AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2021, Vol. 150, No. 9, 1772–1799
https://doi.org/10.1037/xge0001012

Planning When to Say: Dissociating Cue Use in Utterance Initiation Using Cross-Validation

Laurel Brehm[1] and Antje S. Meyer[1, 2]

[1] Department of Psychology of Language, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
[2] Donders Institute for Brain, Cognition, and Behavior, Radboud University Nijmegen

In conversation, turns follow each other with minimal gaps. To achieve this, speakers must launch their utterances shortly before the predicted end of the partner's turn. We examined the relative importance of cues to partner utterance content and partner utterance length for launching coordinated speech. In three experiments, Dutch adult participants had to produce prepared utterances (e.g., *vier*, "four") immediately after a recording of a confederate's utterance (*zeven*, "seven"). To assess the role of corepresenting content versus attending to speech cues in launching coordinated utterances, we varied whether the participant could see the stimulus being named by the confederate, the confederate prompt's length, and whether within a block of trials, the confederate prompt's length was predictable. We measured how these factors affected the gap between turns and the participants' allocation of visual attention while preparing to speak. Using a machine-learning technique, model selection by *k*-fold cross-validation, we found that gaps were most strongly predicted by cues from the confederate speech signal, though some benefit was also conferred by seeing the confederate's stimulus. This shows that, at least in a simple laboratory task, speakers rely more on cues in the partner's speech than corepresentation of their utterance content.

Keywords: language production, picture naming, turn-taking, statistical modeling

Supplemental materials: https://doi.org/10.1037/xge0001012.supp

To speak, one needs to plan what to say and time when to say it. Planning what to say requires formulating an utterance that is sensible and grammatical, while planning when to say it—utterance *launch*—typically requires estimating when a currently speaking interlocutor will finish talking. Existing research has established that turn-taking in conversation tends to be remarkably efficient and effective, but less is known about the cognitive mechanisms that allow speakers to coordinate with conversation partners over time. In the current work, we ask what information speakers track to launch speech successfully: does producing coordinated speech require mentally representing what the conversation partner will say, or can it be done by simply tracking the cues marking the onset and offset of partner utterances? In three experiments using a novel application of a machine learning technique, model selection by cross-validation, we contrasted the role of these factors in utterance launching. We show that while coordinated launch is improved by corepresenting upcoming content, it can still be done successfully by attending to when the confederate prompt starts and ends.

Existing literature demonstrates that individuals are good at coordinating speech in natural conversations. The field of Conversation Analysis highlights this best: individuals are skilled at taking turns, forming a coherent discourse and minimizing interruptions or long gaps in the flow of speech (for foundational work: Sacks et al., 1974; for a recent handbook: Sidnell & Stivers, 2012). Quantitative studies of conversational corpora in a variety of languages underscore how successful this coordination is. Though there is variability in the duration of gaps between speakers' turns (henceforth *turn gap*), estimates of modal turn gaps range from 100 to 400 ms (e.g., Heldner & Edlund, 2010; Levinson & Torreira, 2015; Weilhammer & Rabold, 2003), and a corpus study of polar (yes/no) questions suggests an overall mode across 10 languages around 200 ms (Stivers et al., 2009).

One prevailing psycholinguistic model of conversation, proposed by Pickering and Garrod (2004, 2013) suggests that to create tight coordination with their interlocutors, individuals mentally represent the contents of the partner's utterance in addition to their own plans to speak. This framework considers language to be a form of joint action, similar to playing a piano duet or collaboratively lifting a piano (e.g., Clark, 1996; Knoblich et al., 2011; Vesper et al., 2017, 2010). Theories of joint action assume that individuals need to mentally represent their own action plans and

This article was published Online First March 18, 2021.

Laurel Brehm https://orcid.org/0000-0003-0424-8735

Portions of these data were presented at ESCOP 2019 and Psychonomics 2019. We thank the MPI-NL Simulation and Modelling User Group and the Department of Psychology of Language for helpful feedback, and Annelies van Wijngaarden for providing voice recordings. Scripts and data are archived at the Open Science Framework (available at https://osf.io/8u647/).

Correspondence concerning this article should be addressed to Laurel Brehm, Department of Psychology of Language, Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, the Netherlands. Email: laurel.brehm@mpi.nl

corepresent the plans of other participants performing the task. Thus, each individual needs to plan what they will do and predict what their partner will do. This is deemed necessary for the precise temporal coordination of actions. Corepresenting a partner's utterance in a conversation as deeply as one's own utterance—the strongest prediction of this framework—would mean that to coordinate over time, interlocutors would form precise predictions of their own and their partners' upcoming content and recruit these to plan an appropriate response and have the plan ready to be launched at a pragmatically appropriate time, affording a smooth dialogue (e.g., Levinson & Torreira, 2015).

However, recent experimental evidence in simple laboratory tasks suggests that speakers form a sparse—not full—corepresentation of the partner utterance (e.g., Brehm et al., 2019; Gambi et al., 2015; Hoedemaker & Meyer, 2019). In these studies, participants represented whether or not their partner was about to speak, but did not represent what the partner would say. This was the case even when this information could readily be obtained by looking at the picture being described by the partner. These studies show that in experiments in which participants' primary task was to launch an utterance (naming a picture) at the right time and where this utterance was not contingent on the prior utterance, full corepresentation was not required. While this scenario clearly differs from natural conversation, it raises an important question: how can launching be done without corepresentation?

An obvious alternative is that instead of using corepresentations to time speech onset, speakers simply process the actual speech and gestures of their partner and launch their utterance when cues indicate that the current utterance is about to end. This proposal is also based on several earlier lines of work (for review: Garrod & Pickering, 2015). There is, first, a substantial literature on the linguistic and para-linguistic cues that signal upcoming ends of turns and the willingness of the current speaker to yield the floor. These cues include gestures (Duncan, 1972; ten Bosch et al., 2005), prosodic properties of utterances (Bögels & Torreira, 2015; Cutler & Pearson, 1985; Duncan, 1972; Gravano & Hirschberg, 2011; Grosjean & Hirt, 1996; Schaffer, 1983), and syntactic or lexical markers (de Ruiter et al., 2006; Duncan, 1972; Magyari & de Ruiter, 2012). While many cues are correlated with occurrence of ends of turns, none appear to be a uniquely strong predictor (see Bögels and Torreira, 2015; de Ruiter, et al., 2006; Gravano & Hirschberg, 2011). In actual conversations, listeners most likely rely on a combination of cues.

Second, a number of studies have investigated which cues listeners use to identify the ends of turns and respond to them. Some of these studies used metalinguistic tasks, asking participants to press a button at anticipated utterance end or to estimate how many words would be used to continue a sentence (e.g., Corps et al., 2018, 2020; de Ruiter et al., 2006; Grosjean & Hirt, 1996; Magyari & de Ruiter, 2012). Other studies have used highly constrained production paradigms (e.g., Barthel et al., 2017, 2016). The latencies obtained in these studies (with participants often responding before the end of the utterance) are sufficiently fast to indicate that participants typically predict rather than detect ends of turn and use a number of cues to do so including lexical content, syntax, and prosody.

This literature also points to ways that cues used for launching and planning speech could be dissociable. Barthel et al. (2017) used a list completion paradigm where a confederate first named a variable number of objects and the participant then had to name any remaining objects on the screen. The authors found that presence or absence of a lexical cue (the word *and*, indicating that the next word would be the final object name) affected when participants began to look at the first object they had to name, implying an effect on the onset of planning. Presence or absence of a late prosodic cue (a boundary tone on the last word) did not affect gaze, but shortened response time, suggesting it might have impacted when participants launched their utterances. Combined, these results show the importance of speech cues for response launching but not response planning.

In two recent articles, Corps and colleagues (2018, 2020) used a combination of methodologies to isolate turn end prediction from response planning. In these studies, participants heard polar questions, such as "Are dogs your favorite animal?" or "At university, do you study math?" Participants either answered the questions or indicated with a button-press when they believed the utterance would end. This allowed the authors to dissociate what affects preparing simple utterances, versus what affects prediction of turn end, a precondition for timely launching. For both measures (button press and speech onset), they recorded the response latencies and response precision (the absolute value of the time interval between response onset and the end of the question).

Corps et al. (2018) manipulated the predictability of the cue to the answer ("animal" vs. "math" in the above example) and the length of the question. They found that the predictability of the cue affected latencies to answer questions, but did not affect the precision of indicating via button-press when an utterance would end. However, longer utterances diminished the precision of button-press judgments. The authors suggest that this pattern emerged because speakers used the content of the partner utterance to prepare their response early, but not to plan when to launch. This implies that corepresentation of utterance content may not be necessary for utterance launch.

Corps et al. (2020) varied the global (sentence-level) speech rate of the questions and the duration (local speech rate) of the utterance-final word in a question-answering paradigm. Participants were sensitive to both speech rate manipulations, with responses being faster (measured from the onset of the final word in the question) when the global or local rate was fast than when it was slow. Participants also responded faster when the utterance-final word was predictable, though, this did not interact with the local speech rate when the two were manipulated in tandem. The implication is that prior utterance length affected response latencies more than content predictability did, which is consistent with the earlier finding that the timing of prior utterances impacts launch time, but corepresented content may not. However, as the authors acknowledge, the response latencies in this task were much longer than the turn gaps typically found in conversation. An obvious account is that the time participants needed to plan and launch their utterances extended until after the end of question. More generally, while the authors' interpretation of the additive effects of predictability and speech rate is plausible, the study did not clearly separate the processes needed for response planning and launching.

In the present study, we used a novel paradigm to isolate launching from planning and to examine the relative importance of the cues speakers use in deciding when to launch an utterance. On each trial, two stimuli had to be named—these were numerals in

Experiment 1 and pictures in Experiments 2 and 3. The first stimulus was a prompt named by a prerecorded confederate that had a variable onset time and variable duration, reflecting the time it took the confederate to name pictures in isolation; the second stimulus was named by the participant. The participants' task was to produce their utterance as soon as the confederate stopped talking. Both stimuli appeared at trial onset, with the mean prompt offset across experiments of 1,241 ms allowing participants ample time to prepare their utterance. The two utterances were independent in terms of content, with confederate and participant naming different pictures. The focus on the timely launching of utterances and the lack of semantic cohesion between turns are important deviations from natural conversation. However, by making participants' goal the timely launching of speech, allowing more than enough time for utterance planning, and decoupling the content of the turns, we were able to experimentally assess what factors contribute to timely utterance launching while setting aside issues of response planning.

One question to be addressed was what time interval participants would aim for in coordinated responding: given the ease of the task, would the central tendency of launching be immediate (0 ms), consistent with the instruction to respond without a gap, or would it reflect estimates taken from corpora (200 ms)? Responding immediately would be possible, but would require participants to predict the offset of the confederate prompt and prepare to launch their utterance before the prompt offset. Responding at a short delay would be more consistent with earlier literature, and would suggest that participants observed rather than predicted the prompt offset, or that they predicted the offset but responded at a delay.

Another question was which cues participants would use to launch their utterances. To address it, we varied whether the participant could see the confederate's picture (occluded vs. overt). This manipulation tested the role of corepresentation of utterance content in coordinated launching. If corepresentation of the confederate's utterance content contributes to precise launching of a response, seeing her picture and being able to covertly generate its name should greatly facilitate launching compared with a situation where the picture is occluded and full corepresentation of utterance content is not possible.

We also varied two properties of the confederate prompts: the prompt length (long vs. short), and the block-level predictability of prompt length (mixed-length vs. pure blocks). If participants used these types of cues in the speech signal, we expected shorter gaps after long than short prompts. As the prompt utterance began at a variable interval after picture presentation, the earliest moment that the participant could use to time utterance launch would be the prompt onset. This means that long prompts offered more launch preparation time than short ones, and by virtue of being richer speech signals, also offered more cues to the end of the utterance; both properties should improve coordination for long versus short prompts. We also expected shorter gaps after pure than mixed blocks. This is because if given sufficient preparation time, participants respond faster when the timing of a response signal relative to a precue is fixed than when the timing is variable (e.g., Niemi & Näätänen, 1981; Rolke & Ulrich, 2010; Teichner, 1954; Woodrow, 1914). In our paradigm, the onset of the participant utterance might function as a precue; hence responses should be faster in the (more uniform) pure condition, than in the (more variable) mixed condition. This effect might be particularly pronounced for long prompts that offer more preparation time.

We indexed the success of launching speech by measuring the turn gap between the offset of the prompt and the onset of responding. This was the primary measure of interest. We used a direct measure of time between turns for all analyses rather than a transformed or derived measure (e.g., precision, Corps et al., 2018; or entropy, de Ruiter et al., 2006) because the direct measure of turn gap makes the fewest assumptions about the underlying distribution of response times. This makes it most suitable for determining where the distribution is centered and for building predictive models.[1]

We designed the paradigm to isolate launching from planning; however, launching speech is contingent on having successfully planned an utterance when launch needs to happen. Therefore, we also measured the time interval that participants took to plan their speech to establish whether we had made response planning easy in all conditions. We did so by using eye-tracking to time when participants fixated their own picture and the confederate's picture in relation to speech onset. The pattern of fixations to pictures captures how participants chose to allocate their attention during the trial, and the time elapsed between the onset of the gaze to their own picture and the onset of speech (eye-voice lead) captures how much time speakers needed to plan and launch their utterances. We predicted that occlusion would increase both turn gaps and eye-voice lead to a similar degree because it hinders launching, affecting both measures, but has no additional impact on planning. We also predicted an effect of confederate utterance length on eye-voice lead in the opposite direction to the pattern for turn gaps: If participants launch their utterances to coincide with the confederate's turn offset, launching should occur later, and eye-voice-lead should be longer, for long versus short prompts; this effect is predicted to be stronger in pure than in mixed blocks.

The primary analyses quantified how the experimental variables impacted launching. To investigate this question, we performed two types of analyses. The first set used a standard linear mixed effect regression approach, where combinations of cues are tested as fixed-effect predictors. This showed which cues reliably contributed to turn gap in a manner that is directly comparable with earlier literature. The second set built models from combinations of predictors using *k*-fold cross-validation to show what factors captured the variability in turn gap best. This allowed us to directly rank the relative importance of content and speech cues in launching speech in the form of a predictive model.

In summary, we examined how much corepresenting a partner's utterance content contributed to the timely launching of a response, relative to attending to the acoustic signal of a confederate prompt alone. We did this in three production experiments by varying the visibility of the confederate's to-be-named picture, the length of the prompt, and the predictability of prompt length. We used mixed effect modeling and a novel application of statistical modeling to estimate how much these types of cues—accessibility of picture knowledge, prompt predictability, and the timing of the prompt itself—contributed to timely utterance launching.

[1] Analyses contrasting turn gap and precision for all experiments appear in the online supplemental materials.

Experiment 1

The goal of this experiment was to assess how successfully individuals can time their speech launch to the offset of simple prerecorded confederate prompts based upon the predictability of the prompt's length and content. The stimuli were numbers, which we selected because these are frequent, vary in length and are portrayed with visually simple symbols (numerals). We manipulated the length predictability of the prompts by presenting long and short prompts in pure or mixed blocks. The prediction was that longer prompts would afford tighter coordination, and that this would be accentuated when they were embedded in pure blocks. We also manipulated whether the confederate's numeral was visible to the participant, allowing participants to easily form a corepresentation of the prompt only on a subset of trials. If corepresentation affords tight coordination, occlusion should hinder timely launching.

