CHAPTER TWO

What do we know about the mechanisms of response planning in dialog?

Ruth E. Corps*
Psychology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
*Corresponding author: e-mail address: ruth.corps@mpi.nl

Contents

Abstract

During dialog, interlocutors take turns at speaking with little gap or overlap between their contributions. But language production in monolog is comparatively slow. Theories of dialog tend to agree that interlocutors manage these timing demands by planning a response early, before the current speaker reaches the end of their turn. In the first half of this chapter, I review experimental research supporting these theories. But this research also suggests that planning a response early, while simultaneously comprehending, is difficult. Does response planning need to be this difficult during dialog? In other words, is early-planning always necessary? In the second half of this chapter, I discuss research that suggests the answer to this question is no. In particular, corpora of natural conversation demonstrate that speakers do not directly respond to the immediately preceding utterance of their partner—instead, they continue an

Psychology of Learning and Motivation, Volume 78
ISSN 0079-7421
https://doi.org/10.1016/bs.plm.2023.02.002

utterance they produced earlier. This parallel talk likely occurs because speakers are highly incremental and plan only part of their utterance before speaking, leading to pauses, hesitations, and disfluencies. As a result, speakers do not need to engage in extensive advance planning. Thus, laboratory studies do not provide a full picture of language production in dialog, and further research using naturalistic tasks is needed.

1. Introduction

Psycholinguists have developed detailed accounts of the cognitive processes underlying speaking (language production) and listening (language comprehension), and they have traditionally studied these mechanisms separately. In fact, we have sophisticated theories of language production during monolog (i.e., when we speak by ourselves; Section 2). However, the majority of language use occurs in dialog, in which we rarely just speak or listen. Instead, we usually take turns at talking, regularly switching between comprehending our partner and producing our own response. But what do we actually know about the mechanisms of language production, and particularly response planning, in dialog?

Dialog has been of interest to corpus linguists for decades, and particularly since Sacks, Schegloff, and Jefferson's (1978) seminal work on the rules governing interaction. They noted that dialog involves at least two people who take alternating turns at speaking. The content of these turns (i.e., what the speaker wants to say) is not specified in advance, and so speakers have to plan their utterances "on the fly." Importantly, only one speaker tends to talk at a time, and transitions (from one speaker to the next) with no gap or overlap are common but any overlap that does occur is very brief. Thus, turns are coordinated in time.

More recently, corpus studies have confirmed the close timing of turns. For example, Stivers et al. (2009) quantified response times to polar (*yes/no*) questions in ten languages. They found that there was variation in the average gap duration across languages, with some having short average gaps (such as Japanese, with an average gap of 7 milliseconds [ms]) and others having longer average gaps (such as Danish, with an average gap of 469 ms). But despite this variation, most answers (i.e., the peak of the distribution, or the mode) were produced within 200 ms of the question end across all languages. Furthermore, Heldner and Edlund (2010) analyzed gap durations in three different corpora—a Dutch dialog corpus, which consisted of face-to-face and telephone conversations, and English and Swedish Map Task

corpora, where speakers worked together to make their way around a map. Although there was a large amount of overlap in all three corpora (40%), the majority of all turn transitions (51–55%) took place within 200 ms, with 70–82% taking place within 500 ms.

The short gaps between turns in conversation contrasts with the much longer latencies in isolated language production. Research has shown that producing a picture name takes between 600 and 1200 ms, depending on factors such as word frequency (e.g., Indefrey & Levelt, 2004), while a complete utterance takes around 1500 ms (e.g., Ferreira, 1991). Thus, if the listener (as the next-speaker) is to achieve a turn gap of 200 ms, then they must begin planning their response before the current speaker reaches the end of their utterance. As a result, the listener must plan while still comprehending the speaker, and comprehension and production processes must overlap (at least momentarily). In fact, there is evidence that responding too slowly is interpreted negatively by the other person in the dialog, and so timely responses are socially desirable (Section 3.1).

This *early-planning* mechanism has been implemented in theories of conversation, which claim that listeners predict what a speaker is likely to say (utterance content) and use these predictions to begin planning a response as soon as possible, even if they are still comprehending the speaker's utterance (e.g., Levinson & Torreira, 2015; see Section 3.2). There is much evidence to support early-planning in dialog (see Sections 3.3 and 3.4). This evidence comes from highly-constrained laboratory tasks that use a variety of techniques, such as question-answering or picture naming, designed to approximate the processes involved in conversation. Much like dialog, these tasks involve planning a response and articulating it at the appropriate moment, so there is little gap or overlap between responses.

This research suggests that early-planning enables interlocutors to closely coordinate their utterances. But this research also suggests that planning a response while simultaneously comprehending is difficult (see Section 3.5). Does language production need to be this difficult during dialog? In other words, does there always need to be this large overlap between comprehension and production processes? Our recent research (Corps, Knudsen, & Meyer, 2022) suggests the answer to this question is no. In a recent corpus analysis, we have shown that speakers do not always respond to each other during dialog (see Section 4.1)—instead, they continue an utterance they produced earlier. In these cases, the listener's response does not depend on the content of the speaker's utterance, and so comprehension and production processes do not always need to extensively overlap.

In addition, studies of monolog suggest that language production is highly incremental, and speakers do not need to plan their full utterance before they actually speak. Theories and studies of dialog have tended to ignore this incrementality. For example, Levinson and Torreira (2015) claim that speakers complete all stages of response planning as early as possible. But incrementality likely makes language production easier than it would be if speakers planned a full sentence before speaking—the cognitive effort of planning is distributed throughout the utterance, rather than concentrated at the start. One consequence of this incrementality is that speakers are often disfluent, producing filled pauses such as *uh* or *um* (see Section 2.2.1). Because these theories of dialog have paid little attention to this incrementality and disfluency, our understanding of language production is incomplete—laboratory studies have focused on idealized situations, in which one speaker plans a full utterance in response to the previous speaker. In Section 5, I report a corpus analysis that investigates the similarity of speech elicited in the laboratory and speech elicited in natural conversations, with the aim of demonstrating that we need to step away from basing theories of language production in dialog on the idealized utterances produced in highly constrained laboratory tasks. In the following sections, I first provide an overview of what we know about language production in monolog before turning to dialog. Note that I limit my discussion to the mechanisms of production. As a result, I do not discuss the extensive literature on priming in dialog, which focuses on what causes speakers to produce one word or syntactic structure over another (see e.g., Garrod & Pickering, 2007, for a review).

2. The mechanisms of language production in monolog

2.1 Producing words

Although producing a word may seem simple, it is no easy feat. Researchers tend to agree that language production is a staged process, typically divided into three steps—deciding what to say (conceptualization), deciding how to say it (formulation), and then finally saying it (articulation; Levelt, 1989). During conceptualization, the speaker decides which message they wish to convey. For example, if the speaker is asked to name a picture of a dog, then they may activate the lexical concept *golden retriever* or *dog*, depending on the context of production. If the speaker names the picture in the context of other dogs, then they will produce *golden retriever*. But if they name in the context of other animals, then they will likely produce

dog (see Clark, 1997), unless the term *golden retriever* has been used recently (e.g., Brennan & Clark, 1996).

The concept is then formulated, and this process of formulation (or lexicalization) involves two steps. First, activation spreads from the concept to connected abstract lexical representations. For the sake of simplicity, I adopt the terminology of Levelt, Roelofs, and Meyer (1999); see also Kempen & Huijbers (1983) and refer to these representations as *lemmas*, but they have also been referred to as *lexical entries*, *lexical representations*, or simply *words* throughout the literature. Theories tend to differ with respect to how lemmas are characterized. Some researchers claim that lemmas are lexical representations specifying the meaning of a word (or its semantics; e.g., Butterworth, 1989), while others claim that lemmas represent the syntactic features of a word, such as its grammatical class (e.g., whether it is a noun or a verb) or gender (e.g., whether it is gendered or gender-neutral; Levelt et al., 1999). But regardless, this lemma is the interface between the conceptual level and the next stage of formulation—word-form retrieval.

During word-form retrieval (or phonetic encoding), the activated lemma is mapped onto its corresponding word-form, which provides the speaker with information about the word's sound and how it should be produced. Constructing this word-form involves retrieving the word's morphological makeup, its metrical shape, and its segmental makeup (phonological encoding). For example, if the speaker is producing the word *dog*, then they will retrieve the morpheme <dog>. They will then spell out the metrical shape of *dog* (that it is monosyllabic) and its segmental information (/d/ /ɒ/ /g/). These representations spread activation to connected phonemes, which specify the word's syllabary and the articulatory gestures for producing the word (such as the necessary mouth movements). Once this process is complete, the speaker finally articulates the word.

It is worth noting that I have painted a rather simplistic view of word production. Although researchers agree that production involves selecting a word's meaning and its form, this is where the agreement tends to end. Some theories claim that production is strictly serial, so that speakers only activate the word-form of a single lemma (e.g., Levelt et al., 1999). For example, if the speaker wishes to produce the word *dog*, then activation will spread to semantically related lemmas, such as *cat* and *bone*, but only the word-form for the selected lemma (*dog*) is actually activated (e.g., Levelt et al., 1991). Others, however, claim that activation flows freely among meaning, lexical, and sound representations, and so speakers activate the word form of partially activated but unselected lemmas (e.g., Dell, 1986).

For example, the speaker would activate the word-form of *dog*, *cat*, and *bone*, even though they selected the lemma for *dog* (e.g., Peterson & Savoy, 1998).

Additionally, most theories accept the existence of an intermediary stage between conceptualization and word-form access, but others have rejected the existence of lemmas completely. For example, Caramazza (1997); Caramazza & Miozzo (1998); Miozzo & Caramazza (1997) suggested that some characteristics (e.g., verb tense or grammatical category), which are often thought to be activated at the lemma level, can be directly activated from a word's concept, while others (e.g., gender features) can be activated from word form. Thus, Caramazza claims that lemmas are not necessary for production.

But regardless of these disagreements, there is clear evidence for separate meaning and word-form representations in word production. In the classic picture-word interference (PWI) paradigm, participants name pictures while ignoring auditory or written distractor words (e.g., Schriefers, Meyer, & Levelt, 1990). Participants are slower to name a picture (e.g., dog) when the distractor word is semantically related (e.g., *cat*) rather than unrelated. They are also faster to name a picture when the distractor word is phonologically related (e.g., *doll*). Importantly, these effects depend on the time interval between the presentation of the distractor word and the presentation of the picture (the stimulus onset asynchrony, or SOA). In particular, a semantically related distractor word interferes with picture naming when presented 150 ms before the picture (an SOA of −150 ms), while a phonologically related distractor facilitates picture naming when presented at the same time as the picture or 150 ms after (an SOA of 0 or +150 ms). These results suggest that lexical access is staged, with meaning accessed separately from form.

