big *house*?
		Have you ever watched a game of *cricket*?
		Have you ever been on a *plane*?
		Would you like to go for a walk in the *forest*?
		Have you ever played a game of *poker*?
		Have you ever broken your *phone*?
		Are you doing anything *important*?
Unpredictable	Varied	Are a lot of your friends *in the same classes*?
		Do you spend a lot of your time *with friends*?
		Is your favourite book *the Hunger Games*?
		Did you do anything you enjoyed and *didn't expect to*?
		IS your favourite film called *The Imitation Game*?
		If I want to stay warm during the winter, should I put on *multiple layers*?
		Do most students finish their *studies after four years*?
		Have you ever been to London to visit *the Imperial War Museum?*
		In a few years, would you like to move to *the mountains*?
		Is your favourite TV show *The Great British Bake Off*?
		Have you ever been to the cinema to watch *the Lion King*?
		Are you going to celebrate *New Year in Edinburgh*?
		During your lunch break, would you like to *grab a bite to eat*?
		Have you ever read a book by *Suzanne Collins*?
		Have you ever read a book called *Blood Diamond*?
		Do you have a lot of *free time*?
		Tomorrow morning, would you like to eat *your breakfast in bed*?
		Should I call the police if there is someone *acting suspiciously*?
		During the evening, do you *eat dinner*?
		When studying, do you like to work *in the library*?
		Next week, would you like to have dinner at *that new restaurant*?
		In your opinion, do you think you are *a nice person*?
		Tomorrow afternoon, would you like to *play football*?

(*continued on next page*)

Table C.1 (*continued*)

Content predictability	Length predictability	Simulus
		When it's raining, should I take an umbrella to *keep myself dry*? Have you ever been to *the zoo*? At University, are you in *lectures a lot*? Would you like to have a *glass of wine*? In your spare time, have you ever listened to *heavy metal*? In the past, have you ever tried *to ice skate*?

Table C2
Stimuli used in Experiments 2a and 2b.

Content predictability	Length predictability	Stimulus
Predictable	Single	Have you passed your driving *test*? Can most fish breathe under *water*? Have you ever read a Shakespeare *play*? Is red your favourite *color*? Have you ever forgotten your keys and been locked out of the *house*? Have you ever seen a spider with less than eight *legs*? At University, are you a psychology *student*? Do you regularly borrow books from the *library*? Should I go to the zoo if I want to see a lot of different *animals*? Do you think surfers are scared of being bitten by a *shark*? Do you think most students will pass their *exams*? Is a Dalmatian dog black and *white*? When meeting someone new, do you shake their *hand*? To pay for your studies, did you take out a *loan*? Are dogs your favourite *animal*? Either at university or school, have you ever failed an *exam*? Did you wake up before 9o'clock this *morning*? To keep the sun out of my eyes, should I wear *sunglasses*? Is spring your favourite season of the *year*? Do genies grant *wishes*? Does the Queen live in Buckingham *Palace*? Have you ever dyed your *hair*? Do you enjoy watching horror *movies*? To grow, do plants need *water*? Can you type without looking at the *keyboard*? Is a Unicorn a horse with a *horn*? Do you wash your hair every *day*? To pay for your tuition fees, did you have to take out a student *loan*?
Unpredictable	Single	Have you ever visited the city of *Paris*? Today, do you think I should wear a *tie*? Do you need to go to the supermarket to buy some *crisps*? In the past, have you had a lot of different *cars*? Would you like to see a picture of my *spider*? Have you ever injured your *eye*? Have you ever seen a wild *bear*? During the summer, do you like spending time at the *library*? Do you live far away from the *beach*? If you could get a pet, would you like to get a *tortoise*? Should I buy a new suit for my *dance*? Do you live in a house with other *animals*? Are you happy with your *grades*? Do most people have two *siblings*? Have you got a big *house*? Would you like to go for a walk in the *forest*? Have you ever played a game of *poker*? Have you ever broken your *phone*? Do you participate in a lot of *experiments*? Do you have two *homes*? Have you ever had to visit the hospital after injuring your *body*? Are you allergic to *fish*? In your opinion, do you think you are a good *cook*? Is chocolate your favourite *treat*? Are you in a *society*?
Unpredictable	Varied	When you're studying, do you like to work *silently*? Should I call the police if there is someone *suspicious*? In a few years, would you like to move to *Japan*? Do you spend a lot of your time *revising*? Before starting your studies at University, did you take a *loan*? When it's raining, should I take an umbrella to *university*? Is your favourite book *religious*? Did you do anything you enjoyed *today*?

