Predictions in conversation

Lilla Magyari

Norwegian Reading Centre for Reading Education and Research, Faculty of Arts and Education, University of Stavanger, Stavanger, Norway

This is an Author Accepted Manuscript version of the following chapter: Predictions in conversation, published in A life in cognition: Studies in cognitive science in honor of Csaba Pléh, edited by J. Gervain, G. Csibra, & K. Kovács, 2022, Springer Nature Switzerland AG. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-66175-5_5

Abstract

Natural conversations usually run smoothly and effortlessly with small gaps and overlaps between the turns of the interacting participants. Measurement of turn-taking timing revealed that interactants often reply within 200 ms after the end of the other's turn. Such precise timing has challenged current theories of language processing. Several models proposed in recent years have been trying to explain the way short turn-transitions are enabled by the interaction between speech production, comprehension, and particular cognitive processes. Most of these models propose that participants must predict the end of the current speaker's turn in order to prepare an answer in advance to be able to produce the next turn on time. In this paper, I discuss models of turn-taking and possible predictive processes during conversations.

Keywords: natural conversation, turn-taking, prediction, timing, language processing, speech production, speech comprehension

Introduction

In everyday interactions, natural conversations are often effortless and managing them seems to be easy. However, a closer look at the timing pattern of conversations implies a complex cognitive architecture underlying turn-exchanges. Corpus studies of recordings of natural conversations have shown that the duration of turn-transitions – the time between the end of a turn and the beginning of the new one – is most frequently around 200 ms (Stivers et al., 2009; Heldner & Edlund, 2010; Levinson & Torreira, 2015). Several studies have tried to explain how interactants manage conversations since the 1960's (Pléh, 2012). Recent models of turn-taking consider the cognitive processes of the speakers and listeners during conversations, and most agree that listeners need to predict the end of the turn of the current speaker, in order to produce a timely response. Different models, however, suggest different predictive mechanisms underlying smooth turn-transitions. Some accounts do not agree about when listeners who speak next start to prepare their turn, i.e. towards the end of a current turn or as soon as it is possible to formulate a response. Some models also do not agree about how listeners can predict the end of a turn. In this paper, it will be discussed whether the different accounts can be reconciled.

Spontaneity and constraints in conversations

Natural conversations are everyday verbal interactions between two or more participants. A seminal paper of conversation analysis (CA) by Sacks, Schegloff and Jefferson (1974) outlined the specific properties that distinguish informal, everyday conversations (e.g., a conversation between family members at the dinner table; hereinafter referred to as conversation) from other verbal exchange systems (e.g., teacher-student interaction in classroom, interviews, public speeches, etc.). Conversations are not organized in advance. There are no constraints regulating what is said, who speaks and for how long. In other type of interactions (i.e. in more formal or institutional interactions), the speaker's and the listener's role or the content of the interaction might be more regulated. For example, teachers speak more often and for longer than students in a class, and teachers control when students are allowed to speak.

Despite the absence of explicit constraints, the internal structure of conversation and the temporal pattern of turn-taking exhibit a systematic underlying organization (Sacks et al., 1974). Sack and his colleagues noted for example that

1) “Turn order is not fixed, but varies”

2) “Turn size is not fixed, but varies”

3) “Length of conversation is not specified in advance”

4) “What parties say is not specified in advance”

5) “Relative distribution of turns is not specified in advance”

6) “Number of parties can vary.” (Sacks et al., 1974, p. 700-701.).

Besides the varying features of conversations, they also observed stable aspects that later were claimed to be strongly universal (Stivers et al., 2009; 2010):

10) “Overwhelmingly, one party talks at a time”,

11) “Occurrences of more than one speaker at a time are common, but brief”,

12) “Transitions (from one turn to the next) with no gap or overlap are common. Together with transitions characterized by slight gap or overlap, they make up the vast majority of the transitions.” (Sacks et al., 1974, p. 700-701.).

Sacks and his colleagues proposed that the phenomenon that interactants try to minimize longer gaps and longer overlaps is a social norm in conversation. Later, corpus studies confirmed that short gaps and short overlaps are more common than longer ones. Sacks and his colleagues also claimed that no gaps or no overlaps (i.e. turn-transitions around 0 ms) are also common. However, corpus studies measuring turn-transitions with millisecond precision showed that the most frequent turn-transitions are not zero ms long. Most common turn-transitions are characterized by a short, circa 200 ms long gap (e.g. Stivers et al., 2009; Heldner & Edlund, 2010, Levinson & Torreira, 2015). For example, Stivers and colleagues (Stivers et al., 2009) examined ten languages ranging from those of traditional communities to major world languages and found that the distribution of turn-transition durations were unimodal with a mode offset between 0 to 200 ms for all the studied languages.