Tip-of-the-tongue (TOT) states also support the separation of meaning and form representations. A TOT state occurs when the speaker cannot recall a particular word (even though they know it), but can recall information about the word. For example, speakers can report information about the word's form, such as its length in syllables or its word onset (e.g., Brown & McNeill, 1966). They can also recall syntactic information, such as the word's grammatical gender (e.g., Vigliocco, Antonini, & Garrett, 1997), its grammatical class (e.g., Iwasaki, Vigliocco, & Garrett, 1998), and whether it is a count or mass noun (e.g., Vigliocco, Vinson, Martin, & Garrett, 1999). These findings suggest that speakers are able to correctly report syntactic and semantic information about the word, even though they cannot retrieve the word's full form for articulation, suggesting that form information is accessed separately from meaning. Thus, we know that speakers produce a word by selecting its meaning separately from its form.

2.2 Incrementality in sentence production

Words are often produced as part of larger sentences, and so speakers have to activate multiple words and order them in an appropriate structure. How many words can speakers activate in parallel? In other words, how far ahead do they plan? Existing theories of sentence production generally assume that speakers do not plan an entire sentence before they begin to speak (e.g., Ferreira & Slevc, 2007). Instead, planning proceeds incrementally—as soon as one piece of the sentence (such as the first word) is processed at one stage of production, it is passed onto the next stage. As a result, the complete sentence does not need to be planned at the conceptual level before it is formulated—later parts of the sentence can be planned while earlier parts are simultaneously formulated. For example, a speaker who wishes to say *Dogs chase cats* could activate the concept for the word *dog*, which triggers retrieval of its corresponding lemma and word-form. This word then takes the first spot in the sentence's syntactic frame, and the speaker can articulate *dog* without necessarily knowing how the sentence will end. Two sources of evidence support incrementality in sentence production—research that has shown that speech is often disfluent (Section 2.2.1), and research manipulating the ease of sentence planning (Section 2.2.2).

2.2.1 Incrementality and disfluencies

Most of the research investigating disfluencies in language production comes from dialog, but these studies are relevant for understanding incrementality during monolog and so I discuss these results here. We know that speakers do not plan their full sentence before they speak because they often produce disfluencies. For example, consider excerpt (1) below from the Santa Barbara corpus of American English (Du Bois et al., 2000), where Lynne is talking about shoeing a horse.

(1) Lynne: But uh what was I gonna say. Oh and it's really tiring though. And it—you know like, you get so—I've only done like, well, at the end of the year, now see I took the second half of the course.

It is clear from (1) that speech can be disfluent in many different ways: Utterances can be incomplete, contain pauses (which may be silent or filled with words like *uh* or *um*), hesitations, repetitions, discourse markers (such as *like*), and utterance restarts (e.g., Fox Tree & Clark, 1997). Many of these disfluencies are present in Lynne's utterance—for example, she produces

filled pauses such as *uh*, and her utterance includes incomplete units (e.g., *you get so*) that are abandoned and never resumed. These different types of disfluencies occur at different rates. For example, Eklund and Shriberg (1998) found that 32% of sentences and 5% of words were disfluent in a corpus of American-English telephone conversations, with filled pauses occurring more often (59% of the time) than any other type of disfluency. Similar results were found by Bortfeld, Leon, Bloom, Schober, and Brennan (2001) in a corpus of task-oriented conversations. Furthermore, Branigan, Lickley, and McKelvie (1999) found that 31% of disfluencies in the English Map Task corpus were repetitions, 42% were deletions, 10% were hesitations, and 13% were substitutions.

Much debate has focused on the meaning of these disfluencies in language production (e.g., Fox Tree, 2010; Fox Tree & Schrock, 1999; Fraser, 1999). This debate is primarily of interest to researchers investigating speaking in dialog, since they are largely based on corpora of conversational speech, but I briefly summarize their results here since they are useful for understanding incrementality during sentence planning in monolog. According to the *signal account*, speakers produce disfluencies to signal upcoming difficulty or delay to the listener (e.g., Clark, 1994; Fox Tree & Clark, 1997; Smith & Clark, 1993), perhaps so they can hold the floor or encourage the listener to allocate their attention to forthcoming information. For example, Smith and Clark (1993) found turn gaps of 2230 ms when utterances began without a filler, gaps of 2560 ms when utterances began with *uh*, and a gap of 8830 ms when utterances began with *um*. Additionally, research suggests that comprehenders expect speakers to refer to objects that have not been mentioned before (discourse-new objects) when the speaker produces a disfluency (e.g., *Now put thee uh…*), but objects they have referred to before (discourse-old objects) when they do not produce a disfluency (e.g., *Now put the…*; e.g., Arnold, Fagnano, & Tanenhaus, 2003; Arnold & Tanenhaus, 2011; Arnold, Tanenhaus, Altmann, & Fagnano, 2004). Furthermore, disfluencies can trigger attention to upcoming words (e.g., Bosker, Tjiong, Quené, Sanders, & De Jong, 2015; Collard, Corley, MacGregor, & Donaldson, 2008), making them easier to remember later (e.g., Corley, MacGregor, & Donaldson, 2007). Thus, speakers may intentionally produce disfluencies to signal an upcoming delay or important information to the listener.

Alternatively, these disfluencies could be a *symptom* of difficulty planning to speak (e.g., Levelt, 1989). Although speakers may produce disfluencies to signal discourse-new objects, they may also produce disfluencies before referring to these objects simply because they find them harder to name

than discourse-old objects. Consistent with this argument, Schachter, Christenfeld, Ravina, and Bilous (1991) found that speakers hesitated more when they had choice in what they could say, which presumably made planning difficult. Similarly, Hartsuiker and Notebaert (2009) found that participants produced more disfluencies, pauses, and self-corrections when naming pictures with low name agreement (such as *sofa*, which could also be referred to as *couch* or *settee*) than pictures with high name agreement (such as *arm*). They also found that participants made more self-corrections and repetitions when they named gender neuter pictures (which use the infrequent determiner *het* in Dutch) than when they named common gender pictures (which use the more frequent determiner *de*), suggesting speakers produce disfluencies when they experience difficulty during lexical access. However, note that these results could be interpreted in line with the signal account. For example, speakers could be aware that they will have difficulty naming pictures with low name agreement rather than high name agreement and could produce a disfluency to signal this difficulty to the listener. As a result, it is difficult to determine whether speakers produce disfluencies as a signal or a symptom of difficulty during speaking.

Regardless of the meaning of these disfluencies, they demonstrate that speakers do not plan their full sentence before they produce it. If they did, then we would expect disfluencies to rarely occur, unless they were a deliberate signal, and we would expect any that do occur to primarily be located at the beginning of the speaker's utterance, where most of the planning difficulty occurs. But inconsistent with this prediction, Clark and Fox Tree (2002) found that disfluencies are distributed throughout the speaker's utterance. They defined three locations in utterances from the London-Lund corpus of British English, which consists of face-to-face conversations. Corpus analyses often focus on intonation units, which are stretches of speech produced under a continuous intonation contour (e.g., Chafe, 1992). These intonation units can consist of sentences, phrases, parts of phrases, or even single words. Clark and Fox Tree defined three locations in intonation units: (1) at the boundary; (2) after the first word; and (3) later in the utterance. An example of these locations can be found in (2), where commas are used to mark intonation unit boundaries.

(2) and then uh somebody said, . [1] but um—[2] don't you think there's evidence of this, in the twelfth—[3] and thirteenth centuries?

If speakers plan their full sentence before speaking, then disfluencies should primarily occur at location 1, where the speaker begins a new intonation

unit, and should not occur at location 2, where the speaker is part way through an intonation unit, and especially not at 3, where they have almost finished the intonation unit. Although speakers produced more *uhs* and *ums* at location 1 (43 per 1000 opportunities), they still occurred at location 2 (27 per 1000 opportunities) and location 3 (13 per 1000 opportunities), suggesting disfluencies are not confined to the start of the speaker's sentence.

Although I have focused on research investigating whether disfluencies are produced as a consequence of difficulties during language production, there is also evidence that disfluencies can serve pragmatic functions, such as signaling new information (e.g., Arnold et al., 2003; Arnold & Tanenhaus, 2011) or discourse structure (e.g., Swerts, 1998). Importantly, however, the occurrence of disfluencies demonstrates that speakers do not plan their full sentences before speaking. Instead, they plan incrementally, and so do not necessarily know how their sentence will end before they start speaking.

2.2.2 Experimental studies on advance planning and incrementality

Although studies investigating the occurrence of disfluencies during production provide evidence that speakers plan their sentences incrementally, these studies were not designed to explicitly test this claim. Experimental studies of sentence planning have investigated the scope of advance planning (i.e., how much of their sentence the speaker plans before speech onset), and provide more direct support for incrementality during sentence production (e.g., Brown-Schmidt & Konopka, 2015; Brown-Schmidt & Tanenhaus, 2006; Griffin & Bock, 2000; Griffin, 2001; Smith & Wheeldon, 1999; see Wheeldon, 2013, for a review). For example, Griffin (2001) conducted an eye-tracking experiment in which participants described objects displayed on-screen using the sentence frame *The A and the B are above the C*. Objects B and C varied in their name agreement—sometimes they had high agreement and only one plausible name (e.g., *apple*), while other times they had low agreement and multiple plausible names (e.g., *sofa* or *couch*). These objects also varied in the frequency of their dominant name—sometimes the dominant name was highly frequent, while other times it was less frequent. Participants spent longer looking at objects B and C when they had low rather than high agreement names. Participants also spent longer looking at these objects if their names were low rather than high frequency. But the agreement and frequency of these objects did not affect how quickly participants named object A, suggesting participants

began speaking when they had planned object A's name, before they selected object B and C's names. In other words, speakers did not plan their full sentence before they began articulation.