Method

Participants

Data were collected from 56 individuals recruited from the participant database of the Max Planck Institute for Psycholinguistics. Eleven participants were excluded because of a mismatch between the list files presented to the two experimental computers, one participant was excluded because of a computer crash, and four were excluded because of eye-tracking calibration issues. This left a final sample of 40 participants. This sample size was chosen based upon power simulations of the ability to reliably observe condition differences of 25 ms or larger (a small but meaningful effect in word production studies) with a 63 to 91 ms *SD* per condition (pooled $SD = 79$ ms) in a $2 \times 2 \times 2$ design with 128 trials at 80% power; estimates of mean and *SD* were based upon Experiment 3 in Meyer et al. (2003), which manipulated length predictability in production.

The final sample of 40 participants (31 female) was on average 23.4 years old and ranged from 20 to 33 years old. All participants reported normal or corrected to normal vision and hearing and were native speakers of Dutch. They gave informed consent for participating in the study and were paid 6 € for their participation. The study was approved by the ethics board of the Faculty of Social Sciences of Radboud University.

Materials and Design

On each trial, participants saw two stimuli. The item named by the prerecorded confederate ("Confederate Item") was on the left side of the screen. This was a numeral, or, in the occluded condition, a single '#'. The item to be named by the experimental participant ("Participant Item") was on the right side of the screen.

Confederate items were divided into a set of four short duration numbers, consisting of 1 (één), 2 (twee), 10 (tien), and 11 (elf), and a set of four long duration numbers, consisting of 7 (zeven), 9 (negen), 12 (twaalf), and 13 (dertien). These were selected out of the numbers below 20 so that the number of digits on the screen (one or two) was fully crossed with the length of the number in Dutch (long or short).

All confederate items were recorded by a female Dutch native speaker in a simple experimental paradigm where one number was presented on the screen per trial to generate prompts for the main experiment. The confederate produced all numbers 17 unique times (once per experimental item, plus once per practice item) in random order. Prompt onsets reflected the amount of time it took the confederate to plan and produce the picture name. Prompts were randomly assigned to experimental trials such that the same prompt was presented in the same experimental condition for all participants. As Table 1 shows, prompts in the long and short condition differed in onset latencies (by 96 ms; $t(141) = 10.26$, $p < 0.001$) and durations (by 138 ms; $t(135) = 18.29$, $p < 0.001$), so that the offset of the prompt occurred, on average 234 ms later in the long than the short condition. The minimum offset time, leaving the shortest planning interval for the participant, was 938 ms.

In the experiment, prompts were presented in four counterbalanced blocks, two of which were pure blocks containing all long or all short stimuli, and two of which were mixed blocks, containing long and short stimuli interleaved. Confederate items appeared half the time in an overt form, with the number displayed to the participant, and half the time in an occluded form, with the number replaced with a #. This manipulation was randomized across trials within blocks such that each participant saw a unique ordering of items.

Participant items were also half short duration numbers, consisting of 4 (vier), 5 (vijf), 6 (zes), and 8 (acht), and half long duration numbers, consisting of 14 (veertien), 15 (vijftien), 16 (zestien), and 18 (achttien). We did this to provide variability to participants. Given that the participants had ample time to prepare their utterance during the trial, we did not expect an effect of utterance length on any dependent measure.

Each participant number repeated 16 times within the experiment for a total of 128 trials. Each of the confederate numbers was paired twice with each of the participant numbers, and each block contained two tokens of each confederate and participant number,

Table 1
Mean Onset, Offset, and Duration of Prerecorded Experiment 1 Prompts (ms)

Short condition				Long condition			
Confederate item	*M* onset	*M* offset	*M* duration	Confederate item	*M* onset	*M* offset	*M* duration
1 (één)	681 (55)	1,108 (63)	427 (17)	7 (zeven)	779 (52)	1,373 (54)	594 (29)
2 (twee)	660 (56)	1,086 (65)	426 (20)	9 (negen)	783 (63)	1,350(69)	567(30)
10 (tien)	699 (56)	1,059 (62)	360 (22)	12 (twaalf)	785 (63)	1,334 (65)	549 (25)
11 (elf)	672 (44)	1,161 (48)	489 (23)	13 (dertien)	750 (55)	1,295 (75)	545 (52)
M	678	1,104	426		774	1,338	564

Note. Standard deviation in parentheses.

with one of each confederate number occluded, and with one of each participant number paired with an occluded confederate number. Block type was counterbalanced across participants using a Latin Square design. Trials within blocks appeared in a different random order for each participant.

Apparatus and Procedure

Visual stimuli were displayed on a 24″ monitor (1920 × 1080 pixels). The confederate and participant numerals were displayed in size 60 Arial font 820 pixels apart (center-to-center visual angle of 14.56°), centered vertically and horizontally.

Stimulus presentation was controlled by SR Research Eyelink Experiment Builder software (Version 1.10.1630; Eyelink Experiment Builder, 2015) on two BenQ computers. One computer played the prerecorded confederate prompts over internal speakers, and the other presented visual stimuli and recorded sound files using a Shure SM10A head-mounted microphone. The two computers were synchronized to each other to time-lock sound presentation with visual stimulus presentation. The participant's eye movements were tracked with an EyeLink 1000 Plus Desktop Mount, Version 5.09. The participant's right eye was tracked at 500 Hz with a spatial accuracy of about 0.25° to 0.5°. Areas of interest were the 150 pixel squares centered on each number.

Each trial began with a drift check, during which a small circular target appeared in the center of the screen that the participant needed to fixate. Once a fixation was registered, the screen displayed the confederate item on the left side of the screen and the participant item on the right side of the screen. A sound file for the prompt was simultaneously played, with the onset (measured from presentation of visual stimuli) and duration of each word varying as in Table 1; note that the prompt utterance always began at a delay from trial onset. The participant's utterance was recorded. Participants were given a 2,500 ms interval from trial onset within which to provide their response.

The experiment began with calibration, followed by instructions and 16 practice trials. Participants were reminded after the practice trials to do their best to start speaking as soon as the recording was finished. The experiment consisted of four blocks with 32 trials each (128 trials total). Participants could take breaks between blocks.

Analysis

Before analysis, the onsets and offsets of the confederate's and participant's utterances were annotated and transcribed by trained research assistants using Praat (Boersma & Weenink, 2017). Fixations to the two interest areas were extracted with EyeLink Data Viewer software (Version 3.1.1; EyeLink Data Viewer, 2017). Gaze duration was calculated from fixation duration. When there was only one fixation to an interest area, gaze duration matched fixation duration. When there were several successive fixations to the same interest area, gaze duration was defined as the time between the start of the first and end of the last of all consecutive fixations in the same interest area, including intervening blinks and saccades.

Mixed-Effect Regression. The first set of analyses used mixed-effect regression models calculated using R (Version 3.5.1; R Core Team, 2018) with package lme4 (Version 1.1–20; Bates et al., 2015) to examine differences between conditions in speech timing. We performed these analyses with eye-voice lead (interval between first fixation to own image and onset of speech) and turn gap (interval between offset of prompt and onset of own speech) as dependent measures. All analyses used the same three crossed fixed effect predictors: Occlusion (Occluded, Overt), Block Context (Pure, Mixed), and Confederate Length (Short, Long). Each predictor was coded with the contrasts (.5, −.5). Production Length (Short, Long) and its interactions were also added to both models to test the role of planning difficulty in launching, though we did not expect it to systematically affect utterance timing. The results of this analysis are reported in Appendix A.

For all analyses, random intercepts were added for Participant and Participant Item. The maximal model justified by the data included random slopes for all predictors and their two-way interactions for Participant and Participant Item; when this model failed to converge, models were refitted after removing random slope terms that accounted for the least variance, beginning with higher-order terms (interactions) before lower-order terms (main effects). Random slopes were also removed if they correlated at .9 or above with any other term to avoid overfitting. The final random effect structure is reported at the bottom of each model table. For linear models, *t*-values above 2 should be considered significant following the field's convention (e.g., Baayen et al., 2008).

Cross-Validation. The second set of analyses assessed which predictors generalized best to new data. To do this, we used stratified 10-fold cross-validation. This method splits the data into *k* parts (*folds*), sampling randomly within conditions to divide data evenly by condition across folds. A model is fit to the data in *k*-1 of the folds and this model is used to predict the remaining fold's data. The mean squared difference between the predicted and observed values of the dependent measure for the last fold mean squared prediction error (MSPE) reflects predicted error. MSPE is averaged across folds over multiple iterations with different randomizations (here, 500); this is known as cross-validation by averaging (e.g., Zhang & Yang, 2015). Models that minimize MSPE have the best fit to the out-of-sample (held out) data, allowing the modeler to select the best combination of predictors from a set of similar options. We used 10 folds as it is currently considered best practice (see, e.g., Arlot & Celisse, 2010).

Model fit was evaluated using two methods. The first method examined Bayesian Information Criterion (BIC), an information criterion that penalizes the improvement in model log-likelihood by the number of parameters in the model such that models with higher log-likelihood and fewer parameters are preferred. BIC can be used to compare two models fit to the same data that are not necessarily nested (Schwarz, 1978). We chose to evaluate BIC rather than Akaike's Information Criterion (AIC), another information criterion often used for cross-validation, because BIC penalizes more strongly against complex models (see, e.g., Liu & Yang, 2011; Zhang & Yang, 2015). The first measure of model fit involved ranking models by their BIC and examining differences in BIC between pairs of nested models. Following convention (Kass & Raftery, 1995), an improvement in BIC of over six is strong evidence for a better model. The second measure of model fit compared the MSPE across models. This allowed us to assess the fit to held-out data, taking into account model runs with particularly high and particularly low error. To evaluate model fit, we report the MSPE for each model and the difference in MSPE for two nested models. For ease of interpretation, we also report

the mean absolute value of error for each model (M|Err|), averaging across folds and iterations. This provides an estimate of model fit on the same scale as the dependent measure.

We used cross-validation to ask two questions about turn gap. First, we examined whether cues from the specific confederate prompt (its onset and offset time) predicted turn gap better than measures of recent experience (the average of the last five recent prompt onsets/offsets). Note that prompt duration, another predictor of potential interest (e.g., de Ruiter et al., 2006), is a linear combination of onset and offset and was covered by including both other predictors. We compared these models to each other and to a model containing only a random intercept by participant. By contrasting models with matching numbers of predictors, we can assess the informativity of different cues in the signal; by comparing models to the random intercept only model, we can assess how much more variance is accounted for by each cue beyond individual differences in planning and launching.

Next, we assessed whether adding two trial-level predictors, Occlusion status and Block Context, improved model fit. This allowed us to see whether these accounted for variance beyond the timing of the prompt itself. Confederate prompt length was not included in these analyses because it covaried with the confederate offset measure. We also report the optimal model created from the cross-validation analyses. This demonstrated how well we could recover turn gap from the original data.

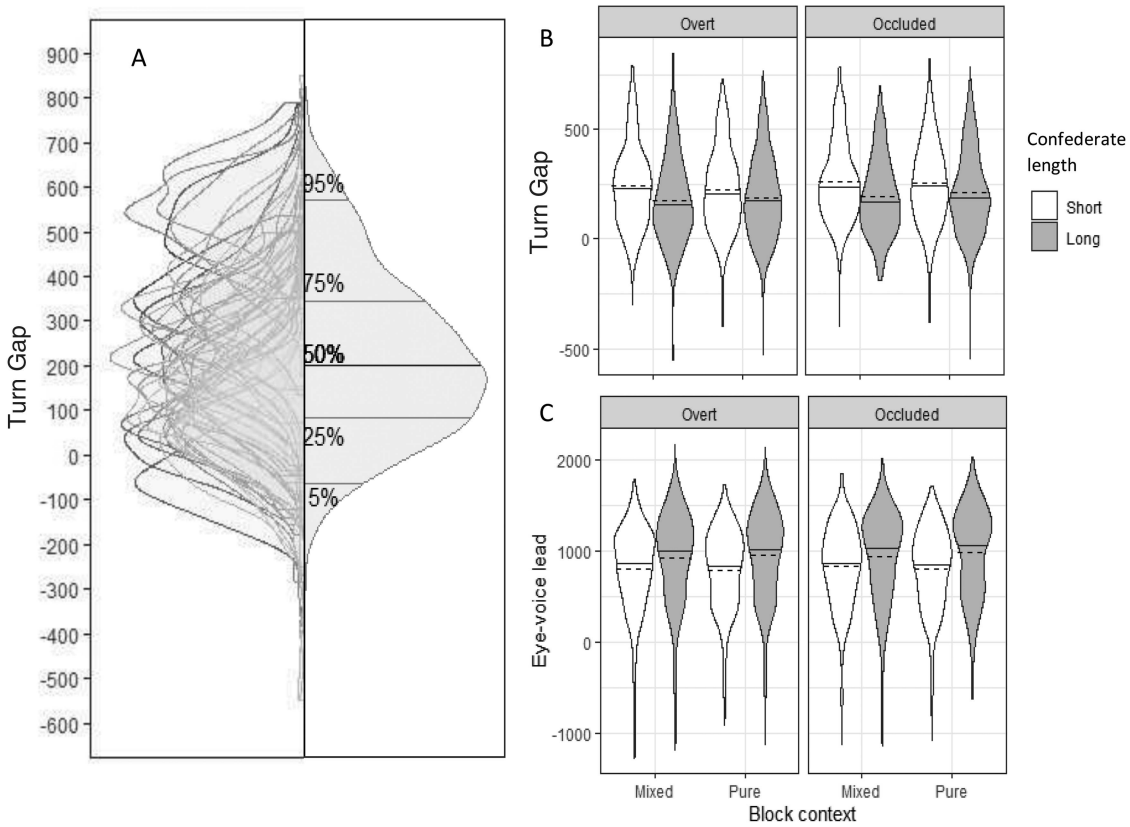
Results and Discussion

Trials were excluded from further analysis if the participant provided the wrong label (48 trials), provided no response (14 trials), started speaking before the confederate (four trials), or if logs indicated that the two computers went out of synch (21 trials). This left 5,033 trials for analysis of turn gap. An additional 365 trials were excluded from analyses of eye-voice lead because no fixation was registered to the participant picture, leaving 4,668 trials.

The participant's task was to align the onset of their utterances as precisely as possible with the offset of the confederate's turn. As shown in Figure 1, the median turn gap in the present study was 201 ms (1st quantile = 83 ms; 3rd quantile = 343 ms), with an overall range from −550 to 849 ms. Only 11% of the utterances began before the offset of the prompt. Both in terms of central tendency and distribution, these turn gaps correspond remarkably well to those reported for analyses of conversations (e.g., Levinson

Figure 1

Distribution of Overall Turn Gap Split by Participant (Panel A, Left) and Pooled, With Quantiles (Panel A, Right), Turn Gap Times by Condition (Panel B) and Eye-Voice Lead Times (Intervals Between First Look to Participant's Stimulus and Participant's Speech Onset; Panel C) in Experiment 1 by Occlusion, Block Context, and Confederate Prompt Length



Note. In panels B and C, dashed lines reflect condition mean and solid lines reflect condition median.

& Torreira, 2015; Stivers et al., 2009). As Figure 1 shows, turn gaps were reliably longest in the mixed block, short condition, and shortest in the mixed block, long condition. We return to these findings below.

We begin by assessing what information participants used in responding, reporting qualitatively where attention was directed during the trials. As Table 2 shows, most of the time a number was fixated (87% across all conditions), attention was directed at the number to be named by the participant. Thus, judging from their eye gaze, participants attended much more to their own item than the confederate's.

However, on 42% of the trials, there was at least one fixation to the confederate number. The next column in the table shows when the fixations to the two numbers occurred: Of the trials when the confederate number was fixated, participants looked at it immediately after trial onset on 82% of trials. This means that on 56% of all trials, the participants only looked at their own number, on 33% of the trials, they looked first at the confederate number and then at their own, on 8% of trials they looked at the confederate number after their own, and on 3% of trials they only looked at the confederate number. The fact that participants preferentially looked early rather than late at the confederate number suggests that gazes to the confederate item did support corepresentation, though note that such gazes occurred on a minority of trials. As Table 2 shows, the fixation proportions for both items were similar across conditions. Recall that occluded and overt confederate items were randomly interleaved; therefore, it is not surprising that participants fixated equally often on both.