Sacks and his colleagues’ (1974) account is supported by observations regarding the sensitivity of interactants to the speed of their partner’s response. Deviations from almost

immediate responses, i.e. delays and overlaps, are interactionally consequential. When overlap occurs which is not an interruption, it is normally resolved by speaker withdrawal so that only one speaker remains (Schegloff, 2000). Many studies of CA demonstrate that some of the silences between turns have a functional role. A long gap after a question can indicate that the recipient has uncertainty about the response or find it difficult to answer for other reasons (see Levinson & Torreira, 2015). Silences might also occur preceding 'dispreferred' responses which fail to align with the action suggested in the prior turn (e.g., Pomerantz, 1984; Schegloff, 2007). For example, a rejection is dispreferred in response to an invitation. Extract (1) shows an example from an English telephone call where a rejection is preceded by a 300 ms long gap. A corpus analysis also showed that gaps of 700 ms or longer are associated with dispreferred actions (Kendrick & Torreira, 2015), and an experimental study found that overhearers interpret inter-turn silences of 600 ms or longer as an indication for rejection or disagreement (Roberts et al., 2011).

(1) Erhardt 1 (Schegloff, 2007, p. 68 from Kendrick & Torreira, 2015, p.5)

01 Kar: °Gee I feel like a real nerd° < You can all come up here,

02 → (0.3)

03 Vic: Nah, that's alright we'll stay down here,

Sacks and his colleagues (1974) describe three basic rules which govern the speaker-changes in the course of conversation and coordinate the floor transfer between participants to minimize the gaps and overlaps. These rules operate on variable sizes of syntactic units which can function as full turns. The end of such a unit is called "transition relevant place" where a possibility for speaker-change is present. The rules specify how speakers can take the floor:

(1) "If the current speaker C selects the next speaker N, then C must stop, and N should start. (…)
(2) If C does not select N, then any participant can self-select (…).
(3) If no other party self-selects, C may continue." (Levinson & Torreira, 2015, p. 11)

These rules can be recursively reapplied until someone will continue with speaking. Interestingly, these rules also predict the relative duration of silences within and between turns.

If intra-turn silences are generated by rule 3, i.e. no other participant took the floor after applying rule 2, silences should be longer than silences between turns of different speakers where the speaker started to speak by applying rule 1 or 2. A large corpus study found that gaps within a turn of the same speaker was 140 ms longer than inter-turn gaps (ten Bosch et al., 2005; but cf. Markó, 2006 cited in Pléh, 2012).

Sacks and colleagues' model has consequences for language processing as well. They noted that such fast turn-transitions can only be achieved if the listener who is going to be the next speaker can "project" the end of the current turn. They suggested that listeners predict the type of construction that a speaker is going to produce (e.g., a word, a phrase, clause, or a multi-clausal construction) to estimate when the turn will likely end. In contrast, others argued that no projection is necessary because observable cues appear in the speaker's speech or behaviour (e.g., eye-gaze) shortly before turn-endings that signal to the listener that the turn is coming to an end (e.g. Duncan, 1974; Duncan & Fiske, 1977).

Recently the temporal patterning of turn-taking and how it is achieved by conversational partners also have intrigued researchers. Short turn-transition times are surprising from a cognitive processing point of view. Listeners who are to speak next must accomplish several tasks in a short time window. They must, at the very least, sufficiently comprehend the turn in order to produce a relevant answer. Moreover, several decades of speech production literature show that speech production has a relatively long latency. Studies using picture naming tasks indicate that about 600 ms of preparation time is needed before the articulation of a single word begins (Levelt, 1989; Indefrey & Levelt, 2004; Indefrey, 2011) – far more than the 0-200 ms turn taking interval observed in natural interactions. This suggests that participants do not simply wait until the other has finished speaking and then start to speak, but must start speech planning in advance, before the current turn has finished (Levinson, 2013; Levinson & Torreira, 2015). Therefore, listeners probably execute parallel cognitive tasks for an immediate response. Listeners' comprehension and production processes might overlap towards the end of a turn, and predictive mechanisms must be at play. The short transitions also suggest that the next speaker times the production of the answer to the end of the current turn. The intriguing psycholinguistic question is, then, how does the speech production and comprehension system work together to achieve smooth and fluent turn-transitions?

Prediction of the content

Speech production is relatively slow

Although the time needed to prepare a conversational turn might differ from the time needed to prepare words or sentences in well-controlled laboratory settings, experiments still can provide information on how long it takes to formulate a response. The literature is extensive on speech production experiments, and it provides insight into the latencies of the different stages of speech production. There is some consensus about four major stages of speech production, although the architectural details are debated: 1) conceptual preparation, 2) lexical access, 3) phonological processing, and 4) articulation (Dell, 1986; Levelt, 1989; Caramazza, 1997). Word production studies showed that, given a picture to name, it takes from 420 ms to 2000 ms to retrieve and code a word for articulation (Indefrey, 2011). Naming latencies can vary with variation in the task or stimuli. For example, repetition of the same word, word length, familiarity, word-frequency, priming effects or cognate statues all influence naming times (e.g., Jescheniak & Levelt, 1994; Jescheniak, Schriefers & Hantsch, 2003; Strijkers, Costa & Thierry, 2010).

Natural language use, however, is characterized not only by the production of single words, but by grammatical constructions of varying length. If speakers must prepare larger units before articulation, we might expect a significant drag on preparation for speech production. In the speech production literature, there is a discussion about the typical size of the planned units before articulation. Proposals vary from radical incrementality, where only one word is planned at a time (e.g., Levelt & Meyer, 2000; Gleitman, January, Nappa & Trueswell, 2007), to the generation of the structural frame before production (Griffin & Bock, 2000; Bock, Eberhard & Cutting, 2004). Some studies argue that the time-course of speech formulation might be flexible and speakers might use planning units of different size under different circumstances (e.g., Wagner, Jescheniak & Schriefers, 2010; Konopka & Meyer, 2014).