Other eye-tracking studies also suggest that speakers tend to look at each object as they mention it, only shifting their gaze to the next object prior to articulation of the previous object (e.g., Gleitman, January, Nappa, & Trueswell, 2007; Griffin & Bock, 2000; Griffin & Spieler, 2006; Meyer, Sleiderink, & Levelt, 1998). If participants plan more than one word at a time, then the delay between fixating an object and naming it should be shorter for words occurring later in the sentence. But Griffin and Bock found that this delay was the same for all objects, regardless of their position in the sentence. Furthermore, Meyer et al. had participants name pairs of objects using noun phrase conjunctions, such as *scooter and hair*. Participants shifted their gaze from the current object to the next object only once they had retrieved the word-form of the object they were naming. In other words, they only fixated *hair* once they had retrieved the word-form of *scooter*. Together, these findings suggest that speakers plan only one word before beginning articulation.

However, other research using less predictable sentences suggests speakers can activate more than one word at a time. For example, Smith and Wheeldon (1999) had participants produce sentences about moving objects. Participants were slower to produce sentences beginning with complex noun phrases (e.g., *The dog and the kite move above the house*) rather than a simple noun phrase (e.g., *The dog moves above the kite and the house*), suggesting participants dedicated more resources to planning a later word (*kite*) before the onset of the first word (*dog*) when sentences were more complex. Furthermore, Meyer (1996); see also Wagner, Jescheniak, & Schriefers (2010) had participants name pairs of pictures with either noun phrase conjunctions (e.g., *the arrow and the bag*) or locative sentences (e.g., *the arrow is next to the bag*). While planning their utterance, speakers heard a distractor word, which was semantically related, phonologically related, or unrelated to the first or the second noun. Participants were slower to initiate their sentences when the distractor word was semantically related rather than unrelated to either of the nouns, suggesting that the speaker planned the meaning of both nouns. Participants were also faster to initiate their sentences when the distractor word was phonologically related rather than unrelated to the first but not the second noun, suggesting that participants planned the word-form of only the first noun. These findings suggest that the scope of planning is different for different stages of production.

Thus, although theories of sentence production tend to agree that speakers plan their utterances incrementally, there is disagreement about the scope of this incrementality. Some studies suggest that speakers plan word-by-word (e.g., Griffin, 2001), while others suggest they plan in larger chunks (e.g., Smith & Wheeldon, 1999). One way of reconciling these findings is by assuming that the scope of planning is different for meaning and word-form (e.g., Meyer, 1996). Relatedly, planning is likely flexible, and so the degree of incrementality is under the speaker's control (e.g., Konopka, 2012; Swets, Jacovina, & Gerrig, 2013).

Consistent with this suggestion, there is evidence that the scope of planning is influenced by time pressure. Ferreira and Swets (2002); see also Swets et al. (2013) had participants produce answers to two digit sums (e.g., 9 + 7 = ?) when time pressure was absent (Experiment 1) or present (Experiment 2). In both experiments, initiation times increased as problem difficulty increased. However, problem difficulty influenced utterance duration only in Experiment 2, suggesting that speakers simultaneously planned and articulated when they were encouraged to produce their utterance immediately. When there was no time pressure, participants made use of more extensive advance planning. Similarly, Wagner et al. (2010; Experiment 1) measured planning scope using a PWI task, in which participants produced simple sentences consisting of two nouns (e.g., *the frog is next to the mug*). While producing these sentences, participants heard distractors that were unrelated or semantically related to the first or the second noun. The authors determined whether each participant was a fast or a slow speaker based on their average latencies when they produced the sentence in the presence of an unrelated distractor. Both fast and slow speakers experienced an interference effect for the first noun—they were slower to initiate their sentences when the distractor word was semantically related rather than unrelated. But the interference effect on the second noun was larger for the slow than the fast speakers, suggesting slow speakers had a tendency to plan further in advance than fast speakers.

Planning scope is also sensitive to linguistic factors, such as ease of structural assembly. In their second PWI experiment, Wagner et al. (2010) asked participants to produce simple sentences (e.g., *the frog is next to the mug*) or to switch between producing simple and complex sentences (e.g., *the red frog is next to the red mug*). They found that the additional cognitive load of switching sentence structure eliminated any interference effect for the second noun, regardless of whether speakers were fast or slow. Similarly, Konopka (2012); see also Konopka & Meyer (2014) had participants describe three pictures using a complex noun phrase (e.g., *The axe and the saw are above/below the cup*). On some trials, targets were preceded by primes

that elicited the same or a different sentence structure. The results showed that repeating sentence structure extended speakers' planning scope from one to two nouns. Together, these findings suggest speakers reduce their planning scope when structural assembly is difficult. But regardless of how much speakers plan in advance, these studies demonstrate that speakers plan incrementally—they do not need to plan a full sentence before they speak.

3. The mechanisms of language production in dialog

It is clear from Section 2 that we have sophisticated theories of speech production during monolog. But the majority of language use occurs in dialog, where we rarely just speak. Instead, we usually take turns at talking, regularly switching between comprehending our partner and producing our own response. What do we know about the mechanisms of language production, and particularly response planning, in dialog?

3.1 Why is timely language production so important in dialog?

Before discussing the mechanisms of language production, and timely turn-taking, in dialog, it is worth understanding why it is important that the listener responds to the speaker so quickly. Dialog seems difficult—most theories agree that the next-speaker has to juggle comprehension and production processes if they are to achieve turn gaps of 200 ms (e.g., Levinson & Torreira, 2015). The next-speaker could avoid this issue by beginning response planning only once the speaker has reached the end of their turn. So why are short gaps so important? Research suggests they are important for maintaining the flow of conversation, and there is evidence that delayed responses tend to be interpreted negatively by the listener. For example, if you invite someone for dinner then a delayed response may indicate the other person's reluctance. This issue is illustrated in an excerpt from a telephone conversation (3), in which C interprets a pause of 1.86 s as a negative response to his question (Levinson, 1995):

(3)

C: So um I was wondering if you would be in your office on Monday by any chance?

(1.86 s)

C: Probably not.

Experimental studies have investigated the consequences of these delayed responses. For example, Bögels, Kendrick and Levinson (2015), (see also Bögels, Kendrick, & Levinson, 2020) measured Dutch participants'

brain activity while they listened to telephone conversations, in which one speaker produced an initiating action (e.g., a request, an offer, or a proposal) and the other speaker produced either an acceptance (*ja* or *yes*) or a rejection (*nee* or *no*). The gap between these two turns was either long (1000 ms) or short (300 ms). Participants displayed a larger N400, which is associated with semantic processing (see Kutas & Federmeier, 2011, for a review), when they encountered a rejection following a short rather than a long gap. This effect suggests that the listener expects an immediate response to be positive, and so they experience processing difficulty when this response is actually negative. Thus, long gaps can indicate that the speaker will produce a rejection, which the listener may interpret negatively.

Research also suggests that gap length affects how listeners view their partner. In one study, Templeton, Chang, Reynolds, Cone LeBeaumont, and Wheatley (2022) investigated whether response times (which are equivalent to turn gaps) provide a useful measure of social connection. Participants held a ten-minute casual conversation with a stranger (Experiment 1) or a friend (Experiment 2) and then rated their social connection. In both experiments, participants felt more connected to their partner and enjoyed the conversation more when their partner responded more quickly. In Experiment 3, participants listened to audio clips in which the gap between the turns was manipulated so it was either short or long. As in the previous experiments, participants thought the interlocutors were more socially connected when they responded more quickly to each other, suggesting overhearers perceive short gaps positively, even if they are not involved in the conversation.

In another study, Koudenburg, Postmes, and Gordijn (2013) had participants interact with each other naturally or with a one second delay between turns in the second half of the conversation. Participants who had a conversation with a delay of one second felt less solidarity with their partner than those who conversed naturally. Furthermore, Roberts and Francis (2013); Roberts, Francis, and Morgan (2006); Roberts, Margutti and Takano (2011) found that listeners' ratings of the speaker's willingness to comply decreased as the length of the gap between turns increased. Together, these findings suggest that long gaps do not only disrupt the flow of conversation—they are also socially undesirable. In the next sections, I discuss the mechanisms that enable interlocutors to avoid long gaps.

3.2 Levinson and Torreira's (2015) theory of language production in dialog

Although other psycholinguistic models of turn-taking exist (e.g., Garrod & Pickering, 2015), I focus my discussion on Levinson and Torreira's (2015)

theory because it the most influential model in the literature (see e.g., Bögels & Levinson, 2017; Corps, Gambi, & Pickering, 2018, for a review). They proposed that the production system (supporting speaking) and the comprehension system (supporting listening) are simultaneously engaged in conversation. In particular, the listener (B) focuses on determining the gist of the current speaker's (A) utterance. B can determine the gist by identifying A's speech act (i.e., what type of utterance they are producing, such as a question; e.g., Gisladottir, Chwilla, & Levinson, 2015), or by using the context of A's utterance to predict what she is likely to say (e.g., Altmann & Kamide, 1999).

As soon as B has identified A's speech act or has predicted enough of A's utterance, B begins planning a response. Thus, the content of B's response and the moment he begins planning it both depend heavily on the content of A's utterance. While planning this response, B simultaneously listens to the rest of A's utterance and waits for cues that signal she will soon finish speaking. If B finishes planning before A has reached the end of her utterance, he holds his response in an articulatory buffer (presumably at the phonological level) until he can articulate. Once there is sufficient evidence that the end of the utterance is imminent, B launches his planned response.

This model explains short turn gaps by claiming that next-speakers are highly proactive and begin planning their utterances as soon as the response-relevant information has been provided. Under this theory, turns are coordinated in both content and time because (1) the content of B's utterance depends on the content of A's utterance; and (2) B only initiates articulation once A has finished. In the next section, I review evidence that listeners (as next-speakers) can determine the gist of the speaker's utterance by predicting the content of this utterance. I then discuss evidence that suggests speakers use these predictions to plan a response early, in line with Levinson and Torreira's theory.

3.3 Content prediction during language comprehension

This section provides only a brief review of evidence for prediction during comprehension, since more extensive reviews are readily available elsewhere (e.g., Pickering & Gambi, 2018). The important point is that much research has demonstrated that listeners predict what a speaker is likely to say—that is, the content of the speaker's utterance. For example, participants often expect the same continuation (e.g., *spoon*) when presented with sentence contexts such as *At the dinner party, I wondered why my mother wasn't eating her soup. Then I noticed she didn't have a*…. Importantly, this effect does

not only occur in laboratory tasks; in natural conversations, interlocutors sometimes complete each other's utterances (e.g., Howes, Purver, Healey, Mills, & Gregoromichelaki, 2011), suggesting that the listener comprehends the speaker's incoming utterance and predicts what the speaker is likely to say next.