The next column of Table 2 shows when the first gaze to a stimulus began. On trials where both numbers were fixated, the first gaze began around 300 ms after stimulus onset (293 ms for the confederate number, and 308 ms for the participant number). When participants only looked at their own number, the gaze onset was later, at 383 ms. It is possible that participants sometimes deliberately delayed fixating upon any of the items. Alternatively, the delay of the first gaze may have led to the absence of a gaze to the confederate number because there was insufficient time remaining to look it.

The following column in Table 2 shows how long participants looked at each of the two items. We focus on the cases where participants looked only at their own number, or looked at the confederate number and then their own (89% of trials). When participants looked only at their own number, the average gaze duration was around 2 s, most of the duration of the trial. In the remaining time, they fixated areas outside of the two interest areas, looked away from the screen, or blinked. When the participants looked first at the confederate number, the average gaze duration was 496 ms and varied little by occlusion. This is not surprising, as the stimuli were highly frequent and easy to read. The following gaze (to the participant's number) had an average duration of 1,332 ms, lasting almost to the end of the trial. This makes two common gaze patterns: Participants either looked only at their own number, or they looked briefly at the confederate's number and then at their own for a prolonged period of time.

To assess the relationship between looking and speaking time, we measured eye-voice lead, which is the time between the onset

Table 2
Fixations and Gaze Data by Condition in Experiment 1

Condition			All looking time (%)						Gaze in time		Gaze out time	
Block context	Confederate length	Occlusion	C	P	Trials with C fixations (%)		First gaze (%)		C	P	C	P
Mixed	Short	Occluded	11%	89%	Yes	39%	C	86%	281	909	768	2,274
							P	14%	1,219	254	1,498	1,065
					No	61%	P	100%	—	361	—	2,186
		Overt	12%	88%	Yes	42%	C	83%	282	917	751	2,256
							P	17%	1,330	263	1,605	1,070
					No	58%	P	100%	—	364	—	2,159
	Long	Occluded	13%	87%	Yes	42%	C	83%	288	1,012	830	2,312
							P	17%	1,289	286	1,601	1,070
					No	58%	P	100%	—	385	—	2,251
		Overt	13%	87%	Yes	43%	C	82%	302	993	801	2,311
							P	18%	1,233	266	1,566	1,102
					No	57%	P	100%	—	402	—	2,201
Pure	Short	Occluded	13%	87%	Yes	40%	C	80%	277	934	790	2,285
							P	20%	1,032	312	1,440	883
					No	60%	P	100%	—	409	—	2,132
		Overt	13%	87%	Yes	41%	C	82%	291	871	723	2,275
							P	18%	1,097	239	1,496	885
					No	59%	P	100%	—	398	—	2,145
	Long	Occluded	14%	86%	Yes	40%	C	82%	295	1,012	855	2,279
							P	18%	899	247	1,412	714
					No	60%	P	100%	—	380	—	2,219
		Overt	14%	86%	Yes	47%	C	79%	321	956	792	2,275
							P	21%	1,068	305	1,434	895
					No	53%	P	100%	—	364	—	2,207

Note. C = confederate number; P = participant number.

of the participant's first fixation to their own number and the onset of speech. This is an index of the amount of time between beginning to plan the response and beginning to say it—or alternatively, the sum of planning time and launching time.

Recall that we had expected a main effect of occlusion on eye-voice lead, with shorter eye-voice lead in the overt than the occluded condition, a main effect of confederate length, with shorter eye-voice lead for short than long prompts, and an interaction of block context and confederate length, with the length effect being stronger in pure than mixed blocks. As shown in Figure 1 and confirmed with mixed effect analyses (see Table 3), eye-voice lead was significantly shorter (by 18 ms) when the confederate number was overt rather than occluded, and significantly longer (by 144 ms) when the prompt was long than when it was short. Finally, the difference in eye-voice lead for long and short prompts was larger in the pure than the mixed blocks (173 vs. 115 ms) suggesting that as predicted, the uniform-length prompts allowed participants to adjust their speech timing most closely to the confederate's.

For turn gap (see Figure 1 and Table 3), there was again a small (22 ms) but significant effect of occlusion, with turn gaps being shorter when the confederate number was overt than when it was occluded. This means that when the participants could see the confederate stimulus, they initiated their utterances earlier than when it was occluded, shortening both the eye-voice lead and the turn gap. There was also a significant main effect of confederate length, and a significant interaction between confederate length and block context: As predicted, long prompts elicited shorter turn gaps (by 54 ms) than short prompts, and contrary to predictions, the effect of confederate length was weaker in pure blocks (41 ms) than in mixed blocks (67 ms).

In a first series of cross-validation analyses, we examined which properties of the confederate stimulus predicted turn gap, adding in measures related to the timing of the prompt and recent prompts. Results of this analysis are displayed in Table 4. They showed confederate offset to be the best predictor of turn gap, as indicated by the better fit for both models containing confederate offset. Additional independent variance was accounted for by confederate onset, as indicated by the improved fit for the Confederate Onset + Confederate Offset model compared with the confederate onset model. This suggests that individuals estimated when the confederate was likely to end their speech by attending to the phonetic or intonational properties of her utterance and using this information to time their speech launch. The remaining models performed only slightly better than the random intercept only model. This indicates that recent experiences with confederate onsets and offsets are not particularly useful for responding: they are less informative on any given trial than the specific onset or offset of the confederate prompt. This suggests that predictions about when to respond are made at the level of the individual trial.

We added trial-level predictors to this first set of models in a second series of cross-validation analyses. These models are shown in Table 5. All models were improved by the addition of the occlusion parameter, but the confederate offset models were improved most, with more than six BIC units (a reliably large effect) of improvement compared with the bottom-ranked model. This means that being able to see the confederate number may not provide independent information from what can be gained by listening to the confederate prompt. This might mean that the key information gained from both types of cues, the number and the auditory input, is lexical: the word being spoken. In contrast, no models were improved by the block context predictor. This shows that the block-level predictability of the confederate prompt length did not capture variance beyond the speech cues that are available by listening.

Across all cross-validation analyses, the model with the lowest BIC and MSPE was the model predicting turn gap from confederate onset, confederate offset, and occlusion as main effects in addition to a random intercept by participant. In this model, the mean absolute error per point was 105.48 ms. Figure 2 shows the fit for this model for one sample run. For trials between the 5% and 95% quantiles, the observed data have an MSPE of 8770.85 and a

Table 3
Outputs of Linear Mixed-Effect Models for Eye-Voice Lead and Turn Gap in Experiment 1

	Eye-voice lead			Turn gap		
Fixed effects	Estimate	*SE*	*t* value	Estimate	*SE*	*t* value
Intercept	834.36	64.75	**12.89**	219.86	27.52	**8.00**
Occlusion	27.71	9.01	**3.08**	21.29	3.67	**5.81**
Block context	−2.57	16.11	−0.16	1.78	4.69	0.38
Confederate length	−146.47	16.39	**−8.94**	53.96	5.43	**9.94**
Occlusion × Block Context	19.23	18.01	1.07	5.15	5.64	0.92
Occlusion × Confederate Length	8.38	18.01	0.47	9.82	5.64	1.74
Block Context × Confederate Length	−53.58	18.03	**−2.98**	−25.36	5.64	**−4.49**
Occlusion × Block Context × Confederate Length	−9.68	36.01	−0.27	1.43	11.29	0.13
Random effects	Term		*SD*	Term		*SD*
	Participants	Intercept	398.93	Participants	Intercept	219.00
		Block context	83.89		Occlusion	14.80
		Confederate length	86.08		Block context	23.65
	Item	Intercept	40.31		Confederate length	29.33
	Residual		307.24	Item	Intercept	34.55
				Residual		100.06

Note. Bold values reflect *t*-values above 2.

Table 4
Prediction of Turn Gap by Cross-Validation for Experiment 1, Ranked by Model Fit

Rank (BIC)	Rank (MSPE)	Model parameters	BIC	ΔBIC	MSPE	ΔMSPE	M|Err|
1	1	Confederate onset + Confederate offset	51874.97	452.27[a]	11258.89	1349.47	105.95
2	2	Confederate offset	51978.38	348.85[a]	11568.64	1039.72	107.40
3	3	Confederate onset	52224.09	103.15[a]	12267.7	340.66	110.62
4	5	Average last 5 confederate offsets	52296.73	30.50[a]	12480.14	128.22	111.58
5	4	Average last 5 confederate onsets + Average last 5 confederate offsets	52303.19	24.04[a]	12473.04	135.32	111.55
6	6	Average last 5 confederate onsets	52310.56	16.68[a]	12526.1	82.26	111.78
7	7	Random intercept only	52327.24	—	12608.36	—	112.15

Note. All models included a random intercept by participant. Delta Bayesian information criterion (BIC) and delta mean squared prediction error (MSPE) are calculated by comparison to a random-intercept only model.
[a] Models with reliably improved BIC.

mean absolute error per point of 72.65 ms; in the tails of the distribution, the observed data are predicted less well, with an MSPE of 34,751.87 and mean absolute error per point of 153.58 ms for the left tail and an MSPE of 30,313.12 and mean absolute error per point of 120.85 ms for the right tail.

In summary, Experiment 1 showed that individuals were highly successful at synchronizing the onset of their own utterance with the offset of a prerecorded prompt. The modal turn gap was approximately 200 ms, which is very similar to the modal turn gap in conversation corpora (Levinson & Torreira, 2015; Stivers et al., 2009). Turn gaps were shortest when the participants could see the number to be named by the confederate, consistent with the view that tight utterance timing may be achieved by corepresentation of upcoming utterances. Yet, participants chose to visually attend the confederate number only on a minority of trials, and very good coordination was also achieved when such information was not available, as evidenced by the relatively small improvement from including this predictor in cross-validation models. This suggests that full corepresentation of the prior utterance's content can be useful, but is not necessary to achieve swift turn taking in a simple language task.

Experiment 2

The goal of Experiment 2 was to replicate Experiment 1 with new materials, line drawings rather than numbers, and a different task, picture naming rather than reading aloud. The use of line drawings afforded greater experimental control of the phonetic properties of the confederate and participant utterances. In Experiment 1 there were only eight confederate prompts, and the stimuli became unique at an early point within the word. Aside from the onset tʊ, occurring for *twee* (2) and *twaalf* (12), each confederate item had a unique onset. This may have allowed participants to form precise predictions about the upcoming word after hearing the onset of the prompt. In Experiment 2, we used mono- and disyllabic picture names matched on the first consonant and vowel

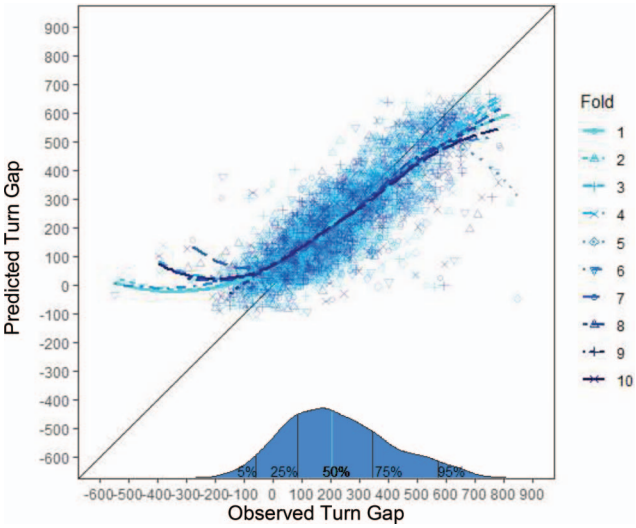
Table 5
Prediction of Turn Gap by Cross-Validation for Experiment 1 With Addition of Occlusion and Block Context Parameters, Each Series Ranked by Model Fit

Rank (ΔBIC)	Rank (ΔMSPE)	Model parameters	BIC	ΔBIC	MSPE	ΔMSPE	M|Err|
		Occlusion models					
1	2	Confederate onset + Confederate offset + Occlusion	51839.78	35.19[a]	11159.89	99.00	105.48
2	1	Confederate offset + Occlusion	51943.76	34.62[a]	11468.51	100.13	106.94
3	3	Confederate onset + Occlusion	52192.81	31.28[a]	12171.23	96.47	110.18
4	4	Average last 5 confederate onsets + Average last 5 confederate offsets + Occlusion	52273.88	29.31[a]	12380.62	92.42	111.13
5	6	Average last 5 confederate offsets + Occlusion	52267.70	29.03[a]	12388.53	91.61	111.17
6	5	Random intercept + Occlusion	52298.47	28.77[a]	12516.49	91.87	111.74
7	7	Average last 5 confederate onsets + Occlusion	52281.86	28.70[a]	12435.07	91.03	111.37
		Block context models					
1	1	Random intercept + Block context	52330.88	−3.64	12613.26	−4.64	112.17
2	4	Average last 5 confederate onsets + Average last 5 confederate offsets + Block context	52306.85	−3.66	12478.31	−4.73	111.57
3	5	Confederate onset + Block context	52227.78	−3.69	12273.2	−4.81	110.64
4	3	Average last 5 confederate offsets + Block context	52300.43	−3.70	12485.52	−4.85	111.60
5	2	Average last 5 confederate onsets + Block context	52314.28	−3.72	12531.6	−4.92	111.80
6	6	Confederate offset + Block context	51982.13	−3.75	11573.62	−4.69	107.43
7	7	Confederate onset + Confederate offset + Block context	51878.73	−3.76	11263.5	−4.56	105.97

Note. All models included a random intercept by participant. Delta Bayesian information criterion (BIC) and delta mean squared prediction error (MSPE) are calculated by comparison with a nested model without the added parameter (occlusion, block context).
[a] Models with reliably improved BIC.

Figure 2

Observed Versus Predicted Turn Gaps for Best-Fitting Cross-Validation Model of Turn Gap in Experiment 1 Across 10 Randomized Folds for a Sample Iteration



Note. Model is: Turn Gap ~ Confederate Onset + Confederate Offset + Occlusion + (1 Participant). Points reflect trial-level data; lines reflect loess smooths for each cross-validation fold. Density plot at bottom reflects distribution of observed data (on arbitrary Y scale), with vertical lines reflecting quantiles. See the online article for the color version of this figure.

to make the confederate's speech onset less informative. In addition, we doubled the size of the stimulus set. Both of these changes should make gazes to the confederate pictures more helpful for utterance timing than in Experiment 1. In other words, full corepresentation of confederate utterances might be more likely to be observed in this than in the preceding experiment, with correspondingly larger effects of occlusion.

Method

Participants

Data were collected from 41 individuals (33 female, average age 23 years, range = 20 to 31 years) recruited from the participant database of the Max Planck Institute for Psycholinguistics. This sample size was chosen as described in Experiment 1. All participants were native speakers of Dutch and reported normal or corrected to normal vision and hearing. Informed consent was obtained from each participant. They were paid 6 € for their participation.

Materials and Design

As before, two stimuli were shown on each trial. Confederate items were presented on the left side of the computer screen and named by a prerecorded confederate, and participant items were presented on the right side and named by the participant. All items were colored drawings of common objects (see Appendix B). Most items came from the Multipic database (Duñabeitia et al., 2017). The image for *nagel* (finger *nail*) was cropped from the original Multipic picture. Images for *tang* (pliers) and *ladder* (ladder) came from Rossion and Pourtois (2004). *Panda* came from Severens et al. (2005); *zee* (sea) came from an internal database, and *hemd* (undershirt) came from Wikimedia commons. The latter three drawings were colored in Photoshop by the first author. In the occluded condition, the confederate item was replaced by a Gabor patch. All pictures were displayed at a size of 200 pixels square, 900 pixels apart (center-to-center visual angle of 15.82°), and centered vertically and horizontally on the screen.