Even if speakers need to prepare only the first word of a turn before articulation, the time-course of a single word production is still relatively slow compared to the tight timing of conversational turns. However, the time needed for the conceptual preparation of speech in picture naming studies may differ from the time needed for the conceptual preparation in conversational turns. The duration of the conceptual preparation could be either very fast due

to contextual constraints or very slow due to the difficulty of the response. For example, initiated greetings can narrow down the number of the possible appropriate responses for the next conversational partner. Table 1 shows which greeting is appropriate in response to an initial greeting in Hungarian. In contrast, refusals of invitations might be difficult to formulate due to its consequences. According to meta-analyses of picture naming studies, the average duration from the end of conceptual preparation to articulation is at least 400 ms (Indefrey & Levelt, 2004; Indefrey, 2011). This duration is still twice as long as the average gap between turns in conversation. In other words, it is likely that planning a next turn often precedes the ending of the current turn in natural conversations.

First turn \ *Second turn*	*Csókolom!*	*Jó napot!*	*Szervusz!*	*Szia!*
Csókolom!			+	+
Jó napot!	+	+		
Szervusz!	+		+	+
Szia!	+			+

Table 1. Appropriate adjacency pairs of Hungarian greetings (from Pléh, 2012, p.70)

Language comprehension is predictive

It is not obvious how listeners prepare an answer to a turn which has not yet been fully completed. Speakers might start their turn with fillers or particles (e.g. *well, um, uh*) that may be used independently of the content of a previous turn. In this case, speakers start preparing their turn without the need for listening to the current turn until the end. However, CA and corpus studies (e.g., Pomerantz, 1984; Clark & Fox Tree, 2002; but cf. O'Connell & Kowal, 2005; Kendrick & Torreira, 2014) have robustly shown that there are no speech components which are entirely unrelated to the action that a given turn implements. For example, hesitations and particles like *well* in English are likely to appear at the beginning of turns which do not fully conform to the expectations set by a preceding turn. These turns express "dispreferred" actions in CA terms (Pomerantz, 1984;. Kendrick & Torreira, 2014). This suggests that speakers design turn-beginnings already with an idea of what sort of action they will implement during their turn.

A more likely possibility is that listeners predict the trajectory of the turn in progress. Eye-tracking and EEG studies revealed that listeners make predictions at different levels of

language comprehension. In the visual word paradigm of eye-tracking studies, participants' eye-movements are recorded as they observe a visual scene and listen to sentences that refer to objects in that scene (Tanenhaus et al., 1995). These studies demonstrate that participants look to possible referents in the visual scenes prior to those being mentioned in the sentence (e.g., Kamide, Altmann & Haywood, 2003; Knoeferle et al., 2005; Altmann & Kamide, 2007). This implies that participants combine the semantic analysis of the incoming speech with the visual context early on, which narrows the possible trajectories speech might take and affords predictive understanding. ERP studies have also demonstrated that listeners not only predict developments in the situation under discussion but can also predict the lexical gender or the phonological form of specific words (Wicha, Moreno & Kutas, 2004; DeLong, Urbach & Kutas, 2005; Van Berkum et al., 2005).

It has been suggested that predictions facilitate the speed of language comprehension and can help to disambiguate noisy input during natural language use (see e.g., Kutas, DeLong & Smith, 2011; Pickering & Garrod, 2007). For example, contextually based predictions greatly reduce the number of activated lexical candidates during the processing of an incoming acoustic signal. Consequently, word recognition happens quickly, and it is completed within a few hundred milliseconds (see Hagoort & Poeppel, 2013). Kutas and colleagues argue that the comprehension system is not only able to facilitate the processing of linguistic information, but it also processes information that has already been predicted but not yet encountered in the linguistic input (Kutas et al., 2011). This suggest that listeners could prepare their response to predictable turns which have not yet fully been uttered.

Studies of predictive comprehension are in line with Sacks and colleagues' model of turn-taking (1974) which assumes that listeners project the type of construction that will end the turn. Most recent models of turn-taking (e.g., Levinson & Torreira, 2015; Garrod & Pickering, 2015; Magyari & de Ruiter, 2012) also acknowledge that listeners probably predict the content of turns and start speech production before the turn-end in order to produce a smooth transition. These models also assume that the speech act expressed in the turn might be recognized early during the turn or even before the turn (see Bögels & Levinson, 2017). A few experimental studies showed action recognition in a turn early on. Gisladottir, Chwilla, and Levinson (2015) presented auditory two-turn interactions to participants. The target utterances (e.g. *I have a credit card.*) were preceded by three different utterances (e.g., *I can lend you money for the ticket.*, *I don't have any money to pay for the ticket*., *How are you going to pay for the ticket?*). Therefore, the targets performed three different actions depending on the

previous utterance (e.g., a declination in response to an offer, a pre-offer in response to the statement of a problem, and an answer in response to a question). Participants' event-related potential (ERP) for utterances expressing declinations were more positive at frontal channels than the ERP for answers already around 200 milliseconds after the onset of the target. One other EEG study (Egorova, Sthyrov & Pulvermüller, 2013) compared overhearers' ERPs for naming (i.e. referring to an object) and for requesting (i.e. uttering the name of the object in order to get it). The ERPs of the two action types diverged as early as ~120 ms after the onset. In these experiments, during the test phase the same utterance expressed different actions depending on the context. In conversations, action types often differ already in their linguistic form, for example, in directness or length (e.g., Csató & Pléh, 1987/88). This suggest that even if an action cannot be anticipated based on the context, it might be recognized early on after a few words.