Some research exploring prediction during language comprehension has used the *visual-world paradigm*, in which participants view a visual scene (usually consisting of many objects) while simultaneously listening to sentences. Predictive looking is thought to occur when listeners attend to an object before it is actually mentioned. In one of the first studies using this method, Altmann and Kamide (1999), see also Kamide, Altmann, and Haywood (2003) recorded participants' eye movements while they viewed visual scenes (e.g., a picture of a boy, a cake, a toy car, a toy train set, and a ball) and simultaneously listened to sentences. In one condition, these sentences (e.g., *The boy will eat…*) could apply to only one object in the scene (e.g., the cake), thus making the mention of the cake predictable. In the other condition, the sentences could apply to any of the objects (e.g., *The boy will move…*), making it impossible for the listener to predict how the sentence would continue. When participants heard the verb *eat*, they fixated the cake earlier and for longer than when they heard the verb *move*, suggesting they used the semantics of the verb to predict which of the objects was most likely to be mentioned next.

There is also evidence that listeners predict syntax. In an electroencephalogram (EEG) experiment, Ito, Gambi, Pickering, Fullenbach, and Husband (2020) presented Italian participants with sentences (e.g., *The traffic on the motorway came to a standstill because… [Il traffic in autostrada è rimasto bloccato a causa di…*]) that predicted a particular article and noun combination of a particular syntactic gender (e.g., *an incident [un*$_{\text{masculine}}$ *incidente*$_{\text{masculine}}$]). These sentences continued with the expected article and noun combination, or they continued with an article and noun combination that mismatched the syntactic gender of the expected continuation (e.g., *a flooding [un'*$_{\text{feminine}}$ *inodazione*$_{\text{feminine}}$]). Participants showed a greater negativity around 250 ms after the article when they encountered the unexpected article + noun combination compared to when they encountered the expected article + noun. These findings suggest that listeners can predict the syntactic gender of upcoming words (see also Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Bates, Moreno, & Kutas, 2003).

Finally, there is some evidence that listeners predict word-form (but see DeLong, Urbach, & Kutas, 2017; Ito, Martin, & Nieuwland, 2017; Nieuwland et al., 2018; Urbach, DeLong, Chan, & Kutas, 2020, for

interesting discussions of this evidence). In addition to manipulating syntactic gender, Ito et al. (2020) included a condition where the sentences continued with an article matching the gender of the expected article and noun, but mismatching the word-form (e.g., *a collision* [*uno*$_{\text{masculine}}$ *scontro*$_{\text{masculine}}$]). Participants showed a greater negativity around 450 ms after the article when they encountered the form mismatch article compared to when they encountered the expected article, suggesting they predicted the form of upcoming words.

It is worth noting that word-form predictions occurred later (around 450 ms) than syntactic predictions (around 250 ms) in Ito et al.'s (2020) study, consistent with theories that word-form predictions are delayed relative to semantic and syntactic predictions (e.g., Pickering & Gambi, 2018). However, other studies have found that word-form predictions show a similar time-course of activation to semantic predictions. DeLong, Chan, and Kutas (2018) recorded ERPs while participants read highly constraining sentence contexts (e.g., *The woman stashed her wallet in her* ***purse*** *for safety*) which were continued with a highly predictable word (*purse* in this example), an unpredictable word semantically related to the predictable word (*snatcher* rather than *purse*), or an unpredictable word orthographically related to the predictable word (*nurse* rather than *purse*). They found that both semantically related and orthographically related unpredictable words elicited similarly reduced N400s, suggesting word-form predictions show a similar time-course to semantic predictions (but see also Ito, Corley, Pickering, Martin, & Nieuwland, 2016).

In sum, there is evidence that listeners predict what a speaker is likely to say. Once the listener makes this prediction, they can begin the process of planning their response. For example, if the speaker says *Is the boy going to fly his…*, then the listener could predict the meaning of the word *kite* and use this prediction to plan their answer. The next section reviews evidence that supports such early-planning.

3.4 Evidence for early response planning

After having heard or predicted a sufficient part of the speaker's utterance, listeners can begin planning their own response. Studies investigating the time-course of response planning have used a variety of methods, including picture naming and question-answering, which are designed to be highly controlled while still approximating the mechanisms involved in conversation. In particular, participants' responses are generally highly constrained,

but they still have to prepare this response and articulate it so they avoid extensive gap or overlap with the previous speaker. Many of these studies support Levinson and Torreira (2015) claim that listeners are highly pro-active and begin planning their response early while still comprehending.

In one of the first studies, Bögels, Magyari, and Levinson (2015) measured EEG correlates during a question-answering task, in which the information (here *007*) needed for response planning was available either early (e.g., *Which character, also called 007, appears in the famous movies*?) or late (e.g., *Which character from the famous movies is also called 007*?). Participants were quicker to answer when the critical information was available early (mean (M) = 640 ms) rather than late (M = 950 ms). EEG correlates showed a larger positivity to the critical word when participants planned a response rather than when they simply listened to the questions. This effect was localized to the middle frontal and precentral gyri, which overlap with brain areas involved in speech production (Indefrey & Levelt, 2004). This effect occurred around 500 ms after the onset of the critical information necessary for planning, suggesting that listeners planned their own response as soon as they could determine the likely answer to the question.

However, follow-up studies suggest that Bögels, Magyari and Levinson (2015) EEG findings could also indicate that participants were monitoring the speaker's utterance to determine when they could initiate articulation. Jongman, Piai, and Meyer (2020) found that the large positivity reported by was also linked to attention to the sequence end in a task where participants had to prepare and maintain an answer until they were given a cue to speak. Furthermore, Bögels, Magyari and Levinson (2015) used general knowledge questions, and so the answers likely had to be retrieved from episodic memory. Although previous research has found that the middle frontal and precentral gyri are associated with language production (Indefrey & Levelt, 2004), other studies report that the middle frontal gyrus may also be involved in episodic memory retrieval (e.g., Cabeza, 2002; Rajah, Languay, & Grady, 2011; Raz et al., 2005). Bögels, Magyari and Levinson (2015) did not observe the same pattern of activation in a control study, in which participants memorized the questions, but their results may still reflect the processes of retrieving the answer from memory for production.

Nevertheless, other studies provide converging evidence for early-planning. Magyari, De Ruiter, and Levinson (2017) used a similar paradigm to Bögels, Magyari and Levinson (2015) and had participants view pictures while answering questions such as *Which animal has a light switch and also a battery?*. In the late condition, both of the animals on-screen had a light

switch and another object, and so participants could not plan a response until they heard the final object name. In the early condition, only one of the animals had objects, and so participants could plan a response even before they heard the question. Much like Bögels, Magyari and Levinson (2015), participants answered more quickly in the early ($M = 320$ ms) than the late condition ($M = 361$ ms), suggesting participants planned a response earlier when they knew the likely answer to the question compared to when they did not. But note that the difference between the two conditions was much smaller than in Bögels, Magyari and Levinson (2015) study, suggesting that the gain in response planning was not particularly large.

Results leading to a similar conclusion were reported by Meyer, Alday, Decuyper, and Knudsen (2018), who had participants answer questions (e.g., *Do you have a green sweater?*) while viewing four objects on-screen (e.g., a cake, a branch, a sweater, and a barrel). In the early condition, all the objects were the same color, and so participants could start planning an answer as soon as they understood the color adjective—for example, they knew as soon as they heard *green* that the answer would be *yes* if the objects were green and *no* if they were a different color. In the late condition, the objects were different colors and so participants could not plan an answer until the speaker produced the object name. Participants answered more quickly in the early ($M = 215$ ms) than the late ($M = 297$ ms) condition.

Similarly, we tested whether content prediction facilitates response planning in a set of *yes/no* question-answering studies (Corps, Crossley, Gambi, & Pickering, 2018). In one condition, the final words of the question were predictable (e.g., *Are dogs your favorite animal?*) because the majority of participants agreed on this final word as a continuation in a cloze pre-test. In the other condition, the final words were unpredictable (e.g., *Would you like to go to the supermarket?*) and participants provided different continuations in the cloze pre-test—even though some participants completed the question with the word *supermarket*, others responded with different words like *cinema* or *dentist*. We found that participants answered more quickly when the final words of the question were predictable ($M = 379$ ms; Experiment 2b) rather than unpredictable ($M = 536$ ms), suggesting they predicted the speaker's final word and used this prediction to plan a response. In other words, content prediction facilitated response planning.

Support for early-planning also comes from picture naming studies. Barthel, Sauppe, Levinson, and Meyer (2016); see also Barthel, Meyer, and Levinson (2017) used a task in which German participants completed a confederate's pre-recorded utterances. Participants had to name any

on-screen objects that the confederate had not already named, and so they could (in principle) plan their response as soon as the confederate began uttering their last object name (indicated by the use of the word *and*; e.g., *I have a door and a bicycle*). Both eye movements and response latencies suggested that participants planned their response as soon as possible—they were faster to speak when there was a clear lexical cue (i.e., *and*) to the end of the list ($M = 761$ ms) than when there was not ($M = 867$ ms).

In sum, there is good evidence that listeners (as next-speakers) engage in early-planning during laboratory tasks designed to approximate the mechanisms involved in dialog. As a result, the listener plans their response while simultaneously comprehending the current speaker's utterance. In the next section, I discuss the cognitive demands of dual-tasking comprehension and production.

3.5 Early response planning is cognitively demanding

Although there is much experimental evidence to suggest that listeners plan a response early, as claimed by Levinson and Torreira (2015), participants' average response times were always longer than the 200 or 300 ms typically reported in corpus studies (e.g., Stivers et al., 2009). This difference is not particularly interesting—in some studies, participants had to answer general knowledge questions or name pictures, which likely involved memory search processes or object recognition before a response could actually be planned. What is interesting, however, is that the average gain in response times in the early relative to the late condition was much less than the time difference between the occurrence of these two cues. For example, participants in Bögels, Magyari and Levinson (2015) study responded around 300 ms earlier when the critical information necessary for answer planning occurred early rather than late. But the cue that enabled response planning (e.g., *007*) occurred on average 1700 ms earlier in the early than the late condition. Thus, the gain in response time did not match the gain in information, and 1400 ms were "lost."