Confederate items were a set of eight pictures with monosyllabic names and eight pictures with disyllabic names; a different set of eight pictures with monosyllabic names and eight pictures with disyllabic names made up the response items named by the participant. The mono- and disyllabic item sets were matched for average word frequency as well as their onset consonant(s) and first vowel (see Table 6).

The names of the confederate pictures were recorded by the same female native speaker of Dutch as in Experiment 1 and these prompt utterances were assigned to trials as described in Experiment 1. The resulting speech onset times (measured from trial onset, reflecting the amount of time the confederate took to plan, and launch her picture name), offset times, and durations are shown in Table 6. The minimum offset time, reflecting the participant's minimum preparation interval, was 937 ms. Note that the average onset times were similar for the mono- and disyllabic prompt sets, differing only by 9 ms, $t(118) = 0.22$, $p = .82$, though

Table 6

Mean Onset, Offset, and Duration for Prerecorded Prompts (ms) Used in Experiment 2

Monosyllabic					Disyllabic				
Confederate item	*M* onset	*M* offset	*M* duration	Word frequency	Confederate item	*M* onset	*M* offset	*M* duration	Word frequency
Hemd (undershirt)	612 (65)	1114 (80)	502 (24)	11.98	Herder (shepherd)	639 (34)	1181 (38)	542 (24)	5.92
Mand (basket)	606 (59)	1054 (58)	439 (16)	4.30	Masker (mask)	628 (55)	1131 (64)	503 (31)	19.23
Naald (needle)	769 (151)	1245 (144)	476 (43)	8.51	Nagel (fingernail)	814 (97)	1319 (111)	505 (18)	4.05
Pan (pan)	717 (80)	1017 (88)	299 (21)	9.38	Panda (panda)	770 (50)	1224 (51)	454 (15)	0.96
Rits (zipper)	621 (95)	1169 (113)	548 (32)	4.37	Ridder (knight)	625 (55)	1194 (72)	569 (43)	13.58
Schaar (scissors)	582 (94)	1108 (115)	526 (29)	6.36	Schaduw (shadow)	641 (79)	1292 (84)	651(31)	20.92
Vlieg (fly)	633 (134)	1174 (147)	541 (25)	29.07	Vliegtuig (airplane)	546 (79)	1340 (102)	794 (38)	89.92
Zee (ocean)	658 (65)	1143 (76)	485 (23)	67.80	Zebra (zebra)	611 (70)	1227 (109)	616 (43)	3.06
M	650		477	17.72		659		579	19.71

Note. English translations follow item labels; standard deviations follow *M*s. Frequencies (per million words) are from SUBTLEX-NL ("SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles," by E. Keuleers, M. Brysbaert, and B. New, 2010, *Behavior Research Methods, 42*(3), pp. 643–650. Copyright 2010).

their durations differed by 102 ms, $t(126) = 6.66$ $p < .001$. The lack of an onset difference contrasts with the materials in Experiment 1, where long and short prompts differed in onset times and durations. This difference in the materials of the two experiments was not intended, but may have resulted from using small sets of pictures and a single speaker in both experiments.

Confederate prompts were presented in mixed and pure blocks as described in Experiment 1 and the associated images again appeared half the time (randomized across trials) in an overt form, with the picture displayed to the participant, and half the time in an occluded form, replaced by a Gabor patch. As before, a mix of monosyllabic and disyllabic participant items was used to introduce variability in utterance planning and to mirror the set of items used for the confederate's responses. No systematic effect of participant utterance length was expected since the participants engaged in a delayed naming task; moreover, any effects would be difficult to interpret because the pictures were not matched for visual complexity or ease of recognition. Analyses containing this predictor are reported in Appendix A for completeness.

Each item repeated eight times within the experiment, for a total of 128 trials. Each of the confederate items was paired once with each of the participant items, and each block contained one token of each confederate and participant item, with occlusion balanced for confederate items across blocks, and with occluded and overt confederate items occurring equally often with each participant item. As in Experiment 1, block type was counterbalanced across participants using a Latin Square design; trials within blocks appeared in a different random order for each participant.

Apparatus and Procedure

The apparatus and stimulus presentation were as described in Experiment 1. Trials followed the same procedure as in Experiment 1 with the addition of a familiarization phase added before the experimental instructions. In this phase all images were presented one by one for 4,000 ms with their names written below in Times New Roman font in size 24. Participants were asked to use these names to refer to the pictures.

Analysis

As in Experiment 1, the participants' utterances were transcribed and their onsets determined using Praat (Boersma & Weenink, 2017). The region of interest around each picture was a 200 × 200 pixel square; fixations and gazes were defined as in Experiment 1. The same mixed effect modeling and cross-validation procedures were used as in Experiment 1.

Results and Discussion

Trials were excluded from further analysis if the participants provided a wrong label (111 trials), provided no response (15 trials), or started speaking before the confederate began (four trials). This left 5,118 trials for analysis. An additional 231 trials were excluded from analysis of eye-voice lead because no fixation was registered to the participant's picture, leaving 4,887 trials.

Similar to Experiment 1, utterances were well synchronized to the offset of the prompt. As shown in Figure 3, the median turn gap was 217 ms (1st quantile = 111 ms; 3rd quantile = 358 ms), with an overall range from −466 to 1,119 ms. The length of the confederate prompt again affected coordination, with gaps being shorter after long than short confederate prompts, whereas the effects of occlusion and block context were reduced compared with Experiment 1.

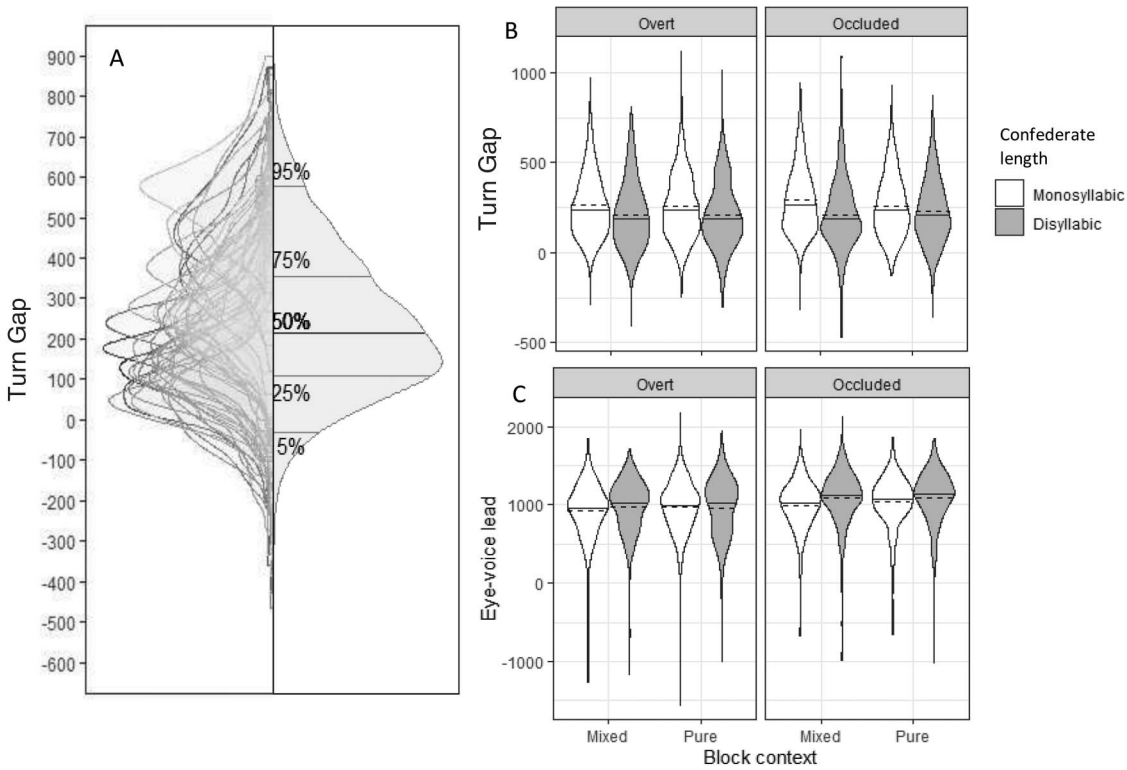
As shown in Table 7, attention while preparing to respond largely mirrors Experiment 1. In total, 86% of the fixations to one of the two interest areas were directed at the participant's own picture, and on 50% of the trials, participants only looked at their own picture. On the 50% of trials that featured at least one fixation to the confederate picture, these fixations were, again, more likely to occur early than late. When participants fixated both pictures, they looked first at the confederate picture on 53% of trials, and first at their own picture on the remaining 47% of trials. On the trials where participants only looked at their own picture, their gaze began 330 ms after trial onset and lasted nearly until the end of the trial. When participants looked first at the confederate picture, the gaze began around 260 ms after trial onset, lasted about 340 ms, and was followed by a longer gaze (around 1450 ms) to the participant's picture starting around 720 ms after trial onset. Finally, when participants looked first at their own picture, the gaze began around 240 ms after trial onset and lasted about 600 ms; the gaze to the confederate picture began around 1,010 ms after trial onset and lasted about 420 ms, or until approximately 1,400 ms after trial onset. During the remaining time, participants typically looked at their own picture again. As in Experiment 1, the order and timing of gazes were not strongly affected by the experimental conditions, though unlike Experiment 1, more fixations were made to the confederate picture when it was overt (vs. occluded). This difference is likely related to the increased visual complexity and size of the stimuli in Experiment 2, as it may have been easier for participants to see from the central fixation cross if the confederate picture was occluded.

As in Experiment 1, eye-voice lead was significantly shorter when the confederate picture was overt (959 ms) than when it was occluded (1,052 ms). While the difference was not statistically significant, eye-voice lead was also numerically shorter when the confederate prompt was monosyllabic (981 ms) than when it was disyllabic (1,029 ms), presumably reflecting the fact that the prompts ended 112 ms later in the disyllabic condition. The length effect was also numerically stronger when the confederate picture was occluded (74 ms) than when it was overt (22 ms), and was reliably stronger in mixed blocks (74 ms) than in pure blocks (23 ms). The direction of the block context by length effect opposes Experiment 1 and shows that in Experiment 2, the eye-voice lead advantage for long utterances was reduced, rather than increased, by any additional information—when confederate's pictures could be seen, and when information about the utterance length was implicitly provided in pure blocks.

As in Experiment 1, turn gaps were significantly shorter when the confederate picture was overt (234 ms) than when it was occluded (247 ms), and significantly shorter after disyllabic confederate prompts (214 ms) than after monosyllabic ones (267 ms; see Table 8 and Figure 3). The length effect highlights the role of prior turn length in launching, compared with planning a response. Opposing Experiment 1, the effect of confederate length on turn gaps was stronger in mixed blocks (67 ms) than pure blocks (40 ms). In other words, the disyllabic advantage (shortening the turn gap relative to monosyllabic prompts) was strongest when, apart from the phonetic input, no other information about the confederate prompt was available. This was evidenced by a two-way

Figure 3

Distribution of Overall Turn Gap Split by Participant (Panel A, Left) and Pooled, With Quantiles (Panel A, Right), Turn Gap Times by Condition (Panel B) and Eye-Voice Lead Times (Intervals Between First Look to Participant's Stimulus and Participant's Speech Onset; Panel C) in Experiment 2 by Occlusion, Block Context, and Confederate Prompt Length



Note. In panels B and C, dashed lines reflect condition mean and solid lines reflect condition median.

interaction between block context and confederate prompt length, and a main effect of occlusion.

In a first series of cross-validation analyses, reported in Table 9, we examined which properties of the confederate stimulus predicted turn gap. The model that used both the onset and offset times of the confederate prompt to predict turn gap performed the best on all metrics, followed by the model that used only confederate offset time. Models predicting turn gap from recent confederate onsets and offsets performed worse than the random intercept only model. This provides further evidence that recent confederate onsets and offsets are not particularly useful for planning when to respond.

In a second series of cross-validation analyses (see Table 10), we added trial-level predictors to the better-than-baseline performing models reported above. Only the models containing confederate offset time were reliably improved by the addition of occlusion. This provides strong evidence that occlusion is not independent from other factors, consistent with the idea that speech cues and overt stimuli both allow participants to access the confederate picture's name. As in Experiment 1, no model was improved by adding block context.

Across all cross-validation analyses, the model with the lowest BIC and MSPE was again the model predicting turn gap from confederate onset, confederate offset, and occlusion as main effects in addition to a random intercept by participant; one sample run of this model appears in Figure 4. In this model, the mean absolute error per point was 118.05 ms. For trials between the 5% and 95% quantiles, the observed data have an MSPE of 9457.05 and a mean absolute error per point of 74.20 ms; for data in the tails of the observed distribution, the observed data are predicted less well, with an MSPE of 33,223.92 and mean absolute error per point of 140.36 ms for the left tail and an MSPE of 77,049.15 and mean absolute error per point of 212.67 ms for the right tail. These metrics are comparable with the final model selected in Experiment 1.

In summary, Experiment 2 replicated the main findings of Experiment 1: Participants timed their utterances to begin about 200 ms after the offset of the confederate prompt. The gap between utterances was shortened when participants could see the confederate's stimulus, implying a role for corepresentation in coordination, but good temporal coordination was also achieved when they could not see this confederate item but had to rely only on hearing the input. Again, the onset and offset times of the confederate prompt were the most important predictors in cross-validation models. This shows that the timing of the confederate's utterance—its onset and offset—provided the most important triggers for launching speech in a simple language task.

Table 7
Fixations and Gaze Data by Condition in Experiment 2

Condition			All looking time (%)						Gaze in time		Gaze out time	
Block context	Confederate length	Occlusion	C	P	Trials with C fixations (%)		First gaze (%)		C	P	C	P
Mixed	Short	Occluded	10%	90%	Yes	40%	C	60%	270	680	571	2,124
							P	40%	1,085	274	1,462	889
					No	60%	P	100%	—	343	—	1,961
		Overt	17%	83%	Yes	57%	C	57%	271	748	638	2,085
							P	43%	957	252	1,396	763
					No	43%	P	100%	—	323	—	2,038
	Long	Occluded	10%	90%	Yes	43%	C	58%	248	662	557	2,223
							P	42%	1,057	242	1,442	785
					No	57%	P	100%	—	349	—	2,015
		Overt	17%	83%	Yes	62%	C	54%	277	735	622	2,194
							P	46%	1,026	251	1,444	831
					No	38%	P	100%	—	322	—	2,113
Pure	Short	Occluded	10%	90%	Yes	40%	C	57%	243	697	574	2,194
							P	43%	1,051	247	1,480	841
					No	60%	P	100%	—	323	—	2,007
		Overt	15%	85%	Yes	54%	C	51%	264	721	612	2,156
							P	49%	1,028	240	1,461	851
					No	46%	P	100%	—	303	—	2,000
	Long	Occluded	10%	90%	Yes	41%	C	55%	271	699	597	2,219
							P	45%	998	227	1,372	811
					No	59%	P	100%	—	317	—	2,057
		Overt	19%	81%	Yes	62%	C	54%	257	774	666	2,177
							P	46%	957	229	1,420	825
					No	38%	P	100%	—	350	—	2,057

Note. C = confederate picture; P = participant picture.