Taken together, there is much evidence for prediction during comprehension. These processes might also help participants in conversations to anticipate the content of turns well in advance and begin to prepare for their response in time. However, this observation also suggests that listeners' speech processing system is engaged in several processes while the other is speaking: 1) comprehension of the current turn, 2) prediction of the not-yet-encountered part of the turn, 3) preparation for the response. There is little consensus about how these tasks are allocated in the speech processing system. Levinson and Torreira (2015) suggest a gradual switch of resources from comprehension of the incoming turn towards the production of the ensuing response. According to Pickering and Garrod's (2013) model of speech processing listeners use the production system to make predictions of the other's turn and prepare their own response. In this model, listeners predict by simulation, i.e. they estimate the speaker's intention using covert imitation and then they use this intention to predict the speaker's completion of the current turn as if they did it for their own utterance. Future research could reveal how these processes intertwine during conversations.

Speech planning might start early

Another debated issue is when listeners who are to be the next speaker start response planning. Levinson and Torreira (2015) suggest that the listener's production system might automatically begin to prepare a response when the speech act of the current turn is recognized, but

articulation of the prepared response starts only at the imminent completion of the current turn. An EEG study showed EEG effects in line with this account (Bögels, Magyari & Levinson, 2015). Participants answered questions in an interactive quiz-paradigm. The experimenter interacted freely with the participant through an intercom system and asked quiz-questions which were answered by the participant. Although the questions were pre-recorded, participants thought they were asked by the experimenter live. In one condition the questions could be answered easily already at the half of the question while in the other condition only after encountering the last word (e.g., *Which character, also called 007, appears in the famous movies?* and *Which character from the famous movies is also called 007?*). ERP responses showed a large positivity localized at language-processing brain areas at the point where the question could be answered. Around the same time, there was a suppression of alpha band activity at the back of the scalp indicating a switch in attention from comprehension to production planning. These results suggested that planning of a turn started as soon as it was possible to formulate a response.

Other studies using a so called dual-task paradigm found evidence for later production planning. In Sjerps and Meyer's (2015) study participants' task was to name pictures on a screen while they simultaneously performed a finger-tapping task. Objects were presented in two rows on the screen. First, participants listened to the pre-recorded names of the pictures in the first row on the screen, then they named the pictures in the second row themselves. Participants' finger-tapping performance deteriorated only in the last 500 milliseconds before the end of the recorded picture-names. Eye-tracking results also showed that participants started to look at the second row around this time. This suggest that participants started to plan their turn (i.e. naming the pictures) only towards the end of the incoming turn. The setting of this experiment was, however, very different from a conversational setting. For example, the content of the participants' turn was not contingent on the previous turn (pre-recorded picture name), and finger-tapping is a non-linguistic task, hence, it might lead to a different load for the attentional and language processing systems.

It is likely that listeners can often predict content of turns during conversation. They might also start to prepare for the response as soon as they are able to predict the content. Most of the recent models of turn-taking suggest that content prediction might still not be enough to achieve timely responses, and speakers also need to time the articulation of their turn close to the turn-end of the previous speaker.

Prediction of the turn-ends

Speakers probably time their turns

Some researchers challenged precision timing in turn-taking. Heldner and Edlund (2010) claim that the most common between speaker interval, a short gap of 200 ms shows that timing of turn-taking is not precise. Moreover, they also conclude that the high distribution of turn-transition times (i.e. many overlaps and gaps) suggest that speakers do not aim for holding turn-transition durations constantly around 200 ms. However, Levinson and Torreira (2015) pointed out that listeners perceive silences between turns only if those are longer than 150-250 ms. Therefore, the most commonly observed gaps of ~200 ms cannot be achieved if speakers just react to the silence at the turn end even if what they want to say is already prepared. Turn-transitions have a unimodal distribution with a peak around 200 ms (Stivers et al., 2009) and longer gaps and silences have an interactional significance (e,g, Schegloff, 2000; Kendrick & Torreira, 2015) as described earlier. This suggest that short gaps are the "default mode" of conversations. Speakers not only prepare their response in advance but they aim to produce well-timed responses.

Three main schools of thought have emerged with regards to predicting turn-ends. Some researchers argue that interactants predict or "project" the overall structure and even the words of the turns, and so they can predict when a turn is going to end (Sacks et al., 1974; de Ruiter, Mitterer & Enfield, 2006; Magyari & de Ruiter, 2012). Others characterized possible turn-yielding cues that appear just before turn-ends and that are assumed to signal that the speaker will finish the current turn soon (Duncan, 1974; Local & Walker, 2012; Levinson & Torreira,R2 5 2015). A few studies propose that speakers time their turn by entrainment to the other's rate of syllable production (Wilson & Wilson, 2005; Garrod & Pickering, 2015).