This inefficiency likely occurs because listeners who plan early must represent both the speaker's utterance (using comprehension mechanisms) and their planned response (using production mechanisms). Both production and comprehension require central attention (see Jongman, 2021, for a review), and so dual-tasking them should be cognitively demanding. As a result, planning early may interfere with simultaneous comprehension (and vice versa). In fact, research suggests that all stages of response planning

are cognitively demanding (e.g., Cook & Meyer, 2008; Ferreira & Swets, 2002; Roelofs, 2008; Roelofs & Piai, 2011).

In addition, comprehension and production are two very similar tasks, relying on similar neural circuits (e.g., Menenti, Gierhan, Segaert, & Hagoort, 2011; Silbert, Honey, Simony, Poeppel, & Hasson, 2014). For example, Segaert, Menenti, Weber, Petersson, and Hagoort (2012) found that the same brain areas (the left inferior frontal gyrus, the left middle temporal gyrus, and the bilateral supplementary motor area) were sensitive to syntactic repetition during comprehension and production. Furthermore, the representations for lexical concepts and lemmas are shared between production and comprehension. In the classic picture-word interference (PWI) paradigm, participants name pictures while ignoring simultaneously presented auditory or written distractor words (e.g., Schriefers et al., 1990). These studies have shown that participants are slower to name a picture (e.g., a dog) when the distractor word is semantically related (e.g., *cat*) rather than unrelated, suggesting that there is competition between shared representations of concepts during production (the target) and comprehension (the distractor).

This representational similarity is important because speakers' adjacent utterances are thought to be highly related in conversation, and research suggests that performance on one task suffers more when the other task is more rather than less similar (e.g., Wickens, 2008). For example, Fairs, Bögels, and Meyer (2018) used a psychological refractory period (PRP) paradigm, in which participants completed two separate tasks (Task A and Task B). The authors manipulated the interval between the start of Task B and the start of Task A by varying the stimulus onset asynchrony (SOA), so that participants sometimes completed the tasks in overlap. Participants experienced more interference when performing a picture-naming task alongside a syllable-identification task than when they performed a picture-naming task alongside tone-identification. These results suggest that the phonological representations used during syllable identification were also used during picture-naming, and competition occurred between comprehension and production when participants needed to use them simultaneously.

Thus, planning a response early may interfere with simultaneous comprehension. Research has recently begun to investigate this issue. In one study, Jongman and Meyer (2017) used a picture-naming task, in which half of the participants named the pictures (e.g., apple) while the other half listened to a pre-recorded speaker name the picture (i.e., planning condition was manipulated between-participants). In addition, pictures were preceded

by auditory primes, which were either identical to (*apple*), associatively related to (*peel*), or unrelated to the target picture (*nail*). The authors found fastest naming latencies for pictures preceded by an identity prime, intermediate latencies for those preceded by an associatively related prime, and slowest latencies for those preceded by an unrelated prime. This priming pattern was the same regardless of whether or not participants named the non-target picture, suggesting that speech planning did not interfere with comprehension of the prime.

Jongman and Meyer replicated the identity priming effect in a second experiment, in which participants had to decide whether or not to name the picture at the start of each trial (i.e., planning condition was manipulated within participants). However, in this experiment they found an associative priming effect only when participants did not have to name the picture, suggesting that response planning interfered with comprehension. The lack of associative priming in the planning condition was likely related to the difficulty of the task. In Experiment 1, participants' task was predictable and they knew whether they would need to plan a response before picture onset. In Experiment 2, however, participants had to switch between planning and listening, which was likely cognitively demanding. This task-switching is particularly relevant for natural conversation, since the cognitive load is likely to be greater than in Jongman and Meyer's task given that participants often have to plan (and comprehend) longer, more complex utterances.

In another study, Bögels, Casillas, and Levinson (2018) used a similar paradigm as their earlier study (Bögels, Magyari, & Levinson, 2015), but participants viewed pictures on-screen (e.g., a banana and a pineapple) while simultaneously answering questions. Much like the previous study, the critical information (here *curved*) necessary for response planning was available either early (e.g., *Which object is curved and is considered to be a type of fruit*?) or late (e.g., *Which object is considered to be a type of fruit and is curved*?). But in addition, the questions contained either an expected or unexpected word (e.g., *healthy* rather than *fruit* in both examples). The authors found that participants responded later to questions with an unexpected rather than expected word, regardless of when the critical information occurred, suggesting that they still comprehended these words even when they planned their response early. In addition, an N400 effect occurred at the unexpected word in both planning conditions. However, the size of this effect varied as a result of participants' response times: Participants with slower response times showed a larger N400 effect than those with faster response times. Based on these results, the authors concluded that fast responders allocated fewer resources to comprehension (leading to a smaller N400) and more to

production (leading to faster response times) when they encountered the information necessary for response planning. In contrast, slow responders allocated more resources to comprehension (leading to a larger N400) and fewer to production (leading to slower response times). Thus, this study provides some preliminary evidence that response planning interferes with comprehension.

Thus far, I have focused on studies that show planning interferes with comprehension. These studies could also demonstrate that comprehension interferes with planning. For example, the slow responders in Bögels et al.'s (2018) study may have been slower than the fast responders because comprehending hindered their response planning. But more direct evidence comes from PWI studies, which have shown that participants are slower to name pictures in the presence of words (even when the words are unrelated) than pseudowords (e.g., Dhooge & Hartsuiker, 2012), noise (e.g., Schriefers et al., 1990), or strings of X's (e.g., Glaser & Glaser, 1989), suggesting that comprehending a distractor word (even when you are told to ignore it) interferes with planning the picture name.

In dialog, however, speakers rarely hear words in isolation—they tend to be produced in sentence context. Recently, He, Meyer, and Brehm (2021) investigated whether unrelated background speech interferes with response planning. Dutch participants named a set of six pictures while they simultaneously ignored speech produced by a Dutch talker (high similarity speech), speech produced by a Chinese talker (moderate similarity speech), or eight-talker babble (low similarity speech). Participants were slower to name the pictures when they had to ignore the Dutch talker compared to the Chinese talker, and pictures in both of these conditions were named slower than in the eight-talker babble condition. These findings indicate that comprehension interferes with planning, but the degree of this interference is affected by the similarity of production and comprehension representations—when these representations are more similar (i.e., the same language), interference is higher than when they are less similar (i.e., a different language).

In sum, it is clear that early-planning is cognitively demanding. Not only is there evidence that planning interferes with comprehension, but comprehension also interferes with planning.

4. Is early-planning really necessary in dialog?

There is clear evidence that speakers plan a response early (Section 3.4), but there is also evidence that planning in this way is difficult (Section 3.5). Does language production need to be this difficult during

dialog? In other words, do listeners always need to plan their response while still comprehending the current speaker's utterance? In the next sections, I discuss research that suggests the answer to this question is no, and language production in dialog may be easier than claimed by theories based on laboratory studies (e.g., Levinson & Torreira, 2015). Note that I am not claiming that early-planning never occurs during dialogue. In fact, early-planning is likely to be particularly useful during highly constrained interactions where speakers have clear expectations about what they are likely to say. Rather, I suggest that the need for early-planning may have been overestimated by theories of dialogue.

4.1 Speakers often do not directly respond to each other

Theories of dialog (and the experimental studies testing them) typically think of dialog as like a game of ping pong: The speaker produces an utterance, and the listener uses the content of that utterance to plan an appropriate response. As a result, the content of the speaker's utterance constrains the content of the listener's utterance, and the length of the speaker's utterance constrains the amount of time the listener has for planning. To formulate a response to the previous speaker's utterance, the next speaker must begin planning a response early if they are to achieve turn gaps of 200 ms. Thus, comprehension and production overlap.

In a recent corpus analysis (Corps et al., 2022), however, we observed that many natural dialogs often involve parallel talk, where each speaker develops their turn in parallel with the other speaker over several utterances (or what we refer to as segments, which are stretches of speech produced by one speaker). For example, in (4) from the Santa Barbara corpus of American English (Du Bois et al., 2000), Phil formulates a lunch invitation while Brad talks about a third party (Pat, referred to as *her*). Note that the square brackets indicate overlap. In (5), which is from the German Corpus (GECO; Schweitzer & Lewandowski, 2013), Speaker 31 describes where they live, while Speaker 32 develops a question. Note that the numbers in the square brackets indicate the length of the gap or overlap between speakers.

(4) Phil: ..W- .. w-.. why don't you call me at least a little bit later [maybe,
Brad: [Yeah].
Phil: and] we can [go do that].
Brad: [Can I] do that? Cause I .. she'll be .. Uh ..
Phil: [Ji- .. Jim and I are gonna] have lunch,
Brad: Uh .. I don't want to get her uh ..]

Phil: I don't know if you have plans or not. But we're gonna have lunch later at noon.

(5) 1. Speaker 32: Ja, (Yes) [0.11].
2. Speaker 31: Also, (Well,) [−0.01].
3. Speaker 32: klar (of course) [−0.13].
4. Speaker 31: Kries Böblingen und (district Böblingen and) [−0.2].
5. Speaker 32: Mhm (Uhm) [−0.35].
6. Speaker 31: ahm¨ …das kleine Dorf daneben Ehningen…da (uhm…the small village next to Ehningen…there) [0.08].
7. Speaker 32: Und (And) [−0.13].
8. Speaker 31: wohnen wir (we live) [−0.19].
9. Speaker 32: Du fährst eine dreiviertel Stunde? (you travel three-quarters of an hour?) [−0.12].
10. Speaker 31: Ja (Yes) [−0.02].

For these stretches of parallel talk, the issue of how the listener responds to the speaker's segment does not actually arise because the listener does not respond to the immediately preceding segment at all. Instead, they continue a segment they produced earlier. In these cases, the listener's response does not depend on the content of the speaker's immediately preceding segment—instead, their response depends on the content of a segment they previously produced. As a result, the duration of the speaker's segment does not limit the listener's planning time for their utterance, and so comprehension and production processes do not need to extensively overlap.