Experiment 3

Evidence from all measures in Experiments 1 and 2 suggest that participants only weakly corepresented the confederate stimulus. This may have been because of the use of occlusion as a predictor: On half of the trials, the confederate stimulus presented to the participant provided no useful information about the confederate's utterance. This may have discouraged the participants from looking at the confederate's stimuli in the first place and it limited the information gleaned from such looks. In Experiment 3, the confederate items were always overt. The question was whether the participants would now be more likely to look at the confederate

Table 8
Outputs of Linear Mixed-Effect Models for Eye-Voice Lead and Turn Gap in Experiment 2

Fixed effects	Eye-voice lead			Turn gap		
	Estimate	*SE*	*t* value	Estimate	*SE*	*t* value
Intercept	988.66	32.14	**30.76**	241.16	21.89	**11.02**
Occlusion	88.05	9.47	**9.29**	18.64	3.77	**4.94**
Block context	12.28	29.55	0.42	−5.95	10.65	−0.56
Confederate length	−41.91	23.07	−1.82	52.10	3.49	**14.91**
Occlusion × Block Context	−18.05	19.72	−0.92	−4.39	7.56	−0.58
Occlusion × Confederate Length	−28.48	18.36	−1.55	13.06	7.51	1.74
Block Context × Confederate Length	51.49	17.26	**2.98**	−26.33	6.99	**−3.77**
Occlusion × Block Context × Confederate Length	78.69	40.58	1.94	−12.06	15.02	−0.80

Random effects	Term		*SD*	Term		*SD*
	Participants	Intercept	190.75	Participants	Intercept	135.03
		Block context	141.47	Item	Intercept	22.47
		Confederate length	69.41		Block context	40.31
	Item	Intercept	44.96	Residual		123.14
		Block context	70.29			
		Confederate length	73.66			
	Residual		310.6			

Note. Bold values reflect *t*-values above 2.

Table 9
Prediction of Turn Gap by Cross-Validation for Experiment 2 Ranked by Model Fit

Rank (BIC)	Rank (MSPE)	Model parameters	BIC	ΔBIC	MSPE	ΔMSPE	M|Err|
1	1	Confederate onset + Confederate offset	53820.79	796.86[a]	14065.33	2972.34	118.34
2	2	Confederate offset	54312.81	304.83[a]	15822.24	1215.43	125.56
3	3	Confederate onset	54590.82	26.82[a]	16881.16	156.51	129.71
4	4	Random intercept only	54617.64	—	17037.67	—	130.31
5	5	Average last 5 confederate onsets	54629.19	−11.55	17041.43	−3.76	130.32
6	6	Average last 5 confederate offsets	54631.41	−13.77	17046.58	−8.91	130.34
7	7	Average last 5 confederate onsets + Average last 5 confederate offsets	54641.71	−24.07	17047.16	−9.49	130.35

Note. All models included a random intercept by participant. Delta Bayesian information criterion (BIC) and delta mean squared prediction error (MSPE) are calculated by comparison with random-intercept only model.
[a] Models with reliably improved BIC.

screen, presumably using visual information to corepresent the confederate's utterance.

A second potentially consequential property of the materials used in Experiments 1 and 2 was that the prompts were always single words, making them all relatively short. This means that turn gaps may have been centered around 200 ms rather than 0 ms because, in the absence of an earlier acoustic "landmark" in the confederate's speech, earlier launching was impossible. Moreover, corepresenting the content of the utterances might not have been very useful as there was little variation in the length of the confederate's turns. To assess whether longer prompts afford tighter coordination and whether corepresentation might be more beneficial when utterance length is more variable, Experiment 3 used confederate items that were single nouns, accompanied by single pictures, or noun pairs, accompanied by picture pairs. These short and long stimuli were again presented in pure and mixed blocks. Longer utterances provide earlier trigger points for planning speech and projecting launch time and might lead to shorter turn gaps if the timing of the prompt is what affords coordination. More important, with increasing uncertainty about the length of the utterance and constant availability of the prompt stimulus, corepresentation may be more valuable than it was in Experiments 1 and 2.

Method

Participants

Data were collected from 44 individuals recruited from the participant database of the Max Planck Institute for Psycholinguistics. Two participants were excluded because of a computer crash, and two were excluded because of eye-tracker calibration issues. This left a final sample of 40 participants (34 female, average age 23.8 years, range = 19 to 34 years). We chose this sample size to match the two previous experiments. All participants were native speakers of Dutch and reported normal or corrected to normal vision and hearing. Informed consent was obtained from each participant, and participants were paid 6 € for their participation.

Materials and Design

As before, confederate items were presented on the left side of the computer screen and named by a prerecorded confederate, and participant items were presented on the right side and named by the participant. On half the trials, one confederate item was presented in the same location as described in Experiment 1. On the other half of trials, two confederate pictures were presented with the first appearing 25 pixels above the second (3.96° center-to-

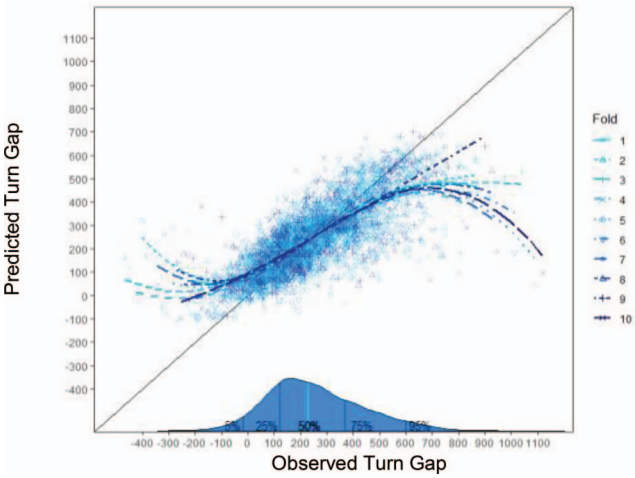
Table 10
Prediction of Turn Gap by Cross-Validation for Experiment 2 With Addition of Occlusion and Block Context Parameters, Each Series Ranked by Model Fit

Rank (ΔBIC)	Rank (ΔMSPE)	Model parameters	BIC	ΔBIC	MSPE	ΔMSPE	M|Err|
		Occlusionmodels					
1	1	Confederate offset + Occlusion	54275.86	36.95[a]	15680.17	142.07	125.00
2	2	Confederate onset + Confederate offset + Occlusion	53801.84	18.95[a]	13997.25	68.08	118.05
3	3	Random intercept + Occlusion	54612.78	4.86	17011.47	26.20	130.21
4	4	Confederate onset + Occlusion	54591.12	−0.30	16875.47	5.69	129.68
		Block context models					
1	1	Confederate offset + Confederate onset + Block context	53818.24	2.55	14052.86	14.56	118.29
2	2	Confederate onset + Block context	54588.91	1.91	16868.68	14.44	129.66
3	3	Random intercept + Block context	54616.28	1.36	17027.55	12.34	130.27
4	4	Confederate Offset + Block Context	54312.08	0.73	15815.16	9.16	125.53

Note. All models included a random intercept by participant. Delta Bayesian information criterion (BIC) and delta mean squared prediction error (MSPE) are calculated by comparison with a nested model without the added parameter (occlusion, block context).
[a] Models with reliably improved BIC.

Figure 4

Observed Versus Predicted Turn Gaps for Best-Fitting Cross-Validation Model in Experiment 2 Across 10 Randomized Folds for a Sample Iteration



Note. Model is: Turn Gap ~ Confederate Onset + Confederate Offset + Occlusion + (1 Participant). Points reflect trial-level data; lines reflect loess smooths for each cross-validation fold. Density plot at bottom reflects distribution of observed data (on arbitrary Y scale), with vertical lines reflecting quantiles. See the online article for the color version of this figure.

center visual angle); the midpoint of the two pictures was centered vertically on the screen and was again 900 pixels (15.82° visual angle) from the participant's picture.

Confederate items used the same pictures as in Experiment 2. In pure one-word blocks, all trials had one confederate picture, which was either mono- or disyllabic, making these blocks analogous to the mixed blocks in Experiment 2. In pure two-word blocks, all trials had two confederate pictures, both mono- or disyllabic. In the mixed blocks, one- and two-word prompts using mono- and disyllabic nouns were randomly interleaved. Participant items were single mono- or disyllabic nouns, as in the preceding experiments. As previously, we predicted that this variability would have no systematic effect on the participants' speech and report analyses containing this factor in Appendix A.

New recordings were made to create prompts with the same trained speaker as in Experiments 1 and 2. A simple experimental paradigm was used where the speaker saw on each trial either a single picture or a picture pair. She was asked to name the pairs in a single fluent utterance, so that the onset time of the confederate prompt reflected the amount of time the confederate needed to plan and launch her utterance. All single items were repeated eight times and all picture pairs were repeated twice. We used these recordings to assign a unique prompt to each trial. The resulting average onset times (defined from trial onset, reflecting the confederate's planning interval), offset times, and durations for each condition are shown in Table 11; the minimum offset time, reflecting the participant's shortest planning interval, was 856 ms. Again, the prompts reliably differed in their temporal properties across conditions: a linear model predicting prompt onset from whether the item was a one- or two-word trial, whether the first word had one or two syllables, and their interaction, showed only a reliable effect of word number, $t(140) = 5.06$, $p < .001$; a similar model predicting prompt duration from the same factors showed reliable effects of first word syllable number, $t(140) = 35.65$, $p < .001$ and word number, $t(140) = 21.41$, $p < .001$.

One-word trials were composed as in Experiment 2; each word appeared four times as a one-word trial. Two-word trials were composed by pairing nouns with the same syllable number pseudorandomly, with the constraints that the same pairing of items (e.g., *hemd*, *pan*) did not appear more than once, no item was duplicated in the same trial (e.g., *hemd*, *hemd* did not appear), and items of the same semantic category (e.g., *panda* and *zebra*) did

Table 11

Mean Onset, Offset, and Duration of Experiment 3 Prerecorded Confederate Prompts Split by First Noun (in ms)

	Monosyllabic				Disyllabic			
Word number	Confederate item	*M* onset	*M* offset	*M* duration	Confederate item	*M* onset	*M* offset	*M* duration
One word	Hemd	463 (35)	952 (23)	490 (17)	Herder	549 (85)	1,037 (107)	504 (13)
	Mand	527 (66)	958 (55)	427 (22)	Masker	549 (25)	1,057 (47)	514 (25)
	Naald	603 (104)	1,086 (114)	483 (38)	Nagel	563 (8)	1,061 (39)	499 (34)
	Pan	620 (39)	896 (28)	276 (13)	Panda	593 (24)	972 (49)	379 (32)
	Rits	481 (32)	1,038 (36)	571 (25)	Ridder	589 (95)	1,127 (114)	539 (26)
	Schaar	419 (47)	964 (53)	545 (12)	Schaduw	451 (70)	1,065 (58)	614 (18)
	Vlieg	558 (258)	1,063 (262)	503 (22)	Vliegtuig	435 (24)	1,207 (45)	767 (39)
	Zee	528 (34)	965 (37)	442 (15)	Zebra	581 (74)	1,142 (81)	562 (32)
	M	525	990	467		539	1,084	547
Two words	Hemd	580 (124)	989 (125)	882 (91)	Herder	689 (81)	1,107 (76)	977 (29)
	Mand	573 (17)	949 (28)	867 (77)	Masker	685 (131)	1,143 (106)	997 (86)
	Naald	620 (57)	1,020 (44)	846 (49)	Nagel	672 (65)	1,146 (80)	981 (54)
	Pan	693 (45)	985 (40)	776 (65)	Panda	653 (41)	1,013 (64)	885 (126)
	Rits	615 (63)	1,064 (52)	924 (57)	Ridder	749 (66)	1,220 (63)	1,024 (120)
	Schaar	584 (101)	1,084 (77)	950 (58)	Schaduw	613 (108)	1,195 (113)	1,121 (110)
	Vlieg	547 (27)	995 (48)	910 (50)	Vliegtuig	556 (71)	1,120 (66)	1,068 (38)
	Zee	632 (43)	1,061 (57)	942 (46)	Zebra	637 (50)	1,199 (88)	1,079 (39)
	M	606	1,018	887		657	1,143	1017

Note. Standard deviation in parentheses following each mean.

not appear together. Confederate items were pseudorandomly combined with participant items following the constraints that each of the confederate nouns appeared no more than once with any participant noun across the entire experiment, and such that each participant item appeared an equal number of times with mono- and disyllabic nouns in one- and two-word utterances. The resulting 128 items were then divided into blocks such that each block contained two instances of each noun as the first word in the trial, and such that each block contained one token of each participant noun. The order of experimental blocks was counterbalanced across lists and trials were displayed in a different random order per participant within each of these blocks. Because of experimenter error, 10 participants were run in two lists, nine in one list and 11 in the final list.

Apparatus and Procedure

The apparatus was as described in Experiment 1. The experiment followed the same procedure as described in Experiment 2, but the interval during which participants could make their response was increased to 3,000 ms. The experiment was presented using Presentation software (Version 18.3; Presentation, 2004); otherwise, stimulus presentation was identical to Experiment 2.

Analysis

As in Experiments 1 and 2, the participants' utterances were transcribed and the utterance onsets were determined using Praat (Boersma & Weenink, 2017). Fixations to the confederate picture(s) and participant's picture were classified as those registered within the 200 × 200 area containing each picture. This means that there were two discrete areas of interest for the confederate items in two-word trials; looks to both were pooled for analysis. Gazes were otherwise calculated as described in Experiment 1. Mixed-effect models were calculated for eye-voice lead and turn gap as described in Experiment 1, with the predictors of block context (mixed, pure) and confederate length (mono/disyllabic) contrast-coded as described in Experiment 1 and the predictor of confederate word number (one word, two words) was contrast coded as (0.5, −0.5). Cross-validation analyses were performed as described in Experiment 1 except that the predictor of confederate word number replaced occlusion.

Results and Discussion

Trials were excluded from further analysis if the participants provided a wrong label (66 trials), an incomplete response (two trials), or no response (39 trials). This left 5,013 trials for analysis of turn gaps. An additional 395 trials were excluded from analysis of eye-voice lead because no fixation was registered to the participant's picture, leaving 4,618 trials.

In the aggregate, coordination in utterance launching was slightly improved compared with Experiments 1 and 2. As shown in Figure 5, the median turn gap was 193 ms (1st quantile = 112 ms; 3rd quantile = 288 ms); compared with Experiment 2, this meant that utterances were launched faster (with a smaller median onset) and more precisely (with a smaller interquartile interval). The duration of the prompts again had a noticeable effect, with shorter turn gaps after disyllabic than monosyllabic prompts and after two-word than one-word prompts.

In all conditions, more visual attention was paid to the confederate items than in either of the previous experiments. As shown in Table 12, participants fixated the confederate picture 31% of the time during one word trials and 37% of the time during two-word trials, and as a whole, 76% of trials received one or more fixations to the confederate picture. The mono- and disyllabic one-word trials in the pure one-word blocks in this experiment were identical to the overt trials in mixed blocks in Experiment 2, which received only 17% of all fixations (see Table 7). This suggests that the low rates of fixations to the confederate picture in the two previous experiments were in part because of the fact that the picture was occluded half of the time; therefore, directing a fixation to the confederate picture was not so useful.