Turn-yielding cues

According to Levinson & Torreira's (2015) turn-taking model, speakers prepare early their response during their conversational partner's turn and they hold articulation until they

encounter turn-yielding cues in the partner's turn. Turn-yielding cues are those perceptual features of behaviour that appear towards the end of conversational turns and signal that the current turn is coming to an end. Most of the suggested cues are prosodic, for instance, final syllable lengthening or pitch changes in the last word (e.g., Duncan, 1974; Duncan & Fiske, 1977; Local & Walker, 2012; Schegloff, 1996). Others also proposed non-verbal signals, for example, particular kind of eye-gaze and gesture (Kendon, 1967; Duncan, 1974). However, corpus studies of conversations have revealed that fast turn-transitions are frequent regardless of whether the conversation takes place face-to-face or in a telephone-like situation (Ten Bosch, Oostdijk & de Ruiter, 2005; de Ruiter et al., 2006; Stivers et al., 2009, but already noted in Levinson, 1983, p. 302). Non-speech cues might be used for turn-end predictions in face-to-face interactions, but these cannot account for the precise timing achieved in telephone conversations.

With regard to prosody, some experiments studied whether listeners perceive pitch changes as turn-yielding (e.g. Beattie, Cutler & Pearson, 1982; Schaffer, 1983; Cutler & Pearson, 1986). Recently, Bögels and Torreira (2015) used a button-task paradigm to study the effect of intonational phase boundaries (a pitch rise and syllable-lengthening) on turn-end predictions. The intonational phase boundaries were placed either in the middle or at the end of longer and shorter questions. Participants were asked to press a button exactly when the question ended (similarly to the task in de Ruiter et al., 2006). Hence, this was not a reaction time task, because participants had to predict the end of the question in order to press the button at the same time with the end. When a longer question contained an intonational phase boundary in the middle, participants pressed the button in one-third of the cases even in the middle of the question. Thus, the results of this study suggest that listeners use intonational cues to predict the turn-end.

However, de Ruiter and his colleagues' (2006) study reached a different conclusion. In their study, participants listened to turns from recordings of spontaneous conversations and pressed a button when a turn ended. The recordings were either played without modification or were modified so that either the intonation contour or the lexical information were missing. When participants listened to turns without intonation contour and consequently the words in the turn could still be understood, there was no change in the accuracy of the button-presses (compared to the performance with the original recordings). But when the words were obscured and intonational contour remained intact, participants' performance got worse. De Ruiter and colleagues concluded that intonation was neither sufficient nor necessary to predict turn-ends,

and that syntactic and lexical information played a major role in the timing of turn-taking. This study, however, did not take into account non-pitch prosodic information (see Rühlemann & Gries, 2020 for a list of different types of turn-final cues). For example, word final lengthening was also present at intonational phase boundaries beside a rise in pitch in Bögels and Torreira's (2015) study. This latter study, however, used questions in Dutch that were recorded as part of semi-spontaneous conversations. An experimenter blind to the purpose of the recordings read the questions from a paper to participants. Although the study showed that listeners use intonation for turn-end predictions, it is not sure that the cues used in their study also appear in natural conversations. A study of English natural conversations found, for example, that changes of word duration (lengthening) does not affect only the last word but larger proportions of a turn (Rühlemann & Gries, 2020). According to the authors, lengthening is not a "one-off" cue marking the turn-end, but it projects the durational envelope of the turn. This result support models of turn-taking (see later) which assume "long distance" predictions, i.e. that the length of turns is estimated by the speakers.

The early development of turn-taking shows that infants engage in turn-taking like interactions, so called "protoconversations" with their caregiver even before they could speak (Trevarthen, 1977; Bruner, 1983). Preverbal infants probably understand little of the semantic and syntactic content of turns, therefore, intonation might play a crucial role in early turn-taking. However, the average of the turn-transition durations is about 1.5 s at 3-months (Bateson, 1975). Although the duration of gaps starts to decrease in the following months but it starts to increase again at 9 months (Hilbrink, Gattis & Levinson, 2015), and it remains at around a second even for 5-years-olds (Garvey & Berninger, 1981, see Levinson & Torreira, 2015 for a short review). Turn-transitions between children remain even slower for a long time than turn-transitions between adults and children (see Pléh, 2012). Hence, early forms of turn-taking are not as precisely timed as conversations between adults. The slowness of the interactions could arise from difficulties with understanding and with formulating sounds and speech, but it is also possible that different mechanisms support turn-taking in children compared to adults.

A long tradition of interaction research has identified a set of prosodic features that typically coincide with turn-ends. Experimental studies (e.g. Bögels & Torreira, 2015; Barthel, Meyer & Levinson, 2017) have shown that speakers probably use such turn-yielding cues if those are available. However, it also seems that speakers are also able to produce well-timed turns relying on other sources of information, i.e. on the semantic and syntactic content of turns.

While turn-final cues are assumed to signal imminent turn-ending, long-range predictions might also be involved in turn-end predictions.