We determined the occurrence of such parallel talk by analyzing corpora of conversations in German, Dutch, and American English. The German Corpus (GECO) consisted of 24 face-to-face conversations between two strangers, the Dutch Corpus (Corpus Gesproken Nederlands; CGN) consisted of 18 face-to-face conversations between two friends or family members, and the English corpus (Santa Barbara Corpus of Spoken American English) consisted of 11 face-to-face conversations between two friends or family members. In all corpora, participants were free to talk about anything they liked, and so there was minimal constraint on their utterances. In the German and Dutch corpora, each row in the transcript represented a single word produced by a speaker. In the English corpus, each row in the transcript represented an intonational unit, which is a "stretch of speech uttered under a coherent intonation contour" (Du Bois, Schuetze-Coburn, Paolino, & Cummings, 1992, p. 17). We created segments by collapsing all words or intonation units produced by one speaker in a stretch of speech before a speaker switch (i.e., a same-speaker stretch of speech).

We determined the occurrence of parallel talk by coding whether or not each segment was a continuation of an earlier segment produced by the same speaker. We considered a segment to be a continuation if it contributed to completing an earlier, syntactically incomplete segment. For example, in (5) the segments in lines four, six, and eight were coded as continuations because word meaning and grammatical structure indicated that they belonged to one utterance produced by Speaker 31. If the previous segment was syntactically complete, then we considered the next segment to be a continuation only if the two segments were unambiguously linked by a pronoun or a conjunction.

Although we were primarily interested in segments that were continuations of a previous segment, we also included a number of other categories. For comparison with the continuations, the most important of these are direct responses. These direct responses occurred when one speaker produced an answer to the previous speaker's question (much like those utterances studied in laboratory experiments), expressions of disagreement (e.g., *That's right indeed* or *No, that was before my time*), literal repetitions of parts of the partner's segment (e.g., Speaker A: *...in a boarding school* Speaker B: *In a boarding school!*), segments that referred directly back to the previous speaker's preceding segment, such as with a pronoun (e.g., Speaker A: *I don't have the ambition to speak flawless French one day* Speaker B: *Which actually is almost impossible*), or elaborations and associations (e.g., Speaker A: *My boyfriend's brother had a neighbor who used to cut his lawn meticulously* Speaker B: *With nail scissors*). These direct responses are very similar to the utterances elicited in laboratory studies, where one speaker asks a question and the participant responds.

In the German corpus, we found that 43% of the segments were continuations. These continuations occurred either after the previous speaker had produced a backchannel, such as *uh huh* or *yeah* (19%), or after the previous speaker had produced a segment of their own (24%). In contrast, only 17% of the segments were direct responses. In the Dutch corpus, 48% of the segments were continuations, with 9% produced after the previous speaker produced a backchannel and 39% produced after the previous speaker produced a segment. Only 21% of the segments were direct responses. Finally, in the English corpus 30% of the segments were continuations, either after a backchannel (16%) or after another segment (14%). In this corpus, the proportion of direct responses was 24%—much higher than in the Dutch and German corpora.

Although there were differences across the corpora, our analysis demonstrates that parallel talk regularly occurs in different languages and

conversational settings. In cases of such parallel talk, speakers continue a segment they have produced previously, rather than directly responding to the immediately preceding segment produced by the previous speaker. As a result, the listener can plan the content of their utterance independently from the content of the current speaker's utterance. In these cases, the question of how speakers manage to respond to each other's utterances so quickly does not arise—the speakers do not directly respond to each other at all, and the duration of the speaker's segment does not limit the listener's planning time. Thus, language production may be particularly difficult in laboratory tasks because speakers are encouraged to directly respond to each other, and produce pragmatically appropriate utterances (e.g., an answer to a question).

Note that these findings do not suggest that each speaker is holding a separate monolog. Informal inspection of the corpora suggested that successive segments in parallel talk may appear unrelated (i.e., segment two may not appear to be a direct response to segment one), but the turns developed by the two speakers are often related. Speakers usually refer to a common theme, as illustrated in (5), where both speakers talk about Speaker 31's home town. Thus, interlocutors are conversing with each other, but there is not necessarily a close content dependency between their utterances.

4.2 Incrementality and disfluency in dialog

Language production may also be difficult in laboratory tasks because speakers are typically encouraged to produce well-formed utterances, which are syntactically complete and do not contain any disfluencies, such as *uh* or *um*. As a result, participants are encouraged to plan a full utterance before speaking—if they do not, then they risk producing disfluent utterances. Planning in this way may make production difficult, given that there is much evidence for incremental planning in monolog (see Section 2.2). One consequence of this incrementality is that speakers are often disfluent, producing filled pauses such as *uh* or *um* (see Section 2.2.1). Although there is much research showing that utterances are disfluent, this disfluency has been underestimated by theories of dialog, and particularly by Levinson and Torreira (2015), because participants in laboratory tasks are discouraged from producing disfluent utterances. In particular, they have focused on fluent, idealized utterances, with the implicit assumption that disfluencies need to be excluded to study the mechanisms of conversation in their purest form (e.g., Bögels, Magyari, & Levinson, 2015). This point is important because it suggests that speech elicited in laboratory tasks designed to understand the

mechanisms of language production in dialog may be very different, and in fact more difficult to produce, from speech as it naturally occurs.

To investigate how much conversational speech deviates from laboratory speech, I conducted further analyses of the Santa Barbara Corpus described in Section 4.1, focusing again on the 11 face-to-face conversations between two people. This corpus has already been used to study disfluency in an analysis by Tottie (2014), but Tottie focused solely on the occurrence of *uh* and *um*. These filled pauses are thought to mark hesitations by the speaker, and could be used to hold the floor while further planning occurs (see Section 2.2.1). I was interested in these filled pauses, but when analyzing the corpus for instances of parallel talk, I noticed that utterances could be disfluent in a number of different ways. For example, speakers often produced discourse markers, such as *well* or *you know*, which are "sequentially dependent elements which bracket units of talk" (Schriffin, 1987). They can be removed from an utterance without altering its meaning or grammaticality (Schourup, 1999). Much like filled pauses, speakers may produce these discourse markers as hesitations, to buy time for further planning. Research suggests that different filled pauses and discourse markers likely have different functions (e.g., Clark & Fox Tree, 2002; Fox Tree & Schrock, 2002; Fuller, 2003). However, my aim was not to determine the different uses of these filled pauses and discourse markers, but rather to illustrate that they occur and contribute to the (dis)fluency of dialog.

Additionally, utterances were often incomplete (6) or contained repetitions (often referred to as self-repairs in the Conversation Analysis literature; e.g., Schegloff, Jefferson, & Sacks, 1977), taking many attempts before successful articulation (7). In these instances, speakers had likely planned part of their utterance, and finished articulating it before they had planned the next part of their utterance. As a result, they abandoned or reformulated their utterance. In other words, incomplete and repeated utterances provide further evidence that planning is incremental. Table 1 provides counts and percentages for the different disfluency categories I considered. I will discuss each of these categories in more detail below, but a full coding criteria (along with examples) can be found at https://osf.io/7aphq/.

(6) Lynne: Cause y- I mean you get so tired.
(7) Lenore: I thought they used the horsehooves in .. for gelatin.

Although previous research has extensively quantified the frequency of discourse markers, repetitions, and filled pauses in corpora (e.g., Crible, 2019; Crible, Degand, & Gilquin, 2017; Crible, Dumont, Grosman, & Notarrigo, 2019; Crible & Pascual, 2020), this work has not considered these findings in

the context of theories of dialog, such as Levinson and Torreira (2015). Furthermore, these corpora have often been based on highly restricted tasks, such as describing a route around a map (e.g., Branigan et al., 1999), and have tended to focus on limited disfluency types. Knowing what people say and how they speak in natural dialog is not only critical for determining whether laboratory speech is a good proxy for natural speech, but also for generating theories of speaking in dialog.

Before I discuss the coding criteria I used for identifying disfluent utterances, it is worth noting that previous research has shown that backchannels are common in spontaneous conversation (e.g., Knudsen, Creemers, & Meyer, 2020). The forms and functions of backchannels have been widely discussed from linguistic and psychological perspectives (e.g., Bangerter & Clark, 2003; Clark & Krych, 2004; Tolins & Fox Tree, 2014). They indicate to the present speaker that they should continue talking either by proceeding in their narrative or elaborating it (e.g., Schegloff, 1982, 2000; Tolins & Fox Tree, 2016). These backchannels are unlikely to contribute to disfluency—in fact, they likely contribute to the flow of dialog by allowing the listener to respond without planning a full utterance. However, I still quantified their occurrence because some discourse markers (such as *hmm*) could be produced as backchannels. Table 1 shows that 17% of the

Table 1 Frequencies (*n*) and proportions (%) of backchannels, incomplete segments, repetitions, resumptions, discourse markers, and filled pauses for segments in the Santa Barbara Corpus of Spoken American English.

	Total segments N = 3190	
	N	**%**
Backchannels	533	16.71
Incomplete segments	912	28.59
Interruptions	300	9.40
Repetitions (or self-repairs)	571	17.90
Resumptions (after interruption)	69	2.16
Segments containing at least one filled pause	531	16.65
Segments containing at least one discourse marker	879	27.55
Disfluent segments, containing at least one category	1854	58.12

Note that these categories were not mutually exclusive, and so a segment could belong to more than one category (i.e., it could contain both a filled pause and a discourse marker). The final row in the table shows the number of segments that were disfluent, and contained at least one category. In particular, a segment was disfluent if it was incomplete, interrupted, repeated, resumed, or contained a filled pause or discourse marker, regardless of how many of these phenomena occurred in the segment.

segments were backchannels (calculated as the number of segments containing a backchannel divided by the total number of segments).

Incomplete segments were those that contained an incomplete word or were abandoned by the speaker and were not resumed in any of the surrounding segments (i.e., the whole segment was incomplete). Incomplete segments also included those in which the speaker was interrupted by their partner and so did not finish their utterance. I also considered segments in which the speaker repeated themselves (e.g., *you have to-to graduate*) to be incomplete because the initial portion was incomplete and subsequently repeated. Note that segments could be incomplete in more than one way. For example, it could contain an incomplete word, be resumed, and then subsequently be abandoned by the speaker so the whole segment is incomplete. I did not determine how many times each segment was incomplete—it was considered incomplete if it belonged to any of these categories. In total, 29% of the segments were incomplete, with 9% of them being incomplete because the other speaker interrupted.

When segments were incomplete, speakers often began a new segment by repeating part of their earlier, incomplete segment. To determine how often speakers repeated part of their segment, I identified segments that contained repetitions or that were resumed after an interruption by another speaker. Again, I did not determine how many times each segment was repeated. Rather I considered an utterance to be a repetition if it was repeated at least once. In total, 18% of the segments were repetitions, while 2% were resumptions of an earlier, interrupted segment.