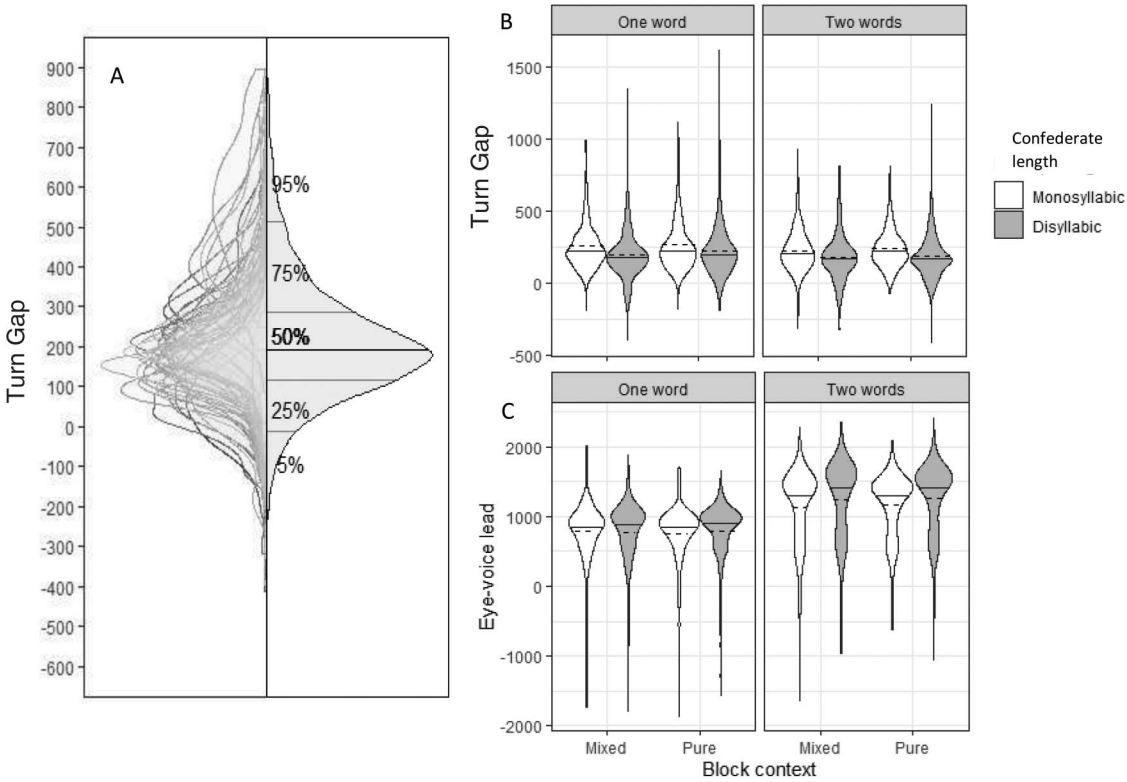
As in Experiments 1 and 2, there were two dominant fixation patterns: on slightly less than half of the trials where the confederate picture(s) received any fixation, the first gaze was directed to the confederate picture(s) at the onset of the trial, and remained there for about 375 ms, at which point a fixation was directed to the participant's own picture for a much longer duration, with attention remaining there for nearly the entire trial. On the remainder of the trials with fixations to the confederate pictures, participants first looked at their own picture, remained there for about 425 ms, and then fixated the confederate picture(s) for 600 to 700 ms; this was often followed by another gaze to the participant's own picture until the end of the trial. These time-courses are remarkably similar to Experiment 2, which meant that participants tended to follow the same apprehension strategies. What differed was which strategy was chosen most often: compared with Experiment 2, more participants in Experiment 3 looked first at their own image and then the confederate image, and fewer participants looked only at their own image.

To pursue the consequences of these apprehension strategies for speech planning, we examined how mean turn gap differed based upon how the participant chose to fixate the images. On trials in which the confederate image(s) received the first fixation, followed by the participant image, the mean turn gap was 261 ms (*SD* 180), while on trials in which the participant image received the first fixation, followed by the confederate image(s), the mean turn gap was 202 ms (*SD* 168 ms), and on trials during which the participants only fixated their own image it was 194 ms (*SD* = 146). This suggests that corepresenting the participant's stimuli is not only unnecessary for tight coordination, but in fact can hinder one's own production if done at the wrong time (see, e.g., Brehm et al., 2019, for similar arguments).

As in Experiments 1 and 2, eye-voice lead was reliably shorter for shorter confederate prompts. As shown in Figure 5, eye-voice lead was reduced by 424 ms for one-word prompts compared with two-word prompts, and by 55 ms for monosyllabic items compared with disyllabic items. These factors interacted such that the eye-voice lead advantage for disyllabic items was small in one-word trials (a 7 ms advantage for disyllabic words), and large in two-word trials (a 102 ms advantage for disyllabic words). This likely follows from the length of these items. As the difference in onset and duration between one and two-word stimuli was largest for the disyllabic items, the disyllabic two-word stimuli left participants the longest interval in which to plan their utterances. An interaction between block context and word number was also observed, such that the word number effect on eye-voice lead was 32 ms larger in pure than mixed blocks. This means that as in Experiment

Figure 5

Distribution of Overall Turn Gap Split by Participant (Panel A, Left) and Pooled, With Quantiles (Panel A, Right), Turn Gap Times by Condition (Panel B) and Eye-Voice Lead Times (Intervals Between First Look to Participant's Stimulus and Participant's Speech Onset; Panel C) in Experiment 3 by Confederate Word Number, Block Context, and Confederate Prompt Length



Note. In panels B and C, dashed lines reflect condition mean and solid lines reflect condition median.

1, participants took more time to plan and initiate their utterances when the prompt was predictably long. These patterns were all confirmed using mixed-effect models (see Table 13).

As in Experiments 1 and 2, turn gaps were reliably shorter after longer prompts, with advantages for two-word versus one-word prompts (27 ms) and disyllabic versus monosyllabic prompts (50 ms). This was confirmed using mixed-effect modeling (see Table 13). Word length and word number did not interact, meaning that the increased planning time afforded by two-word disyllabic items did not confer any additional advantage for launching; this is consistent with the previous experiments and underscores the separation of launching from planning in this experimental paradigm. However, there was an observed interaction between word length, word number, and block context such that for one-word utterances, confederate prompt length mattered more in mixed (68 ms) than pure blocks (38 ms), whereas for two-word utterances, confederate prompt length mattered more in pure (53 ms) than mixed blocks (42 ms). This means that, as in Experiments 1 and 2, the long prompt advantage was greater when there was less information about the prompt. In this experiment this was the one-word prompts in mixed blocks, because for two-word prompts in mixed blocks, the overall length of the item could be estimated from hearing the first word because both words were either mono- or disyllabic. Note that this interaction was not present for eye-voice lead, suggesting the long prompt advantage for one-word prompts differently affected launching and planning.

As in the previous two experiments, we ran two series of cross-validation analyses. In the first series, reported in Table 14, we examined which properties of the confederate stimulus best predicted turn gap. Replicating both previous experiments, the model predicting turn gap from the onset and offset time of the confederate prompt performed the best, followed by the model predicting turn gap from the offset of the confederate prompt only. This again provides evidence that it is the information conveyed within the current utterance—not recent experience—that speakers attend to while preparing to launch speech, even when the number of words in the utterance is variable.

In a second series of cross-validation analyses reported in Table 15, we added trial-level predictors to all models that performed reliably better than the baseline model. Adding confederate word number to all models reliably improved model fit, and it did so especially for the model including confederate onset and offset time. This again confirms that the number of words in the prompt has a strong impact on the turn gap. In contrast, adding the block context predictor reliably improved only one model: the random intercept only model. This suggests again that the predictability of

Table 12
Fixations and Gaze Data by Condition in Experiment 3

Condition			All looking time (%)						Gaze in time		Gaze out time	
Block context	Confederate length	C word number	C	P	Trials with C fixations (%)		First gaze (%)		C	P	C	P
Mixed	Short	One	30%	70%	Yes	78%	C	46%	236	724	612	2,137
							P	54%	863	240	1,578	709
					No	22%	P	100%	—	479	—	2,255
		Two	C1: 17% C2: 20%	63%	Yes	81%	C	49%	293	1,098	674	2,281
							P	51%	819	234	1,430	645
					No	19%	P	100%	—	540	—	2,460
	Long	One	32%	68%	Yes	75%	C	47%	261	804	692	2,230
							P	53%	839	242	1,653	707
					No	25%	P	100%	—	645	—	2,316
		Two	C1:18% C2: 23%	59%	Yes	84%	C	49%	283	1,147	669	2,302
							P	51%	812	233	1,390	689
					No	16%	P	100%	—	570	—	2,524
Pure	Short	One	29%	71%	Yes	69%	C	51%	246	749	612	2,067
							P	49%	901	245	1,789	745
					No	31%	P	100%	—	485	—	2,414
		Two	C1: 15% C2: 20%	66%	Yes	74%	C	47%	276	1,026	643	2,279
							P	53%	824	254	1,462	672
					No	26%	P	100%	—	377	—	2,397
	Long	One	32%	67%	Yes	72%	C	48%	251	789	655	2,101
							P	52%	860	248	1,825	712
					No	28%	P	100%	—	568	—	2,476
		Two	C1: 15% C2: 21%	64%	Yes	76%	C	46%	292	1,111	643	2,323
							P	54%	876	246	1,435	717
					No	24%	P	100%	—	412	—	2,565

Note. C = confederate picture (in two-word condition, this represents combined looks to both pictures); P = participant picture; C1 = first confederate picture; C2 = second confederate picture.

offset time from onset time subsumes the variance that is accounted for by block context in the by-condition mixed-effect analyses.

Similar to Experiments 1 and 2, the best performing model predicted turn gap from the confederate onset, confederate offset, and the number of words in the prompt in addition to a random intercept by participant. One sample run of this model is shown in Figure 6. In this model, the mean absolute error per point was 122.24 ms. For trials between the 5% and 95% quantiles, the observed data have an MSPE of 8341.55 and a mean absolute error per point of 69.55 ms; for data in the tails of the observed distribution, the observed data are predicted less well, with an MSPE of 46,706.29 and mean absolute error per point of 189.10 ms for the left tail and an MSPE of 103007.00 and mean absolute error per point of 261.10 ms for the right tail.[2] This is a slightly better performance in the middle 90% and a slightly worse performance in the tails than in Experiment 2, likely resulting from the closer clustering of observed turn gaps near the mode in Experiment 3.

To summarize, Experiment 3 replicates the key findings of Experiments 1 and 2. Removing the occlusion factor caused participants to look more often at the confederate items, but the increase in looks occurred most reliably after the participant's image was first fixated, and on these trials, synchronization was markedly worse. This underscores that corepresentation is not necessary for coordination, and that it can even hinder one's own speech planning (Brehm et al., 2019; Hoedemaker & Meyer, 2019).

Turn gaps were again clustered near 200 ms, and as seen in the two previous experiments, the most important cue to accurate launching was the length of the prompt utterance. Despite the fact that half of the prompts contained two words, we obtained a predictive cross-validation model with an equally good fit as in the earlier experiments. This shows again how speakers can time their speech onset with minimal corepresentation of the prior utterance, relying instead on cues directly contained in the speech input.

General Discussion

Part of the challenge of having a conversation is to decide when it is appropriate to speak. Timing is everything: it is neither appropriate to interrupt nor to be overly slow to respond. As many authors have pointed out, to achieve smooth turn taking, speakers must plan their utterances on time and launch them at the appropriate moment, such that they begin just after the end of the preceding turn (see Sacks et al., 1974, for foundational work on the topic). The current work explored the cues used to time utterance launching. In a simple production task, participants were asked

[2] There is a cloud of points with mis-predicted values between 600 and 700 ms: these all come from a single participant. This person had slower, more variable responses than the other participants (median gap = 628 ms; 1st quartile = 510, 3rd quartile = 800 ms; see Figure 5A for distributions by participant). Excluding this participant reduced the mean error per point by about 3 ms across the board, but within each series of models, the rankings were identical for any models with improved BIC over their respective baseline.

Table 13
Outputs of Linear Mixed-Effect Models for Eye-Voice Lead and Turn Gap in Experiment 3

Fixed effects	Eye-voice lead			Turn gap		
	Estimate	*SE*	*t* value	Estimate	*SE*	*t* value
Intercept	966.33	44.52	**21.70**	218.35	18.90	**11.55**
Word number	−445.67	35.05	**−12.71**	27.79	3.37	**8.26**
Block context	10.56	11.80	0.89	13.97	13.61	1.03
Confederate length	−60.30	11.77	**−5.12**	49.44	3.36	**14.70**
Word Number × Block Context	−51.48	23.61	**−2.18**	9.43	6.73	1.40
Word Number × Confederate Length	93.16	23.53	**3.96**	3.21	6.73	0.48
Block Context × Confederate Length	−15.63	23.54	−0.66	−8.28	6.73	−1.23
Word Number × Block Context × Confederate Length	−91.79	47.07	−1.95	−39.11	13.46	**−2.91**

Random effects	Term		*SD*	Term		*SD*
	Participant	Intercept	273.83	Participant	Intercept	113.05
		Word number	208.27	Item	Intercept	23.63
	Item	Intercept	33.82		Block context	52.76
	Residual		399.54	Residual		119.08

Note. Bold values reflect *t*-values above 2.

to launch speech immediately following the offset of a prerecorded confederate prompt so as to minimize the gap between the two turns. The content of the participant's utterance was not contingent on the content of the prompt and the recorded prompt began at a naturalistic delay from trial onset. This meant that participants had ample time to plan their utterance, but the onset and offset times of the prior utterance varied enough to make launch challenging.

We performed three experiments using this paradigm, asking how the coordination of utterance launching was impacted by the ease of corepresenting the confederate's utterance content versus the predictability of the prompt's timing. This allowed us to ask whether knowing ahead of time *which* word a person will say is what affords tight coordination—or if one may be instead reasonably accurate in launching by doing nothing more than using cues drawn from the observed utterance as it unfolds.

Corepresentation Is Useful, But Not Required to Launch Preplanned Speech

In the first two experiments, we assessed the role of corepresentation in coordination by manipulating whether participants could or could not see the stimuli the confederate was naming. This tested whether making it easy to mentally represent the prompt utterance improves coordination in utterance launching. The results of both experiments were similar: even without being able to see the confederate item, participants were able to launch their utterances accurately, with an increase in response time of only 15–20 ms for occluded prompts, and turn gaps well within the estimated range from corpus studies (Heldner & Edlund, 2010; Levinson & Torreira, 2015; Stivers et al., 2009; Weilhammer & Rabold, 2003).

Corepresentation reliably improved coordination. In Experiments 1 and 2, turn gaps were reliably shorter when the confederate's stimulus was overt than when it was occluded. Numerically, this effect was stronger in Experiment 2, using picture naming, than in Experiment 1, using reading aloud of numbers. Tallies of participants' eye gaze pattern showed that they looked at the partner's item on a minority of trials but chose to do so most often before looking at their own. Combined, these results suggest that having visual information available supported utterance launching—but only modestly, such that corepresentation was useful but not required for efficiently launching a preplanned utterance.

Table 14
Prediction of Turn Gap by Cross-Validation for Experiment 3 Ranked by Model Fit

Rank (BIC)	Rank (MSPE)	Model parameters	BIC	ΔBIC	MSPE	ΔMSPE	M|Err|
1	1	Confederate onset + Confederate offset	52844.7	242.87[a]	15267.84	1003.43	123.31
2	2	Confederate offset	52857.82	229.75[a]	15357.31	913.96	123.67
3	3	Average last 5 confederate onsets	53066.98	20.59[a]	16148.65	122.62	126.85
4	4	Average last 5 confederate onsets + Average last 5 confederate offsets	53079.04	8.53[a]	16155.43	115.84	126.87
5	5	Confederate onset	53085.91	1.66	16219.25	52.02	127.12
6	7	Random intercept only	53087.57	—	16271.27	—	127.33
7	6	Average last 5 confederate offsets	53093.1	−5.54	16251.63	19.64	127.25

Note. All models included a random intercept by participant. Delta Bayesian information criterion (BIC) and delta mean squared prediction error (MSPE) are calculated by comparison with random-intercept only model.
[a] Models with reliably improved BIC.