Estimation of turn-duration

Sacks and his colleagues (1974) suggested that listeners use the syntactic frame to project the overall structure of the incoming turn, and thereby predict when a turn is going to end. In addition, they acknowledge that intonation as well could project the length of an utterance.

Only a few experimental studies targeted the role of prosody in the predictions of turn-length. Grosjean studied whether listeners can predict how long a sentence will last based on the prosodic information at sentence-beginnings in English (Grosjean, 1983). Grosjean found that prosodic information becomes available for the prediction of sentence length only when semantic and syntactic information cannot help anymore. However, this study also presented sentences read aloud, therefore, we do not know whether the results can be generalized to prosodic information carried by conversational turns.

Prediction of last words and number of words might provide long-range predictions of the turns' content as well. In an experiment, the prediction of the last word of turns and the prediction of the number of words in turns correlated with well-timed turn-end predictions (Magyari & de Ruiter, 2012). In this experiment, a subset of turns was used from de Ruiter and colleagues' (2006) experiment in which participants listened to the turns and tried to press a button right when the turn-ended. The turns were truncated at several points prior to the end. Participants' task was to listen to a whole turn or to a truncated segment and try to guess whether the segment ended. If they guessed that a segment did not end they were asked to guess how the segment continued. Participants' guesses about the last words were more accurate with segments that received more accurate button-presses in the earlier experiment. The number of guessed words also correlated with the button-presses: segments of turns with later button-presses were associated with a larger number of guessed words. Magyari and de Ruiter suggested that anticipation of syntactic frames or words can facilitate the accurate timing of the production of a next turn. Another study (Magyari, Bastiaansen, de Ruiter & Levinson, 2014) provided electrophysiological evidence for long-range predictions of turn-ends based on the anticipation of the content of turns. This study found a neuronal correlate, beta frequency

desynchronization already 1250 ms before the end of turns when the content of the turns could be predicted. The beta desynchronization was localized in the anterior cingulate cortex and the inferior parietal lobule which are associated, respectively, with anticipation, attention, time-processing (Bubic, von Cramon, & Schubotz, 2010; Aarts, Roelofs, & van Turennout, 2008; Lewis & Miall, 2006; Macar et al., 2002; Fuster, 2001) and with language processing (Lau, Phillips, & Poeppel, 2008).

Riest, Jorschick and de Ruiter (2015) studied whether advance knowledge about the turn can help to predict the content of the turn, and hence, whether it facilitates turn-end predictions. They presented participants with turns from natural conversations and employed the button-press paradigm for turn-end predictions. In one condition, the turns presented auditorily were preceded by the visual presentation of the transcription of the turn. There was no difference in the accuracy of the button-presses of the turns preceded by their description compared to the turns presented alone. This finding is in line with the results of Corps and colleagues' study in which participants were presented with questions using the button-press paradigm (Corps, Crossley, Gambi, & Pickering, 2018). The final words of the speaker's question were either predictable (e.g. *Are dogs your favourite animal?*) or unpredictable (e.g. *Would you like to go to the supermarket?*) given the preceding context. They found no effects of content predictability on the timing of button-press responses. They concluded that content predictions are not used for predicting when the turns end. However, an alternative explanation is also possible. When listeners cannot predict the exact words of turns, they might still predict the number of words or the syntactic phrase (e.g. *Would you like to go to [a noun referring to some place]*?) which might be also sufficient for turn-end prediction (see also Magyari & De Ruiter, 2012).

In Riest and colleagues' study, participants were also presented with turns from which closed class words were removed, and with turns from which open class words were removed. Closed class words, e.g., preposition, articles, conjunctions are assumed to have a syntactic role in sentences, while open class words, e.g., nouns, verbs, adjective are assumed to contribute to the meaning of sentences. Participants were able to anticipate the turn-end when closed class words were removed, although their performance was worse compared to the results of button-presses when the original turns were presented. When the open class words were removed, participants' performance deteriorated even more, they pressed the button often after the turns ended. They concluded that both semantic and syntactic information are necessary for turn-end anticipation, but semantic information appears to be more important. Semantic information

could facilitate turn-end predictions, because it enables listeners to predict the words which end the turn. Hence, the results of Riest and colleagues' study is in line with accounts (Magyari & de Ruiter, 2012) that assume that turn-end predictions are also based on anticipation of words.

Although studies claim that prediction of words and content aid turn-end prediction, it is not specified how semantic expectations lead to estimation of turn-length. Garrod and Pickering's (2015) model provides a mechanism for how word predictions lead to well-timed turn-transitions. They suggest that listeners predict the last words of turns by covertly imitating the speaker's utterance. The predicted linguistic representations are combined with the interlocutor's speech rate to which the listener's brain circuitry is entrained by low-level acoustic analysis. In order to speak next, listeners who need to determine the appropriate timing for turn transitions use these predictions. This model has been recently supported by an experimental study (Corps, Gambi, & Pickering, 2020) which showed that speakers use speech rate information to time their articulation.

Conclusion

The tight timing of turn-taking poses a challenge for models of language processing. Most of the recent models agree that speakers need to prepare their response well before the previous turn ends. And for this, the content (or at least overall message) of the previous turn should be anticipated. Experimental studies showed that response preparation can start as early as possible (i.e. when the content of a turn can be anticipated) (e.g., Bögels et al., 2015; Barthel et al., 2017). Other study (Sjerps & Meyer, 2015) found that response preparation started only towards the end of turns when participants also executed a non-linguistic task in parallel. These results suggest that the start of response preparation does not only depend on when the formulation of a response is possible but also on the attentional load during listening to the turn. Further research could study the factors influencing the start of response preparation.