When coding the discourse markers and filled pauses, I considered words (such as *well* or *you know*) and sounds (such as *uh* or *um*) to be discourse markers or filled pauses if they could be removed from the segment without altering the speaker's meaning. For example, *you know* would be considered a discourse marker in a segment such as *And doing it and stuff you know*, but not in a segment such as *Do you know what I mean*?. Table 2 shows the counts and percentages for the individual filled pauses and discourse markers. Segments could contain multiple occurrences of the *same* filled pause or discourse marker. For example, the speaker could produce *uh* multiple times in the same segment. But since I was interested in how many segments contained at least one occurrence of each type of filled pause or discourse marker, Table 1 shows the number of times the speaker produced a particular type of filled pause or discourse marker at least once in a segment. In total, 17% of the segments contained at least one filled pause, and 28% of the segments contained at least one discourse marker.

To determine how often segments were disfluent, I determined how many were incomplete, interrupted, repeated, resumed, or contained a filled pause or a discourse marker. Segments were considered disfluent if they fell into any one of these categories. In total, 58% of the segments were disfluent, and so around only 40% of the segments contained no disfluency and were similar to the idealized utterances elicited in laboratory tasks studying the mechanisms of speaking in dialog.

These findings add to an existing body of research that has shown that spontaneous speech is disfluent (see Section 2.2.1), and suggest that speech planning is incremental. Speakers are likely incremental in this way because planning while comprehending is cognitively demanding (e.g., Oomen & Postma, 2001). Although corpora analyses do not allow us to draw conclusions about the direction of causality, there is some evidence that the fluency of speech is affected when speakers dual-task production and comprehension. For example, Boiteau, Malone, Peters, and Almor (2014) had participants conduct a visuomotor tracking task while simultaneously interacting with a confederate. Participants' tracking performance declined towards the end of the confederate's turn, suggesting they began response planning at this point. Participants' speech rate was also affected by concurrent tracking when they had to plan a response compared to when they just had to listen, but there was no evidence that planning while listening increased the number of disfluencies participants produced. However, the authors considered only *um*s and *uh*s, but it is clear from Tables 1 and 2 that there are many other types of disfluencies.

This incrementality (and disfluencies, by extension) invites parallel talk. Speakers (Speaker A) do not plan their full utterance before they speak, and so they may often pause or hesitate while they plan later parts of their utterance, leading to disfluent speech. This hesitation allows the other speaker (Speaker B) to jump in and articulate their own increment. Speaker A then articulates the rest of their utterance, and so they do not directly respond to the immediately preceding utterance of Speaker B. Thus, incrementality, disfluencies, and parallel talk are closely related to each other.

These findings have important consequences for the way we think about language during dialog. First, they suggest that the utterances we study in the laboratory are very different from the utterances speakers actually produce in natural conversation. This point may seem obvious, but it has important consequences for Levinson and Torreira (2015) theory, which has been used to motivate many studies investigating the mechanisms of speaking during dialog. In particular, Levinson and Torreira (2015) claim that next speakers

Table 2 Frequencies (*n*) and proportions (%) of different types of filled pauses and discourse markets in the Santa Barbara Corpus of Spoken American English.

Filled pause	***N***	**%**
Uh	326	10.22
Oh	134	4.20
Hm	66	2.07
Huh	23	0.72
Ah	11	0.35
Uhuh	4	0.13
Aw	7	0.22
Total filled pauses	571	17.42
Discourse markers		
You know	315	9.88
Well	252	7.90
So	170	5.33
Like	164	5.14
I mean	115	3.61
Kinda	74	2.32
Geez	59	1.85
Man	59	1.85
Oh God	34	1.07
Right	33	1.04
Pretty	28	0.88
See	27	0.85
Really	19	0.60
Now	17	0.53
Sorta	15	0.47
Anyway	13	0.41
Total discourse markers	1394	43.70

Note that these categories were not mutually exclusive, and so a segment could contain more than one filled pause or discourse marker.

must complete all stages of response planning as early as possible (i.e., as soon as they can identify the gist of the current speaker's utterance) if they are to achieve timely turn-taking and respond within 200 ms. But such early-planning may not be necessary in natural conversation—speakers could use disfluencies to hold their turn while planning their utterance, thus minimizing the overlap between production and comprehension processes.

Relatedly, experimental studies investigating production in dialog likely make production harder than it needs to be. First, participants are often encouraged to plan well-formed utterances, and any utterances containing disfluencies are often excluded from analyses. Participants may thus be discouraged from planning incrementally, and may instead plan their complete utterance before they speak in an effort to ensure they produce well-formed utterances. Relatedly, our corpora analyses (Corps et al., 2022) have demonstrated that speakers do not always directly respond to each other—instead, they develop their utterances in parallel and continue an utterance they produced previously. This situation is very different from laboratory tasks, where participants often need to directly respond to the previous speaker and the content of their own utterance depends on the content of the previous speaker's utterance. As a result, speakers likely engage in more extensive advance planning (resulting in a larger overlap between production and comprehension) in laboratory tasks than there needs to be in natural conversation, thus contributing to turn gaps longer than 200 ms.

In sum, it is clear that these theories are missing an important part of natural speech—namely, that speakers are highly disfluent. Thus, these results have important methodological and theoretical consequences, and suggest that we need to study production both in highly controlled laboratory tasks and in natural conversation if we are to build a clear picture of the mechanisms of speaking during dialog (see also De Ruiter & Albert, 2017). In particular, future experimental work could take excerpts from speech corpora and test how disfluencies affect the accuracy of when speakers articulate their responses. Additionally, they could also test how disfluencies affect how participants distribute their attention between response planning and simultaneous comprehension. Finally, research could investigate whether parallel talk is more common in instances where speakers hesitate and produce disfluencies. Testing these hypotheses would provide insight into how comprehension, response planning, and articulation are interwoven during conversation, and would allow researchers to develop theories of language production in natural dialog.

What these findings demonstrate, however, is that we currently do not have a clear picture of speaking in dialog, like we do in monolog, because these studies have tended to focus on highly idealized utterances, often ignoring the fact that production is highly incremental, flexible, and far from perfect.

5. Conclusions

During dialog, interlocutors take turns at speaking with little gap or overlap between their contributions. But language production in monolog is comparatively slow. Theories of dialog tend to agree that interlocutors manage these timing demands by planning a response early, before the current speaker reaches the end of their turn. As a result, there is overlap between production and comprehension processes. Much experimental research supports these theories, but this research also suggests that planning a response early, while simultaneously comprehending, is difficult. Does language production need to be this difficult during dialog? In other words, is early-planning always necessary?

In the second half of this chapter, I discussed research from our lab that suggests the answer to this question is no. In particular, we analyzed corpora of naturally occurring conversations in German, Dutch, and English. We found that speakers often do not directly respond to each other during dialog—instead, they continue an utterance they produced earlier. In these instances of parallel talk, the next speaker's response does not depend on the content of the current speaker's utterance, and so the next speaker's planning time is not constrained by the current speaker's utterance. As a result, comprehension and production do not need to extensively overlap.

This parallel talk likely occurs because speakers are highly incremental. In particular, we also found that speakers are highly disfluent, suggesting they do not plan a full utterance before beginning articulation. This incrementality has not been considered by theories and experimental studies of dialog, which typically focus on idealized utterances. Note that I am not claiming that early-planning never occurs—in fact, it is likely particularly useful in highly constrained interactions (such as question-answering), where speakers do directly respond to each other and must do so in a timely manner. But together, these corpora analyses demonstrate that language production studied in laboratory experiments is very different from how language production actually occurs in natural conversation. Thus, further research using naturalistic tasks is needed to investigate the mechanisms of dialog.

Acknowledgments

I thank Antje Meyer and Martin Pickering for useful comments on earlier versions of this chapter.

References

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

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, *32*, 25–36.

Arnold, J. E., & Tanenhaus, M. K. (2011). Disfluency effects in comprehension: How new information can become accessible. *The Processing and Acquisition of Reference*, 197–217.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*, 578–582.

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, *27*, 195–225.

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*. 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*. https://doi.org/10.3389/fpsyg.2016.01858.

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

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

Bögels, S., Kendrick, K. H., & Levinson, S. C. (2020). Conversational expectations get revised as response latencies unfold. *Language, Cognition and Neuroscience*, *35*, 766–779.

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

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

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

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*, 123–147.

Bosker, H. R., Tjiong, V., Quené, H., Sanders, T., & De Jong, N. H. (2015). Both native and non-native disfluencies trigger listeners' attention. In *Disfluency in Spontaneous Speech: DISS 2015* An ICPhS Satellite Meeting. DISS2015.

Branigan, H., Lickley, R., & McKelvie, D. (1999). Non-linguistic influences on rates of disfluency in spontaneous speech. In *Proceedings of the 14th International Conference of Phonetic Sciences* (pp. 387–390).

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482–1493.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*, 325–337.