Table 15

Prediction of Turn Gap by Cross-Validation for Experiment 3 With Addition of Word Number and Block Context Parameters, Each Series Ranked by Model Fit

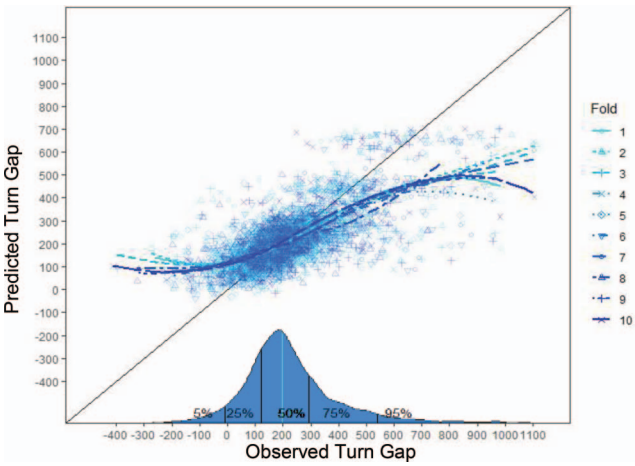
Rank (ΔBIC)	Rank (ΔMSPE)	Model parameters	BIC	ΔBIC	MSPE	ΔMSPE	M|Err|
		Confederate word number models					
1	1	Confederate onset + Confederate offset + Word number	52773.1	71.60[a]	15002.97	264.87	122.24
2	2	Random intercept + Word number	53027.03	60.54[a]	16030.66	240.61	126.38
3	3	Average last 5 confederate onsets + Word number	53034.22	32.76[a]	16017.34	131.31	126.33
4	4	Average last 5 confederate onsets + Offsets + Word number	53046.65	32.39[a]	16025.66	129.77	126.36
5	5	Confederate offset + Word number	52835.66	22.16[a]	15270.12	87.19	123.33
		Block context models					
1	1	Random intercept + Block context	53079.22	8.35[a]	16231.95	39.32	127.17
2	2	Average last 5 confederate onsets + Average last 5 Confederate offsets + Block context	53076.36	2.68	16138.4	17.03	126.8
3	3	Average last 5 confederate onsets + Block context	53065.15	1.83	16134.66	13.99	126.79
4	4	Confederate offset + Confederate onset + Block context	52844.44	0.26	15260.23	7.61	123.28
5	5	Confederate offset + Block context	52858.91	−1.09	15354.74	2.57	123.66

Note. All models included a random intercept by participant. Delta Bayesian information criterion (BIC) and delta mean squared prediction error (MSPE) are calculated by comparison with a nested model without the added parameter (word number, block context).
[a] Models with reliably improved BIC.

The infrequent gazes to the confederate's pictures in Experiments 1 and 2 led us to run Experiment 3, in which the confederate's pictures were always overt. Here, coordination was on average a bit better, improving by about 10 ms at the group level. As expected, participants looked at the confederate's pictures more often than in Experiments 1 and 2. Participants' apprehension strategies also changed in this experiment, such that they fixated the confederate picture nearly as often after fixating their own picture as they did before fixating their own picture. Note though that Experiment 3 also differed from the earlier experiments by including trials where the confederate named two pictures. Thus, it is unclear whether the consistent visibility of the confederate's pictures, the inclusion of the two-picture trials, or both led to the higher rate of gazes to the confederate pictures and the difference in apprehension strategies.

Figure 6

Observed Versus Predicted Turn Gaps for Best-Fitting Cross-Validation Model in Experiment 3 Across 10 Randomized Folds for a Sample Iteration



Note. Model is: Turn Gap ~ Confederate Onset + Confederate Offset + Confederate Word Number + (1 Participant). Points reflect trial-level data; lines reflect loess smooths for each cross-validation fold. Density plot at bottom reflects distribution of observed data (on arbitrary Y scale), with vertical lines reflecting quantiles. See the online article for the color version of this figure.

Comparing the prompts included in both Experiment 2 and 3 further highlights that seeing the confederate's picture had minimal effect on participants' utterance timing. The same prompts were included in both Experiment 2 as "mixed" mono- and disyllable blocks and Experiment 3 as "pure" one-word blocks. While the overall proportions of fixations to confederate pictures in these trials differed between experiments, the mean turn gap was nearly identical: 237 ms (*SD* 189) for Experiment 2 and 241 ms (*SD* 180) for Experiment 3. Furthermore, a comparison of the duration of turn gaps on the Experiment 3 trials where participants looked early versus late at the confederate picture(s) showed that fixating the confederate picture early within the trial hindered rather than helped, and that there was no turn gap benefit compared with trials on which no fixations were directed to the confederate picture(s) at all.

In summary, in line with earlier work by Corps and colleagues (2018, 2020), it is evident from our three studies that access to visual information and, therefore, easy corepresentation of the prompt utterance, played only a small role in utterance launching in this paradigm.

Coordinating via Stimulus Length

In all experiments, we also manipulated the prompt's length and its predictability. The goal was to examine how these cues contributed to coordinated launching. Earlier work suggests that response length impacts the ease of identifying turn end, a precondition for timely launching (e.g., Corps et al., 2020), as do an assortment of speech properties (e.g., Barthel et al., 2017; Bögels & Torreira, 2015; Cutler & Pearson, 1985; Duncan, 1972; Gravano

& Hirschberg, 2011; Grosjean & Hirt, 1996; Schaffer, 1983). Consistent with this work, prompt length had large effects: in all three experiments, participants launched their speech with smaller turn gaps for longer than shorter prompts. This replicates the role of prior turn length for timely launching, and highlights that information provided in the speech signal alone is sufficient to produce simple coordinated utterances.

Beyond the main effects of prompt length, length interacted inconsistently with block context. We had predicted that the length effects on eye-voice lead and turn gap would be most pronounced for pure compared with mixed blocks because participants would benefit from constant preparation periods between precues and response signals (e.g., Niemi & Näätänen, 1981; Rolke & Ulrich, 2010; Teichner, 1954; Woodrow, 1914). Following this prediction, the prompt length effect on turn gaps was stronger in pure blocks in Experiment 1 and for the two-word utterances in Experiment 3, however, it was stronger in mixed blocks in Experiment 2 and for the one-word utterances in Experiment 3. For eye-voice lead, the prompt length effect was stronger in pure blocks for Experiment 1 and Experiment 3, but stronger in mixed blocks in Experiment 2. In other words, while consistently long prompts often improved coordination, the results were not fully systematic within or between experiments. Combined with the lack of a main effect of block context and the fact that block context never improved any cross-validation models, the implication is that while participants might in theory be able to respond faster given uniform preparation intervals, they did not do so consistently in our paradigm.

Across experiments, we varied the duration of the prompts in different ways, in terms of word duration (Experiment 1), syllable number (Experiments 2 and 3), and word number (Experiment 3). This allowed speakers to potentially form predictions of utterance length via multiple cue types—not just in terms of utterance duration, but also in terms of syllable number and word number. However, in our experiments, these cues were always confounded with prompt duration, and they also tended to be confounded with prompt onset time. More work is needed to determine the impact of linguistic structure and temporal duration on the launching of utterances in tightly controlled laboratory contexts. More work is also needed to determine which cues are exploited in conversations where turns vary substantially more in length but where speakers can draw on a host of linguistic and paralinguistic cues not considered here. Our results indicate that in the absence of other information, speakers very effectively exploit low-level speech cues to time their utterances; the question is how this generalizes to other cues and to more naturalistic situations.

A Cross-Validation Approach to Coordination

To directly contrast the use of speech cues and corepresentation based on pictorial information in coordination and move beyond evaluating condition-level differences, we used a cross-validation approach to assess which combination of utterance properties best predicted response launch. This novel approach provides further insight into why predictably long utterances afforded the most coordinated launching and the relative role of corepresentation in launching. In these analyses, we built predictive models from a subset of the data, evaluated which linear combination of cues most accurately predicted the turn gap in held-out data, and assessed which cues accounted for the most unique variance to demonstrate their relative importance.

The cross-validation analyses showed that the best single cue for predicting response launch was the confederate offset time, treated as a continuous measure, with a smaller amount of additional independent variance accounted for by confederate onset time. This suggests that an estimate of the time point at which an utterance will end is the single most useful piece of information for launching timely speech. This result is unsurprising, given that the best way to tell that an utterance has ended is to hear that it has ended. However, our data are also consistent with participants predicting turn end using, for example, the onset of the final syllable of a closed set of words to accurately project offset time by inferring which word is being spoken. Contrasting these possibilities is worthy of future work.

Cross-validation analyses also showed that knowing when the utterance has started also provides unique and important information. As discussed above, the onset of the prompt can be seen as a precue to the offset, or, relatedly, as a trigger for utterance launching; knowing the onset could provide a useful cue. In addition, perhaps knowing the onset also allowed the participant to predict which word out of the relatively small set of items would be said, allowing the participant to project the prompt's offset time.

In the cross-validation analyses, we contrasted trial-level predictors with predictors made out of recent averages to see which built a better model, and found that trial-level cues of onset and offset always contributed more to an accurate model fit. This implies that it is most useful to know specifically how long an utterance is and not simply that the utterance will be relatively long. This meshes with earlier arguments in the literature (see, e.g., Corps et al., 2018, 2020).

For cross-validation models that included utterance offset time, it was also useful to include whether or not the picture was visible. This returns to the question of occlusion: what is it that seeing the picture does? By taking a model-building approach, we observed that the best models of launching are built when both utterance length and utterance content are known. Combined with the small but reliable effect of occlusion observed at the condition level in Experiments 1 and 2, this demonstrates that corepresentation is used to a predictable degree when it is afforded. While speakers do not *need* to corepresent, they *can* do so, and this improves coordination. This may be because corepresentation has beneficial consequences for aligning discourse partners across multiple representational levels (e.g., as in the Pickering & Garrod, 2004, 2013 framework) or because seeing the picture leads to rapid activation of its name, which in turn facilitates understanding the confederate's utterance and ultimately, launching. However, it could also be because knowing utterance content affords precise estimates of launch: knowing exactly what utterance will occur by definition means having a reasonable approximation of how long its duration will be. These are possibilities to be investigated in future work.

In the models of all three experiments, accounting just for properties of trial timing (onset, offset) and content (occlusion) served to build highly accurate models of utterance launch. The best-fitting cross-validation model for each experiment had a mean error of about 100 ms per point. This means that our model performed as well as the limits of human performance. Any movement takes time to program, and this programming time ranges between 100 and 200 ms, depending on the motor response.

Simple auditory response times via button-press take 150 ms on average for healthy young men (as measured by Galton in 1885, and assessed statistically in Johnson et al., 1985). It also takes 150 to 200 ms to launch an eye-movement to a visual stimulus (e.g., Salthouse & Ellis, 1980; Saslow, 1967). In psycholinguistic research, one does not typically observe linguistically mediated differences in eye-movements between conditions earlier than 200 ms from stimulus onset (e.g., Allopenna et al., 1998). Responses that require some decision take longer still: Donders estimated repetition of a known syllable to take 200 ms, and repetition of an unknown syllable to take 284 ms (Donders, 1969).

The limits of human motor movements in turn suggest a reason for why the commonly observed 200 ms turn gap is not a 0 ms turn gap. Achieving a 200 ms gap would require noticing the offset of the prior utterance at most 50 ms after it has actually occurred, and achieving a 0 ms gap would require predicting the time of utterance end at minimum of 150 ms before it ends. Therefore, we suggest that the frequently observed modal gap of 200 ms is approximately as small as humans can achieve if they aim to respond immediately to the offset of a turn, and so, conversations can proceed about as fast as our cognitive and motor processes allow.

Flexibility in Forming Representations for Responding

Situating our findings more broadly within the field of psycholinguistics, our primary finding is that speakers can produce a remarkably accurate launch in a simple paradigm without necessarily drawing upon detailed corepresentations of the discourse partner. This converges with existing work on joint production showing that sparse, not full, corepresentations are often formed, and that even partial corepresentation of the partner's speech task can result in interference to one's own speech production (e.g., Brehm et al., 2019; Gambi et al., 2015; Hoedemaker & Meyer, 2019). This suggests that while corepresenting a partner's utterance may facilitate some aspects of production (e.g., planning what to say, as in Corps et al., 2018), corepresentation may also have detrimental consequences. Speakers might be best served by deploying corepresentations strategically, rather than by default, and may need to change how they respond based upon the task. This makes coordinating the timing of conversation more similar to the joint action of lifting a piano than the joint action of playing a piano duet: It is useful to coarsely represent what the partner is doing, and aiming for very fine-grained temporal predictions might be counterproductive.

The present research also connects with a literature that underscores how listeners do not always predict upcoming utterances in full detail. While listeners and readers predict a set of likely upcoming words based upon their semantic category (see, e.g., Federmeier & Kutas, 1999), there is conflicting evidence concerning the prediction of orthographic or phonological forms, with some work supporting full prediction (e.g., DeLong et al., 2005; Dikker et al., 2010; Laszlo & Federmeier, 2009; Van Berkum et al., 2005), and recent work challenging these findings (e.g., Nieuwland et al., 2018). Thus, it appears that individuals can rely on an assortment of predictive processes, but very few of them are uniquely necessary to achieve adequate comprehension (e.g., Huettig, 2015; Huettig & Guerra, 2019).

Combined, these facets of the literature showcase the importance of thinking about corepresentation in conversation not as a dichotomous factor, but as something that can be turned up or down, perhaps differently across levels of linguistic representation. As such, we posit that the fully aligned, rich corepresentation of a partner's mental state such as in the Pickering and Garrod (2013) or the Levinson and Torreira (2015) framework might reflect only one end of a continuum; the other end might involve little to no corepresentation of the partner. This has precedent in the joint action literature, where minimal corepresentation has been posited to allow tight coordination (e.g., Vesper et al., 2010). An outstanding research question is which linguistic, cognitive, and social variables affect the level of detail interlocutors corepresent about each other's utterance plans in natural conversation.

Linking Experiments With Conversations

In our experiments, the participants' responses were faster (by 100 to 350 ms) than in earlier laboratory studies where participants also aimed to respond as quickly as possible to preceding turns (Corps et al., 2018, 2020; Meyer et al., 2018). As highlighted before, in the present study, the participants had ample time to prepare their utterance during the preceding turn, with an average of 1,241 ms elapsing between trial onset and the offset of the prerecorded confederate speech. An obvious account for the longer gaps seen in the earlier studies is that in those studies participants did not have time to complete utterance planning by the end of the preceding utterance, even though the response set was often small (e.g., "yes" or "no"). Delays could arise for many reasons, for instance because the cue to the answer appeared relatively late in the utterance or because thinking of an appropriate answer and formulating it was time consuming. For example in a "quiz show" paradigm where key information was presented early versus late (Bögels et al., 2015), participants apparently needed several seconds to retrieve the answer regardless of the position of the cue.

What is more remarkable than the differences from earlier laboratory studies is that the central tendency in our experiments was the same as in corpora of conversational speech: around 200 ms. Throughout this article, we have argued that because of the simplicity of our paradigm, participants had usually fully planned their utterance before the end of the confederate's turn, and then only had to launch it. This allowed them to respond with latencies around 200 ms. We suggest that the same may often be true for conversation: 200-ms gap durations arise when speakers have completed planning at least the initial part of their utterance by the end of the preceding turn. An important question for further research is how they manage to do this. There is probably no simple answer to this question. More likely speakers can freely use an assortment of planning strategies, such as drawing upon rich multimodal conversational contexts, planning in a highly incremental fashion, and replying with short, minimally informative utterances (see, e.g., Holler & Levinson, 2019). How speakers use these degrees of freedom in conversation needs to be established in further work.

Conclusion

When asked to speak as soon as a recorded utterance had ended, participants in our experiments based their utterance launching primarily on speech cues from the prior utterance, specifically the offset and onset of a prompt. Knowing ahead of time what the utterance would be also supported fast responding, though to a lesser degree. This meant that listening was more important for precise launching than corepresentation of utterance content in this paradigm. The

implication is that timely utterance launching—at least in simple experimental contexts—can be done with minimal corepresentation of content.