Experimental investigations also showed that listeners use turn-final cues to predict turn-ends (e.g., Bögels & Torreira, 2015; Barthel et al., 2017). However, these studies often used semi spontaneous recordings for stimuli material. Other studies using recordings of natural conversations found that listeners estimate turn durations based on anticipated syntactic

structures, semantic information and words (Magyari & De Ruiter, 2012; Magyari et al., 2014; Riest et al., 2015). Hence, all sources of information (intonation, semantic and syntactic information) might be used in timing of turn-transitions. However, some accounts disagree about whether turn-yielding cues in intonation are used at all. Levinson and Torreira (2015) argued that articulation is launched when turn-yielding cues are detected. In contrast with this, De Ruiter and colleagues (2006) concluded that intonation is not necessary for timing turn-transitions.

Although studies show that semantic and syntactic information facilitate turn-end predictions, these do not explain how anticipation of the linguistic content can lead to precise timing. Garrod and Pickering's (2015) model provides a solution for how predicted linguistic representations can be combined with the other's speech rate for well-timed responses. Future research should develop experiments and a model which could reconcile the competing accounts and experimental results of turn-taking.

Acknowledgements

First of all, I thank Csaba Pléh whose works and talks have been always a source for inspiration for my research. Some parts of this review is based on the Introduction of my doctoral thesis. I am immensely grateful for my promotor, Steve Levinson's advice and comments on the text of the Introduction. I also thank Bálint Forgács for his comments on an earlier version of the manuscript. The writing of this review is part of a project that has received funding from the Europen Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 845343.

References

Aarts, E., Roelofs, A., & van Turennout, M. (2008). Anticipatory Activity in Anterior Cingulate Cortex Can Be Independent of Conflict and Error Likelihood. *Journal of Neuroscience*, *28*(18), 4671–4678. https://doi.org/10.1523/JNEUROSCI.4400-07.2008

Altmann, G.M.T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, *57*, 502–518.

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

Beattie, G.W., Cutler, A., & Pearson, M. (1982). Why is Mrs Thatcher interrupted so often? *Nature*, *300*(5894), 744–747. https://doi.org/10.1038/300744a0

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*(1), 71–89. https://doi.org/10.1080/08351813.2017.1262118

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

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

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

Bruner, J. (1983). *Child's Talk*. Norton.Bubic, A., von Cramon, D.Y., & Schubotz, R.I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*, 1-15. https://doi.org/10.3389/fnhum.2010.00025

Caramazza, A. (1997). How Many Levels of Processing Are There in Lexical Access? *Cognitive Neuropsychology*, *14*(1), 177–208.

Clark, H.H., & Fox Tree, J.E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

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. (2020). How do listeners time response articulation when answering questions? The role of speech rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 781–802.

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

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

Dell, G.S. (1986). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review*, *93*(3), 283–321.

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

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, *3*, 161-180.

Duncan, S., & Fiske, D.W. (1977). *Face-to-face Interaction: Research, methods and theory*. Lawrence Erlbaum.

Egorova, N., Shtyrov, Y., & Pulvermüller, F. (2013). Early and parallel processing of pragmatic and semantic information in speech acts: Neurophysiological evidence. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00086

Fuster, J.M. (2001). The Prefrontal Cortex—An Update: Time is of the Essence. *Neuron*, *30*, 319-333.

Garrod, S., & Pickering, M.J. (2015). The use of content and timing to predict turn transitions. *Frontiers in Psychology*, *6*(751), 1–12. http://dx.doi.org/10.3389/fpsyg.2015.00751

Garvey, C., & Berninger, G. (1981). Timing and turn taking in children's conversations. *Discourse Processes*, *4*(1), 27–57. https://doi.org/10.1080/01638538109544505

Gisladottir, R.S., Chwilla, D.J., & Levinson, S.C. (2015). Conversation Electrified: ERP Correlates of Speech Act Recognition in Underspecified Utterances. *PLOS ONE*, *10*(3), e0120068. https://doi.org/10.1371/journal.pone.0120068

Gleitman, L.R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(4), 544–569. https://doi.org/10.1016/j.jml.2007.01.007

Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the on-line processing of language. *Linguistics*, *21*(3). https://doi.org/10.1515/ling.1983.21.3.501

Hagoort, P., & Poeppel, D. (2013). The infrastructure of the language-ready brain. In M. A. Arbib (Ed.), *Music, language, and the brain*. MIT Press. https://nyuscholars.nyu.edu/en/publications/the-infrastructure-of-the-language-ready-brain

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

Hilbrink, E.E., Gattis, M., & Levinson, S.C. (2015). Early developmental changes in the timing of turn-taking: A longitudinal study of mother-infant interaction. *Frontiers in Psychology*, *6*, 1492. https://doi.org/10.3389/fpsyg.2015.01492

Indefrey, P., & Levelt, W.J.M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

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

Jescheniak, J.D., & Levelt, W.J.M. (1994). Word frequency effects in speech production: Retreival of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843.