(*continued on next page*)

Table C2 (*continued*)

Content predictability	Length predictability	Stimulus
		Have you ever read a book called *Twilight*?
		Have you ever read a book by *candlelight*?
		Have you ever been to the cinema to watch *Wolverine*?
		Have you ever been to London to visit *family*?
		Next week, would you like to have dinner at *six*?
		Have you ever been to *Greece*?
		At university, are you in *psychology*?
		Tomorrow morning, would you like to eat *earlier*?
		In your spare time, have you ever listened to *lectures*?
		In the past, have you ever tried *octopus*?
		Is your favourite film *recent*?
		During the evening, do you *relax*?
		Would you like to learn *Mandarin*?
		Would you like to climb *rocks*?
		Can you play *solitaire*?
		Do you get nervous when speaking *publicly*?
		Have you ever been admitted to hospital to have *surgery*?
		Would you like to have a *snack*?
		Have you ever taken the blame even though you weren't *responsible*?
		After an argument, have you ever slammed a door *shut*?

Appendix D

Ranges of pre-test values for the content and length predictability of stimuli used in all four experiments (see Tables D1 and D2).

Table D1
Ranges of the measures of content predictability (question fragment LSA, completion LSA, completion content cloze, and question fragment entropy) and length predictability (completion length variance, completion length cloze) for stimuli in Experiments 1a and 1b.

Content	Length		Average variance of completion length	Completion length cloze[a]	Question fragment LSA[b]	Completion LSA[c]	Completion content cloze[d]	Question fragment entropy[e]
Predictable	Single	Min	0	90%	0.61	0.76	70%	0
		Max	0.12	100%	1	1	100%	1.14
	Varied	Min	0.32	3%	0.45	0.96	–	–
		Max	3.31	53%	0.93	0.26	–	–
Unpredictable	Single	Min	0	67%	0.58	0.04	3%	1.71
		Max	0.30	100%	0.18	0.37	9%	4.18
	Varied	Min	0.31	0%	0.13	0.05	–	–
		Max	1.93	66%	0.60	0.47	–	–

[a] Percentage of participants who provided the word length of the selected completion used in the main experiment (a single word in the single conditions; multiple words in the varied conditions) as a continuation in the cloze task.
[b] Average over all completion comparisons for that particular fragment.
[c] Average over comparisons between the selected completion and all other completion.
[d] Cloze percentages of the selected completion. If cloze percentage is higher, then participants converged on a completion.
[e] Entropy of question fragments presented to participants in the cloze task. If entropy is lower, then participants converged on a completion.

Table D2
Ranges of the measures of content and length predictability for stimuli in Experiments 2a and 2b.

Condition		Average variance of completion length	Completion length cloze[a]	Question fragment LSA[b]	Completion LSA[c]	Completion content cloze[d]	Question fragment entropy[e]
Predictable single	Min	0	90%	0.61	0.76	67%	0
	Max	0.12	100%	1	1	100%	1.14
Unpredictable single	Min	0	90%	0.17	0.04	3%	1.66
	Max	0.15	100%	0.59	0.37	12%	4.18
Unpredictable Varied	Min	0.26	9%	0.13	0.02	–	–
	Max	2.94	73%	0.49	0.58	–	–

[a] Percentage of participants who provided the word length of the selected completion used in the main experiment (a single word in the single conditions; multiple words in the varied conditions) as a continuation in the cloze task.
[b] Average over all completion comparisons for that particular fragment.
[c] Average over comparisons between the selected completion and all other completion.
[d] Cloze percentages of the selected completion. If cloze percentage is higher, then participants converged on a completion.
[e] Entropy of question fragments presented to participants in the cloze task. If entropy is lower, then participants converged on a completion.