Context of the Research

Speaking in conversations requires juggling one's own linguistic representations for planning with attention to continuously evolving input to produce speech quickly, fluently, and at appropriate times. By focusing on the question of speech launching when planning was trivially easy, our results highlight that speakers can be highly accurate with their speech timing by relying on a very simple set of cues that are more weighted toward prior utterance timing than content. This suggests that the timing of real-world conversations might also be reliant on relatively few cues. Our method provides a bridge between highly controlled experimental work and naturalistic, more ecologically valid conversational contexts. The goal of future research will be to strengthen this bridge, establishing what is minimally required for launching and planning speech depending on the needs of the task by testing predictions from naturalistic scenarios in controlled experimental contexts.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

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

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

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

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

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer program]. http://www.praat.org

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

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

Brehm, L., Taschenberger, L., & Meyer, A. (2019). Mental representations of partner task cause interference in picture naming. *Acta Psychologia*, *199*, 102888. https://doi.org/10.1016/j.actpsy.2019.102888

Clark, H. H. (1996). *Using language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

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

Corps, R. E., Gambi, C., & Pickering, M. J. (2020). How do listeners time response articulation when answering questions? The role of speech rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 781–802. https://doi.org/10.1037/xlm0000759

Cutler, A., & Pearson, M. (1985). On the analysis of prosodic turn-taking cues. In C. Johns-Lewis (Ed.), *Intonation in discourse* (pp. 139–155). Croom Helm.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

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

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science*, *21*(5), 629–634. https://doi.org/10.1177/0956797610367751

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431. https://doi.org/10.1016/0001-6918(69)90065-1

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *6*, 174–215. https://doi.org/10.1080/17470218.2017.1310261

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292. https://doi.org/10.1037/h0033031

EyeLink Data Viewer. (2017). Version 3.1.1 [Computer software]. SR Research Ltd.

EyeLink Experiment Builder. (2015). Version 1.10.1630 [Computer software]. SR Research Ltd.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Gambi, C., Van de Cavey, J., & Pickering, M. J. (2015). Interference in joint picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 1–21. https://doi.org/10.1037/a0037438

Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, *6*, 751. https://doi.org/10.3389/fpsyg.2015.00751

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

Grosjean, F., & Hirt, C. (1996). Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language and Cognitive Processes*, *11*(1/2), 107–134. https://doi.org/10.1080/016909696387231

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

Hoedemaker, R. S., & Meyer, A. S. (2019). Planning and coordination of utterances in a joint naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 732–752. https://doi.org/10.1037/xlm0000603

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. https://doi.org/10.1016/j.tics.2019.05.006

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118–135. https://doi.org/10.1016/j.brainres.2015.02.014

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on pre-

diction in language processing. *Brain Research*, *1706*, 196–208. https://doi.org/10.1016/j.brainres.2018.11.013

Johnson, R. C., McClearn, G. E., Yuen, S., Nagoshi, C. T., Ahern, F. M., & Cole, R. E. (1985). Galton's data a century later. *American Psychologist*, *40*(8), 875–892. https://doi.org/10.1037/0003-066X.40.8.875

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

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

Knoblich, G., Butterfill, S., & Sebanz, N. (2011). Psychological research on joint action: Theory and data. *Psychology of Learning and Motivation*, *54*, 59–101. https://doi.org/10.1016/B978-0-12-385527-5.00003-6

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*(3), 326–338. https://doi.org/10.1016/j.jml.2009.06.004

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*, 731. https://doi.org/10.3389/fpsyg.2015.00731

Liu, W., & Yang, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. *Annals of Statistics*, *39*(4), 2074–2102. https://doi.org/10.1214/11-AOS899

Magyari, L., & de Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, *3*, 376. https://doi.org/10.3389/fpsyg.2012.00376

Meyer, A. S., Alday, P. M., Decuyper, C., & Knudsen, B. (2018). Working together: Contributions of corpus analyses and experimental psycholinguistics to understanding conversation. *Frontiers in Psychology*, *9*, 525. https://doi.org/10.3389/fpsyg.2018.00525

Meyer, A. S., Roelofs, A., & Levelt, W. J. (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language*, *48*(1), 131–147. https://doi.org/10.1016/S0749-596X(02)00509-0

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin*, *89*(1), 133. https://doi.org/10.1037/0033-2909.89.1.133

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., von Grebmer Zu Wolfsthurn, S., Bartolizzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., . . . Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–226. https://doi.org/10.1017/S0140525X04000056

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Presentation. (2004). Version 18.3 [Computer software]. Neurobehavioral Systems, Inc. www.neurobs.com

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rolke, B., & Ulrich, R. (2010). On the locus of temporal preparation: Enhancement of premotor processes. In A. C. Nobre & J. T. Coull (Eds.), *Attention and time* (pp. 227–241). Oxford University Press.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236. https://doi.org/10.1068/p5117

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. *Language*, *50*(4), 696–735. https://doi.org/10.1353/lan.1974.0010

Salthouse, T. A., & Ellis, C. L. (1980). Determinants of eye-fixation duration. *The American Journal of Psychology*, *93*, 207–234. https://doi.org/10.2307/1422228

Saslow, M. G. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *Journal of the Optical Society of America*, *57*(8), 1024–1029. https://doi.org/10.1364/JOSA.57.001024

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, *11*(3), 243–257. https://doi.org/10.1016/S0095-4470(19)30825-3

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Severens, E., Van Lommel, S., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica*, *119*(2), 159–187. https://doi.org/10.1016/j.actpsy.2005.01.002

Sidnell, J., & Stivers, T. (Eds.). (2012). *The handbook of conversation analysis* (Vol. 121). Wiley. https://doi.org/10.1002/9781118325001

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

Teichner, W. H. (1954). Recent studies of simple reaction time. *Psychological Bulletin*, *51*(2), 128–149. https://doi.org/10.1037/h0060900

ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, *47*(1–2), 80–86. https://doi.org/10.1016/j.specom.2005.05.009

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467. https://doi.org/10.1037/0278-7393.31.3.443

Vesper, C., Abramova, E., Bütepage, J., Ciardo, F., Crossey, B., Effenberg, A., Hristova, D., Karlinsky, A., McEllin, L., Nijssen, S. R., Schmitz, L., & Wahn, B. (2017). Joint action: Mental representations, shared information and general mechanisms for coordinating with others. *Frontiers in Psychology*, *7*, 2039. https://doi.org/10.3389/fpsyg.2016.02039

Vesper, C., Butterfill, S., Knoblich, G., & Sebanz, N. (2010). A minimal architecture for joint action. *Neural Networks*, *23*, 998–1003. https://doi.org/10.1016/j.neunet.2010.06.002

Weilhammer, K., & Rabold, S. (2003). Durational aspects in turn taking. *Proceedings of the International Conference of Phonetic Sciences*. Barcelona, Spain.

Woodrow, H. (1914). The measurement of attention. *Psychological Monographs*, *17*(5), i–158. https://doi.org/10.1037/h0093087

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006

(*Appendices follow*)

Appendix A

Supplemental Analyses

We conducted supplemental analyses that added production length and its interactions to the mixed-effect models of eye-voice lead and turn gap for all three experiments. While we made no predictions for this factor, these analyses follow below for completeness.

First, there were no additional effects in Experiment 1 for either eye-voice lead or turn gap. This appears in Table A1.

In Experiment 2, there were three additional interactions for the eye-voice lead model. The interaction between block context and production length showed that in pure blocks only, the eye-voice lead was longer for one-syllable productions. The interaction between confederate length and production length showed that the increase in eye-voice lead for disyllabic p was greatest for monosyllabic productions. The interaction between occlusion, block context, and production length showed that while pure blocks tended to elicit a longer eye-voice lead than mixed blocks, this was not the case when the participant produced a disyllabic word and the confederate item was occluded. There were also three additional interactions in the turn gap model. The interaction between occlusion and production length showed that turn gap was shorter for disyllabic productions with occluded stimuli. The interaction between confederate length and production length showed that turn gap was longer for monosyllabic productions with monosyllabic prompts, and the interaction between occlusion, block context and production length showed that turn gap was longer for monosyllabic productions in mixed blocks with occluded stimuli. These appear in Table A2.

In Experiment 3, there were two additional interactions for the eye-voice lead model. The confederate word number by block context interaction was qualified by production length such that the two-word eye-voice lead increase in pure blocks was carried by cases in which the participant produced a monosyllabic word. There was also an interaction between block context, confederate length, and production length such that the disyllabic eye-voice

Table A1

Outputs of Linear Mixed-Effect Models for Eye-Voice Lead and Turn Gap in Experiment 1 Containing Production Length

	Eye-voice lead			Turn gap		
Fixed effects	Estimate	*SE*	*t* value	Estimate	*SE*	*t* value
Intercept	834.52	64.95	**12.85**	219.85	27.97	**7.86**
Occlusion	28.40	9.08	**3.13**	21.23	2.83	**7.53**
Block context	−2.64	9.09	−0.29	1.74	4.69	0.37
Confederate length	−145.49	16.22	**−8.97**	53.95	5.43	**9.94**
Production length	−25.74	29.70	−0.87	2.00	26.55	0.08
Occlusion × Block Context	19.25	18.16	1.06	5.14	5.66	0.91
Occlusion × Confederate Length	8.36	18.16	0.46	9.77	5.66	1.73
Block Context × Confederate Length	−52.87	18.17	**−2.91**	−25.40	5.66	**−4.49**
Occlusion × Production Length	−13.51	18.16	−0.74	3.6100	5.66	0.64
Block Context × Production Length	−25.68	18.15	−1.42	3.97	5.66	0.70
Context Length × Production Length	8.73	18.16	0.48	0.65	5.66	0.11
Occlusion × Block Context × Confederate Length	−8.40	36.32	−0.23	1.33	11.32	0.12
Occlusion × Block Context × Production Length	44.87	36.31	1.24	2.46	11.32	0.22
Occlusion × Confederate Length × Production Length	19.60	36.31	0.54	8.84	11.32	0.78
Block Context × Confederate Length × Production Length	−26.75	36.31	−0.74	−13.80	11.32	−1.22
Occlusion × Block Context × Confederate Length × Production Length	41.69	72.63	0.57	19.12	22.64	0.85

Random effects	Term		*SD*	Term		*SD*
	Participant	Intercept	399.84	Participant	Intercept	155.70
		Confederate length	84.48		Block context	23.63
	Item	Intercept	39.99		Confederate length	29.29
	Residual		309.90	Item	Intercept	37.34
				Residual		100.37

Note. Bold values reflect *t*-values above 2.

(*Appendices continue*)

Table A2

Outputs of Linear Mixed-Effect Models for Eye-Voice Lead and Turn Gap in Experiment 2 Containing Production Length

	Eye-voice lead			Turn gap		
Fixed effects	Estimate	*SE*	*t* value	Estimate	*SE*	*t* value
Intercept	991.79	31.95	**31.04**	241.34	21.93	**11.01**
Occlusion	90.13	8.92	**10.10**	13.78	3.50	**3.93**
Block context	15.19	8.94	1.67	5.42	3.50	1.55
Confederate length	36.77	9.00	**4.09**	54.66	3.53	**15.48**
Production length	16.82	25.80	0.65	15.94	12.03	1.33
Occlusion × Block Context	5.87	19.60	0.30	5.03	7.72	0.65
Occlusion × Confederate Length	54.50	18.39	**2.96**	1.83	7.23	0.25
Block Context × Confederate Length	48.08	17.97	**2.68**	24.48	7.06	**3.48**
Occlusion × Production Length	25.40	17.85	1.42	19.95	7.00	**2.85**
Block Context × Production Length	58.05	17.86	**3.25**	2.80	7.01	0.40
Confederate Length × Production Length	61.02	18.00	**3.39**	14.41	7.06	**2.04**
Occlusion × Block Context × Confederate Length	57.09	37.34	1.53	28.76	14.69	1.96
Occlusion × Block Context × Production Length	171.99	39.20	**4.39**	72.28	15.45	**4.68**
Occlusion × Confederate Length × Production Length	27.58	36.77	0.75	10.83	14.45	0.75
Block Context × Confederate Length × Production Length	14.33	35.94	0.40	6.68	14.11	0.47
Occlusion × Block Context × Confederate Length × Production Length	68.70	74.66	0.92	25.69	29.38	0.80
Random effects	Term		*SD*	Term		*SD*
	Participant	Intercept	187.07	Participant	Intercept	135.04
	Item	Intercept	48.40	Item	Intercept	23.02
	Residual		309.32	Residual		124.23

Note. Bold values reflect *t*-values above 2.

lead advantage observed overall was weakest in mixed blocks when the speaker produced a monosyllabic utterance. For the turn gap model, there were also three additional interactions. Interactions between production length and confederate word number, and production length and confederate length showed, respectively, that the overall effects of each factor were largest for disyllabic productions. A three-way interaction between block context, confederate length, and production length showed that the shortest turn gaps were in mixed blocks following disyllabic items when the production was disyllabic. These appear in Table A3.

(*Appendices continue*)

Table A3

Outputs of Linear Mixed-Effect Models for Eye-Voice Lead and Turn Gap in Experiment 3 Containing Production Length

Fixed effects	Eye-voice lead			Turn gap		
	Estimate	*SE*	*t* value	Estimate	*SE*	*t* value
Intercept	962.83	45.01	**21.39**	218.36	18.91	**11.55**
Confederate word number	−442.30	33.86	**−13.07**	27.77	3.36	**8.27**
Block context	12.93	32.21	0.40	13.93	14.02	0.99
Confederate length	−60.22	11.49	**−5.24**	49.44	3.36	**14.72**
Production length	3.86	21.05	0.18	10.73	12.39	0.87
Confederate Word Number × Block Context	−42.53	23.08	−1.84	9.43	6.72	1.40
Confederate Word Number × Confederate Length	95.20	22.97	**4.14**	3.28	6.72	0.49
Block Context × Confederate Length	−15.01	22.97	−0.65	−8.25	6.72	−1.23
Confederate Word Number × Production Length	16.37	22.97	0.71	−19.02	6.72	**−2.83**
Block Context × Production Length	−0.51	22.98	−0.02	−11.28	28.03	−0.40
Confederate Length × Production Length	8.23	22.96	0.36	−14.09	6.72	**−2.10**
Confederate Word Number × Block Context × Confederate Length	−87.89	45.94	−1.91	−39.18	13.44	**−2.92**
Confederate Word Number × Block Context × Production Length	−107.37	45.94	**−2.34**	4.70	13.44	0.35
Confederate Word Number × Confederate Length × Production Length	−65.83	45.94	−1.43	3.45	13.44	0.26
Block Context × Confederate Length × Production Length	−103.23	45.93	**−2.25**	34.14	13.44	**2.54**
Confederate Word Number × Block Context × Confederate Length × Production Length	−2.02	91.89	−0.02	−20.31	26.88	−0.76

Random effects	Term		*SD*	Term		*SD*
	Participant	Intercept	276.65	Participant	Intercept	113.03
		Confederate length	200.78	Item	Intercept	23.84
		Block context	189.67		Block context	54.43
	Item	Intercept	35.28	Residual		118.91
	Residual		389.84			

Note. Bold values reflect *t*-values above 2.

(*Appendices continue*)

Appendix B

Pictures Used for Experiment 2 and 3

Confederate items

Hemd

Herder

Mand

Masker

Naald

Nagel

Pan

Panda

Rits

Ridder

Schaar

Schaduw

Vlieg

Vliegtuig

Zee

Zebra

Gabor Patch

Participant items

Been

Baby

Doos

Doolhof

Duif

Duivel

Kast

Cactus

Kip

Kikker

Lamp

Ladder

Tang

Tandarts

Wolf

Wortel

Received February 11, 2020
Revision received September 10, 2020
Accepted October 21, 2020 ■