Jescheniak, J.D., Schriefers, H., & Hantsch, A. (2003). Utterance Format Affects Phonological Priming in the Picture-Word Task: Implications for Models of Phonological encoding in Speech Production. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 441-454.

Kamide, Y., Altmann, G.T.M., & 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.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22–63.

Kendrick, K.H., & Torreira, F. (2015). The Timing and Construction of Preference: A Quantitative Study. *Discourse Processes*, *52*, 255–289.

Knoeferle, P., Croecker, M. V., Scheepers, C., & Pickering, M.J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*, 95–127.

Konopka, A. E., & Meyer, A. S. (2014). Priming sentence planning. *Cognitive Psychology*, *73*, 1–40. https://doi.org/10.1016/j.cogpsych.2014.04.001

Kutas, M., DeLong, K.A., & Smith, N.J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In Bar, M. (Ed.), *Predictions in the Brain: Using Our Past to Generate a Future* (pp. 190-207.). Oxford University Press.

Lau, E.F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. MIT Press. Levelt, W.J.M., & Meyer, A. S. (2000). Word for word: Multiple lexical access in speech production. *European Journal of Cognitive Psychology*, *12*(4), 433–452. https://doi.org/10.1080/095414400750050178

Levinson, S.C. (1983). *Pragmatics*. 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.2015.00731

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

Lewis, P.A., & Miall, R.C. (2006). Remembering the time: A continuous clock. *Trends in Cognitive Sciences*, *10*(9), 401–406. https://doi.org/10.1016/j.tics.2006.07.006

Local, J., & Walker, G. (2012). How phonetic features project more talk. *Journal of the International Phonetic Association*, *42*(3), 255–280.

Macar, F., Lejeune, H., Bonnet, M., Ferrara, A., Pouthas, V., Vidal, F., & Maquet, P. (2002). Activation of the supplementary motor area and of attentional networks during temporal processing. *Experimental Brain Research*, *142*, 475–485.

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

Magyari, L., Bastiaansen, M.C.M., de Ruiter, J.P., & Levinson, S.C. (2014). Early Anticipation Lies behind the Speed of Response in Conversation. *Journal of Cognitive Neuroscience*, *26*(11), 2530–2539. https://doi.org/10.1162/jocn_a_00673

Markó, A. (2006). *Beszélőváltás a társalgásban*. IX. Balatonalmádi Pszicholingvisztikai Nyári Egyetem.

O'Connell, D.C., & Kowal, S. (2005). Uh and Um Revisited: Are They Interjections for Signaling Delay? *Journal of Psycholinguistic Research*, *34*(6), 555–576. https://doi.org/10.1007/s10936-005-9164-3

Pickering, M.J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105-110.

Pléh, C. (2012). *A társalgás pszichológiája*. Libri Kiadó.

Pomerantz, A. (1984). Agreeing and disagreeing with assesments: Some features of preferred/dispreffered turn shapes. In J.M. Atkinson & J. Heritage (Eds.), *Structures of Social Action* (pp. 53-101.). Cambridge University Press.

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

Roberts, F., Margutti, P., & Takano, S. (2011). Judgments Concerning the Valence of Inter-Turn Silence Across Speakers of American English, Italian, and Japanese. *Discourse Processes*, *48*(5), 331–354. https://doi.org/10.1080/0163853X.2011.558002

Rühlemann, C., & Gries, S.T. (2020). Speakers advance-project turn completion by slowing down: A multifactorial corpus analysis. *Journal of Phonetics, 80*, 100976. https://doi.org/10.1016/j.wocn.2020.100976

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

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, *11*, 243–257.

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

Schegloff, E.A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E.A. Schegloff, & S.A. Thompson (Eds.), *Interaction and grammar* (pp. 52-133.). Cambridge University Press.

Schegloff, E.A. (2007). *Sequence Organization in Interaction: A Primer in Conversation Analysis* (Vol. 1). Cambridge University Press.

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

Stivers, T. (2010). An overview of the question-response system in American English conversation. *Journal of Pragmatics*, *42*, 2772–2781.

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

Strijkers, K., Costa, A., & Thierry, G. (2010). Tracking Lexical Access in Speech Production: Electrophysiological Correlates of Word Frequency and Cognate Effects. *Cerebral Cortex*, *20*, 912–928.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634. https://doi.org/10.1126/science.7777863

Ten Bosch, L., Oostdijk, L., & De Ruiter, J P. (2005). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. *Presented at the 7th International Conference of Neural Information Processing ICONIP*.

Trevarthen, C. (1977). Descriptive analyses of infant communicative behavior. In H. R. Schaffer (Ed.), *Studies in mother-infant interaction* (pp. 227-270.). Academic Press.

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

Wagner, V., Jescheniak, J. D., & Schriefers, H. (2010). On the flexibility of grammatical advance planning during sentence production: Effects of cognitive load on multiple lexical access. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(2), 423-440. https://doi.org/10.1037/a0018619

Wicha, N.Y.Y., Moreno, E.M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potentials study of semantic integration, gender expectancy and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272-1288.

Wilson, M., & Wilson, T.P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, *12*(6), 957–968.