Appendix E. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2018.01.015.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Babiloni, C., Carducci, F., Cincotti, F., Rossini, P. M., Neuper, C., Pfurtscheler, G., & Babiloni, F. (1999). Human movement-related potentials vs desynchronization of EEG alpha rhythm: a high-resolution EEG study. *Neuroimage, 10*, 658–665.

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*, 255–278.

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*. http://dx.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*. http://dx.doi.org/10.3389/fpsyg.2016.01858.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Lme4: Linear mixed-effects models using "Eigein" and S4 (R package version 1.1-12). Retrieved from http://CRAN.R-project.org/package=lme4.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology, 3*, 255–309.

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

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

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends. *Journal of Phonetics, 52*, 46–57.

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*, 295–311.

Bürkner, P. -C. (2017). *Brms: Bayesian Regression Models using Stan* (R package version 1.6.1). Retrieved from https://CRAN.R-project.org/package=brms.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

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*, 886–889.

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

Cutler, A., & Pearson, M. (1986). On the analysis of prosodic turn-taking cues. In C. Johns-Lewis (Ed.). *Intonation and discourse* (pp. 139–155). London: Croom Helm.

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*. http://dx.doi.org/10.3389/fpsyg.2014.00781.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversation. *Journal of Personality and Social Psychology, 23*, 283–292.

Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology, 6*. http://dx.doi.org/10.3389/fpsyg.2015.00751.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language, 25*, 601–634.

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

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

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*, 363–374.

Leone, F. C., Nelson, L. S., & Nottingham, R. B. (1961). The folded normal distribution. *Technometrics, 3*, 543–550.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6*. http://dx.doi.org/10.3389/fpsyg.2015.0073.

Magyari, L., Bastiaansen, M. C. M., De Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind speed of response in conversation. *Journal of Cognitive Neuroscience, 26*, 2530–2539.

Magyari, L., & De Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology, 3*. http://dx.doi.org/10.3389/fpsyg.2012.00376.

Magyari, L., De Ruiter, J. P., & Levinson, S. C. (2017). Temporal preparation for speaking in question-answer sequences. *Frontiers in Psychology, 8*. http://dx.doi.org/10.3389/fpsyg.2017.00211.

Näätänen, R. (1971). Non-aging fore-periods and simple reaction time. *Acta Psychologica, 35*, 316–327.

Piai, V., Roelofs, A., Rommers, J., Dahlslatt, K., & Maris, E. (2015). Withholding planned speech is reflected in synchronized beta-band oscillations. *Frontiers in Human Neuroscience, 9*. http://dx.doi.org/10.3389/fnhum.2015.00549.

Pinder, J. E., Wiener, J. G., & Smith, M. H. (1978). The Weibull distribution: A new method for summarizing survivorship data. *Ecology, 59*, 175–179.

Riest, C., Jorschick, A. B., & De Ruiter, J. P. (2015). Anticipation in turn-taking: Mechanisms and information sources. *Frontiers in Psychology, 6*. http://dx.doi.org/10.3389/fpsyg.2015.00089.

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

Sanders, A. F. (1966). Expectancy: application and measurement. *Acta Psychologica, 25*, 293–313.

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension. *Cerebral Cortex, 22*, 1662–1670.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 76*, 379–423.

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.

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

Strömbergsson, S., Hjalmarsson, A., Edlund, J., & House, D. (2013). Timing responses to questions in dialogue. In F. Bimbot. INTERSPEECH 2013. *Paper presented at 14th annual conference of the international speech communication association* (pp. 2584–2588). Lyon: International Speech and Communication Association.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs during reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 443–467.

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