Brown-Schmidt, S., & Konopka, A. E. (2015). Processes of incremental message planning during conversation. *Psychonomic Bulletin & Review*, *22*, 833–843.
Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609.
Butterworth, B. (1989). Lexical access in speech production. In W. Marslen-Wilson (Ed.), *Lexical representation and process*. Cambridge, MA: MIT Press.
Cabeza, R. (2002). Hemispheric asymmetry reduction in older adults: The HAROLD model. *Psychology and Aging*, *17*, 85–100.
Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*, 177–208.
Caramazza, A., & Miozzo, M. (1998). More is not always better: A response to Roelofs, Meyer, and Levelt. *Cognition*, *69*, 231–241.
Chafe, W. (1992). Information flow in speaking and writing. *The Linguistics of Literacy*, *21*, 17–29.
Clark, H. H. (1994). Managing problems in speaking. *Speech Communication*, *15*, 243–250.
Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition*, *64*, 1–37.
Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73–111.
Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81.
Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 696–702.
Cook, A. E., & Meyer, S. (2008). Capacity demands of phoneme selection in word production: New evidence from dual-task experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 886–899.
Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you say it: Disfluency in speech affects the comprehension process. *Cognition*, *105*, 658–668.
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.
Corps, R. E., Gambi, C., & Pickering, M. J. (2018). Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. *Discourse Processes*, *55*, 230–240.
Corps, R. E., Knudsen, B., & Meyer, A. S. (2022). Overrated gaps: Inter-speaker gaps provide limited information about the timing of turns in conversation. *Cognition*, *223*. https://doi.org/10.1016/j.cognition.2022.105037.
Crible, L. (2019). *Discourse markers and (dis)fluency. Forms and functions across languages and registers*. Amsterdam/Philadelphia: John Benjamins.
Crible, L., Degand, L., & Gilquin, G. (2017). The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis) fluency. *Languages in Contrast*, *17*(1), 69–95.
Crible, L., Dumont, A., Grosman, I., & Notarrigo, I. (2019). Application of an interoperable annotation scheme. *Fluency and Disfluency Across Languages and Language Varieties*, *4*, 17.
Crible, L., & Pascual, E. (2020). Combinations of discourse markers with repairs and repetitions in English, French and Spanish. *Journal of Pragmatics*, *156*, 54–67.
De Ruiter, J. P., & Albert, S. (2017). An appeal for a methodological fusion of conversation analysis and experimental psychology. *Research on Language and Social Interaction*, *50*, 90–107.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321.
DeLong, K. A., Chan, W., & Kutas, M. (2018). Similar time courses for word form and meaning preactivation during sentence comprehension. *Psychophysiology*. https://doi.org/10.1111/pstp.13312.
DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: A commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience*, *32*, 966–973.
Dhooge, E., & Hartsuiker, R. J. (2012). Lexical selection and verbal self-monitoring: Effects of lexicality, context, and time pressure in picture-word interference. *Journal of Memory and Language*, *66*, 163–176.
Du Bois, Chafe, W. L., Meyer, C., Thompson, S. A., & Martey, N. (2000). Santa Barbara corpus of Spoken American English, Part 1–4. Philadelphia: Linguistic Data Consortium.
Du Bois, J. W., Schuetze-Coburn, S., Paolino, D., & Cummings, S. (1992). *Discourse transcription (Santa Barbara Papers in Linguistics). Vol. 4*. Santa Barbara: University of California, Santa Barbara, Department of Linguistics.
Eklund, R., & Shriberg, E. (1998). Crosslinguistic disfluency modelling: A comparative analysis of Swedish and American English human–human and human–machine dialogues. In *Vol. 6. 5th International Conference on Spoken Language Processing, 30th November–4th December, 1998, Sydney, Australia* (pp. 2627–2630).
Fairs, A., Bögels, S., & Meyer, A. S. (2018). Dual-tasking with simple linguistic tasks: Evidence for serial processing. *Acta Psychologica*, *191*, 131–148.
Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, *30*, 210–233.
Ferreira, V. S., & Slevc, L. R. (2007). Grammatical encoding. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 453–470). Oxford University Press.
Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*, 57–84.
Fox Tree, J. E. (2010). Discourse markers across speakers and settings. *Language and Linguistics Compass*, *4*, 269–281.
Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*, 151–167.
Fox Tree, J. E., & Schrock, J. C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language*, *40*, 280–295.
Fox Tree, J. E., & Schrock, J. C. (2002). Basic meanings of you know and I mean. *Journal of Pragmatics*, *34*, 727–747.
Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, *31*, 931–952.
Fuller, J. M. (2003). Use of the discourse marker like in interviews. *Journal of Sociolinguistics*, *7*, 365–377.
Garrod, S., & Pickering, M. J. (2007). Alignment in dialogue. *The Oxford Handbook of Psycholinguistics*, 443–451.
Garrod, S., & Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00751.
Gisladottir, R. S., Chwilla, D. J., & Levinson, S. C. (2015). Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PLoS One*, *10*. https://doi.org/10.1371/journal.pone.0120068.
Glaser, W. R., & Glaser, M. O. (1989). Context effects in stroop-like word and picture processing. *Journal of Experimental Psychology: General*, *118*, 13–42.
Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*, 544–569.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, *82*, B1–B14.
Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*, 274–279.
Griffin, Z. M., & Spieler, D. H. (2006). Observing the what and when of language production for different age groups by monitoring speakers' eye movements. *Brain and Language*, *99*, 272–288.
Hartsuiker, R. J., & Notebaert, L. (2009). Lexical access problems lead to disfluencies in speech. *Experimental Psychology*, *57*, 169–177.
He, J., Meyer, A. S., & Brehm, L. (2021). Concurrent listening affects speech planning and fluency: The roles of representational similarity and capacity limitation. *Language, Cognition and Neuroscience*, *36*, 1258–1280.
Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*, 555–568.
Howes, C., Purver, M., Healey, P. G., Mills, G. J., & Gregoromichelaki, E. (2011). On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, *2*, 279–311.
Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*, 101–144.
Ito, A., Corley, M., Pickering, M. J., Martin, A. E., & Nieuwland, M. S. (2016). Predicting form and meaning: Evidence from brain potentials. *Journal of Memory and Language*, *86*, 157–171.
Ito, A., Gambi, C., Pickering, M. J., Fullenbach, K., & Husband, E. M. (2020). Prediction of phonological and gender information: An event-related potential study in Italian. *Neuropsychologia*, *136*. https://doi.org/10.1016/j.neuropsychologia.2019.107291.
Ito, A., Martin, A. E., & Nieuwland, M. S. (2017). Why the a/an prediction effect may be hard to replicate: A rebuttal to DeLong, Urbach, and Kutas (2017). *Language, Cognition and Neuroscience*, *32*, 974–983.
Iwasaki, N., Vigliocco, G., & Garrett, M. F. (1998). Adjectives and adjectival nouns in Japanese: Psychological processes in sentence production. In D. J. Silva (Ed.), *Vol. 8. Japanese/Korean linguistics* (pp. 93–106). Stanford, CA: Center for the Study of Language and Information.
Jongman, S. R. (2021). The attentional demands of combining comprehension and production in conversation. In *Psychology of learning and motivation* (pp. 95–140). Academic Press.
Jongman, S. R., & Meyer, A. S. (2017). To plan or not to plan: Does planning for production remove facilitation from associative priming? *Acta Psychologica*, *181*, 40–50.
Jongman, S. R., Piai, V., & Meyer, A. S. (2020). Planning for language production: The electrophysiological signature of attention to the cue to speak. *Language, Cognition and Neuroscience*, *35*, 915–932.
Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133–156.
Kempen, G., & Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition*, *14*, 185–209.
Knudsen, B., Creemers, A., & Meyer, A. S. (2020). Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.593671.
Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *Journal of Memory and Language*, *66*, 143–162.
Konopka, A. E., & Meyer, A. S. (2014). Priming sentence planning. *Cognitive Psychology*, *73*, 1–40.

Koudenburg, N., Postmes, T., & Gordijn, E. H. (2013). Conversational flow promotes solidarity. *PLoS One*, *8*. https://doi.org/10.1371/journal.pone.0078363.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, Massachusetts: MIT Press.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–38.

Levelt, W. J., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, *98*, 122–142.

Levinson, S. C. (1995). Interactional biases in human thinking. In E. N. Goody (Ed.), *Social intelligence and interaction* (pp. 221–260). Cambridge: Cambridge University Press.

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

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

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

Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture–word interference experiments. *Journal of Memory and Language*, *35*, 477–496.

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*. https://doi.org/10.3389/fpsyg.2018.00525.

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

Miozzo, M., & Caramazza, A. (1997). Retrieval of lexical–syntactic features in tip-of-the tongue states. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1410–1423.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., et al. (2018). Large-scale replication study reveals a limit on probablistic prediction in language comprehension. *eLife*, 7. https://doi.org/10.7554/eLife.33469.

Oomen, C. C. E., & Postma, A. (2001). Effects of divided attention on the production of filled pauses and repetitions. *Journal of Speech, Language, and Hearing Research*, *44*, 997–1004. https://doi.org/10.1044/1092-4388(2001/078).

Peterson, R. R., & Savoy, P. (1998). Lexical selection and phonological encoding during language production: Evidence for cascaded processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 539–557.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*, 1002–1044.

Rajah, M. N., Languay, R., & Grady, C. L. (2011). Agre-related changes in right middle frontal gyrus volume correlate with altered episodic retrieval activity. *The Journal of Neuroscience*, *35*, 17941–17954.

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., et al. (2005). Regional brain changes in aging healthy adults: General trends, individual differences and modifiers. *Cerebral Cortex*, *15*, 1676–1689.

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

Roberts, F., Francis, A. L., & Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication*, *48*, 1079–1093.

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

Roelofs, A. (2008). Attention, gaze shifting, and dual-task interference from phonological encoding in spoken word planning. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 1580–1598.

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

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn-taking for conversation. *Langugage*, *50*, 696–735.

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*, 362–367.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.), *Analyzing text and talk* (pp. 71–93). Georgetown: Georgetown University Press.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, *29*, 1–63.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*, 361–382.

Schourup, L. (1999). Discourse markers. *Lingua*, *107*, 227–265.

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

Schriffin, D. (1987). *Discourse markers. No. 5*. Cambridge: Cambridge University Press.

Schweitzer, A., & Lewandowski, N. (2013). Convergence of articulation rate in spontaneous speech. In *INTERSPEECH 2013, Lyon, France* (pp. 525–529).

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

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, E4687–E4696.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*, 25–38.

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

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

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*, 485–496.

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

Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., & Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, *119*. https://doi.org/10.1073/pnas.2116915119.

Tolins, J., & Fox Tree, J. E. (2014). Addressee backchannels steer narrative development. *Journal of Pragmatics*, *70*, 152.
Tolins, J., & Fox Tree, J. E. (2016). Overhearers use addressee backchannels in dialog comprehension. *Cognitive Science*, *40*, 1412–1434.
Tottie, G. (2014). On the use of uh and um in American English. *Functions of Language*, *21*, 6–29.
Urbach, T. P., DeLong, K. A., Chan, W. H., & Kutas, M. (2020). An exploratory data analysis of word form prediction during word-by-word reading. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 20483–20494.
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*, 443–467.
Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of Italian tongues. *Psychological Science*, *8*, 314–317.
Vigliocco, G., Vinson, D. P., Martin, R. C., & Garrett, M. F. (1999). Is "count" and "mass" information available when the noun is not? An investigation of tip of the tongue states and anomia. *Journal of Memory and Language*, *40*, 534–558.
Wagner, V., Jescheniak, J. D., & Schriefers, H. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 423–440.
Wheeldon, L. (2013). Producing spoken sentences: The scope of incremental planning. In P. Perrier, & P. L. Verlag (Eds.), *Cognitive and physical models of speech production, speech perception, and production-perception integration*.
Wicha, N. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, *346*, 165–168.
Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455.